INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2026IPPAG003

Thèse de doctorat

ENSAE

IP PARIS

On-Policy and Off-Policy Learning for Large Action Spaces

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

École doctorale n°574 École Doctorale de Mathématique Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 13 mars 2026, par

IMAD AOUALI

Composition du Jury :

Vianney Perchet Professeur, CREST, ENSAE, IP Paris	Président
Olivier Cappé Directeur de recherche, CNRS	Rapporteur
Aurélien Garivier Professeur, Ecole Normale Superieure de Lyon	Rapporteur
Claire Vernade Professeur, University of Technology Nuremberg	Examinatrice
Victor-Emmanuel Brunel Professeur, CREST, ENSAE, IP Paris	Directeur de thèse
Anna Korba Assistant professor, CREST, ENSAE, IP Paris	Co-encadrante de thèse
David Rohde Chercheur, Criteo AI Lab	Invité

Acknowledgements

First and foremost, I wish to express my sincere and profound gratitude to my PhD supervisors, Victor-Emmanuel Brunel, Anna Korba, and David Rohde. It has been an immense privilege to learn from and work with them over these years. They shaped my research and personal growth in ways that will stay with me far beyond this thesis.

Victor brought rigor, kindness, and openness to every discussion, profoundly shaping how I approach research and problem-solving. Anna's brilliance, drive, and compassionate mentorship were central to the success of this PhD; her rare ability to combine deep technical insight with empathy and encouragement made her guidance invaluable. David was the best manager I could have hoped for, whose trust, patience, and human approach made all the difference when navigating both professional and personal challenges.

I am profoundly grateful to the members of my thesis jury for the time, care, and expertise they devoted to evaluating this work. I would first like to express my deepest appreciation to Olivier Cappé and Aurélien Garivier for accepting the demanding role of rapporteurs, and for the considerable time and attention they dedicated to reading this manuscript in depth. I am sincerely grateful for their careful assessment, thoughtful comments, and constructive feedback. I would also like to warmly thank Vianney Perchet for the honor of presiding over the jury, and for his invaluable support. Finally, I am deeply grateful to Claire Vernade for serving as examinatrice, and for her generosity, encouragement, and support. It was a true privilege to have such distinguished researchers on my jury, and I deeply appreciate their scientific perspective, insightful remarks, and kindness.

I would also like to thank Criteo, CAIL, and the Performance Science team, as well as CREST, ENSAE, Institut Polytechnique de Paris. Both the company and the laboratory provided outstanding scientific and institutional support throughout this thesis.

I am deeply grateful to several mentors: to Branislav Kveton, whose insights as my first major co-author shaped my approach to research and publication; to Florian Strub, my DeepMind scholarship mentor, whose guidance inspired me to pursue a PhD; and to Flavian Vasile, and Michal Valko, for their invaluable support, both seen and unseen.

My heartfelt thanks also go to all my co-authors, whose contributions have greatly enriched this work: A. Gilotte, B. Heymann, N. Nguyen, A. György, P. Alquier, N. Chopin, S. Katariya, AAS. Hammou, S. Ivanov, A. Benhalloum, M. Bompaire, M. Vono, M. Gartrell, V. Zaytsev, D. Legrand, and O. Jeunen. Special recognition goes to O. Sakhi,

M. Cherifa and A.B. Yahmed who were exceptional companions throughout this journey.

Finally, to my family and friends: my deepest thanks to my mother and siblings for their unwavering love and support. This thesis is dedicated to my late father, whose dream was for me to pursue a PhD. To Basma, Hakim, Achraf, Ismail, Hicham, Ayman, Tayeb, Anas, Nicolas, Youssef, Kini, Song, Issam, Charif, Yassine, Oussama, Abdellah, Ali: thank you for your friendship and presence throughout this journey.

Abstract

Many interactive systems (e.g., recommender systems) can be modeled as contextual bandits. This framework captures the core challenge of decision-making under uncertainty: selecting actions based on context while learning from partial, noisy feedback. Learning in this setting follows two paradigms: *on-policy learning*, in which agents collect data and update their policy simultaneously in real time, and *off-policy learning*, in which the agent's policy is learned offline from static logs collected under a different policy. Standard algorithms for both paradigms struggle to scale to large action spaces, facing either computational intractability or statistical inefficiency. This thesis develops principled and practical methods to make contextual bandit algorithms tractable in large action spaces, advancing both paradigms through novel algorithmic and theoretical contributions.

For on-policy learning, we introduce structured Bayesian models that enable efficient exploration via information sharing. Our first contribution, mixed-effect Thompson sampling (`meTS`) (Chapter 3), couples action parameters through shared latent effects. This reduces Bayesian regret to $\tilde{\mathcal{O}}(\sqrt{TdK_{\text{eff}}})$, where K_{eff} is an effective number of actions. When the number of shared effects is much smaller than the number of actions ($L \ll K$), we have $K_{\text{eff}} \ll K$, yielding significant regret reduction. Moreover, `meTS` achieves dramatic improvements in both memory complexity (from $\mathcal{O}(K^2d^2)$ to $\mathcal{O}((L^2+K)d^2)$) and runtime complexity (from $\mathcal{O}(K^3d^3)$ to $\mathcal{O}((L^3+K)d^3)$). We extend this framework to diffusion Thompson sampling (`dTS`) (Chapter 4), which leverages deep generative models to capture complex action distributions. `dTS` further improves memory complexity to $\mathcal{O}((L+K)d^2)$ and runtime complexity to $\mathcal{O}((L+K)d^3)$, where L denotes the number of layers in the diffusion model. Both methods perform well empirically as analyzed without additional hyperparameter tuning, making them highly practical.

For off-policy learning, we address fundamental bottlenecks through three complementary approaches. First, in Chapter 6, we develop the structured direct method (`sDM`), which models action parameters using a shared latent structure. We prove that `sDM` achieves $\mathcal{O}(1/\sqrt{n})$ convergence in Bayesian suboptimality without requiring the restrictive full logging support assumption. `sDM` performs well in practice, and the performance gap between `sDM` and standard direct methods widens as the action space grows. Second, in Chapter 7, we challenge the conventional wisdom that better reward estimation yields better policies. We demonstrate that optimization intractability, rather than estimation accuracy, becomes the primary bottleneck in large action spaces, and advocate for policy-weighted

log-likelihood objectives that prioritize optimization tractability; these consistently outperform sophisticated estimators on datasets with up to one million actions. Third, in Chapter 8, we address importance sampling variance by combining variance-reducing regularization with principled pessimism. Our exponential smoothing estimators and unified PAC-Bayesian analysis yield tractable learning objectives amenable to stochastic optimization, providing concentration bounds and superior empirical performance.

We validate these theoretical and algorithmic advances through extensive experiments on synthetic and real-world datasets. By developing scalable algorithms for both learning paradigms, this thesis enables the deployment of contextual bandits in modern applications where action spaces routinely exceed thousands or millions of actions.

Contents

Notation

Notation

General Mathematical Notation

Symbol	Definition
$[n]$	Set of first n positive integers: $\{1, 2, \ldots, n\}$
$\mathbb{R}^d$	d-dimensional real vector space
I_d	Identity matrix of dimension $d \times d$
$\Delta(\mathcal{A})$	Probability simplex over action set $\mathcal{A}$
$\mathcal{O}(\cdot)$	Big-O notation for upper bounds
$\otimes$	Kronecker product

Probability conventions. Random variables are denoted with capital letters, and their realizations with the respective lowercase letters, except for Greek letters. With a slight abuse of notation, for random variables X, Y, the distribution (or density) of $X \mid Y = y$ evaluated at x is denoted by $p(x \mid y)$.

Norms and Inner Products

Symbol	Definition
$\|\cdot\|$	Euclidean norm (unless specified otherwise)
$\|a\|_\Sigma$	Weighted norm: $\sqrt{a^\top \Sigma a}$ for $a \in \mathbb{R}^d$, $\Sigma \succ 0$

Contextual Bandit Framework

Symbol	Definition
$\mathcal{X} \subset \mathbb{R}^d$	Context space (d-dimensional)
$\mathcal{A} = [K]$	Finite action set with K actions
T	Number of interaction rounds
n	Number of samples in logged dataset

Policies and Value Functions

Symbol	Definition
$\pi : \mathcal{X} \to \Delta(\mathcal{A})$	Stochastic policy mapping contexts to action distributions
$\pi(\cdot \mid x)$	Probability distribution over actions given context x
π_t	Policy at round t (on-policy setting)
π_0	Logging policy (off-policy setting)
π_*	Optimal policy (off-policy setting)
$\hat{\pi}$	Learned policy (off-policy setting)
$V(\pi)$	Value (expected reward) of policy π
$\hat{V}(\pi)$	Estimated value of policy π

Environment Components

Symbol	Definition
ν	Context distribution over $\mathcal{X}$
$p(\cdot \mid x, a)$	Conditional reward distribution given context x and action a
$r(x, a)$	Expected reward function: $\mathbb{E}_{R \sim p(\cdot \mid x,a)}[R]$
$\hat{r}(x, a)$	Estimated reward function
$r(x, a; \theta)$	Parametric reward model
$r(x, a; \hat{\theta})$	Estimated reward function in parametric case

Random Variables and Data

Symbol	Definition
X_t	Context observed at round t
A_t	Action taken at round t
R_t	Reward received at round t
$A_{t,*}$	Optimal action at round t: $\arg\max_{a \in \mathcal{A}} r(X_t, a)$
H_t	History of interactions: $\{(X_\ell, A_\ell, R_\ell)\}_{\ell < t}$
$\mathcal{D}_n$	Logged dataset: $\{(X_i, A_i, R_i)\}_{i=1}^n$

Performance Metrics

Symbol	Definition
$\mathcal{R}(T) = \sum_{t=1}^T r(X_t, A_{t,*}) - r(X_t, A_t)$	Cumulative regret
$\mathcal{BR}(T) = \mathbb{E}[\mathcal{R}(T)]$	Bayesian cumulative regret
$\textsc{so}(\hat{\pi}) = V(\pi_*) - V(\hat{\pi})$	Suboptimality gap of policy $\hat{\pi}$
$\textsc{Bso}(\hat{\pi}) = \mathbb{E}[\textsc{so}(\hat{\pi})]$	Bayesian suboptimality gap of policy $\hat{\pi}$

Matrix Operations and Concatenation

Symbol	Definition
$[a_1, a_2, \ldots, a_n]$	Horizontal concatenation of vectors into $d \times n$ matrix
$(a_i)_{i\in[n]}$	Vertical concatenation: $(a_1^\top, \ldots, a_n^\top)^\top \in \mathbb{R}^{nd}$
$\mathrm{Vec}(\cdot)$	Vectorization operator
$\mathrm{diag}((\mathrm{A}_i)_{i\in[n]})$	Block diagonal matrix with blocks $\mathrm{A}_1, \ldots, \mathrm{A}_n$
$(\mathrm{A}_i)_{i\in[n]}$	Vertical concatenation of matrices into $nd \times d$ matrix
$(\mathrm{A}_{i,j})_{(i,j)\in[n]\times[m]}$	Block matrix where $\mathrm{A}_{i,j}$ is the (i,j)-th block

Eigenvalues

Symbol	Definition
$\lambda_1(\mathrm{A})$	Maximum eigenvalue of matrix A
$\lambda_d(\mathrm{A})$	Minimum eigenvalue of matrix A

Résumé substantiel en français

Cette thèse étudie l'apprentissage séquentiel et contrefactuel dans des systèmes interactifs où l'espace des décisions est très grand. Ces systèmes sont aujourd'hui omniprésents : moteurs de recommandation, publicité computationnelle, places de marché, systèmes de tarification, robotique ou encore allocation de ressources. Leur fonctionnement repose sur une boucle d'interaction simple mais difficile à optimiser : à chaque étape, le système observe un contexte, choisit une action parmi un très grand nombre de possibilités, puis reçoit une récompense partielle et bruitée qui dépend conjointement du contexte et de l'action choisie. Dans un système de recommandation, le contexte peut représenter l'historique et les préférences d'un utilisateur, l'action correspond à l'item recommandé dans un catalogue contenant potentiellement des millions d'items, et la récompense mesure l'engagement de l'utilisateur, par exemple un clic ou un temps de visionnage. En publicité en ligne, l'action peut combiner le choix d'une annonce et d'un prix d'enchère, tandis que la récompense dépend d'événements successifs tels que le gain de l'enchère, le clic ou la conversion.

Le cadre mathématique central de cette thèse est celui des bandits contextuels. On considère un espace de contextes $\mathcal{X} \subset \mathbb{R}^d$, un ensemble fini d'actions $\mathcal{A} = [K]$, une distribution inconnue de contextes ν, et des distributions conditionnelles de récompenses $p(\cdot \mid x, a)$. La fonction de récompense moyenne est définie par

$$r(x, a) = \mathbb{E}[R \mid X = x, A = a].$$

À chaque tour t, un contexte $X_t \sim \nu$ est observé, l'agent choisit une action A_t selon une politique $\pi_t(\cdot \mid X_t)$, puis reçoit une récompense $R_t \sim p(\cdot \mid X_t, A_t)$. Ce formalisme capture l'essentiel de l'apprentissage interactif : l'agent ne voit que la récompense de l'action qu'il a effectivement choisie, et doit donc apprendre à partir d'un retour partiel. La difficulté principale analysée dans cette thèse est le passage à l'échelle lorsque K est très grand.

La thèse traite deux paradigmes complémentaires. Le premier est l'apprentissage en ligne, ou *on-policy learning*, dans lequel l'agent interagit séquentiellement avec l'environnement et met à jour sa politique au fil des observations. Sa performance est mesurée par le regret cumulé,

$$\mathcal{R}(T) = \sum_{t=1}^{T} \left(r(X_t, A_{t,*}) - r(X_t, A_t) \right),$$

où $A_{t,*}$ désigne l'action optimale dans le contexte X_t. L'enjeu fondamental est le compromis exploration-exploitation : l'agent doit explorer les actions incertaines pour apprendre, tout en exploitant les actions déjà estimées comme performantes afin de limiter la perte de récompense. Le second paradigme est l'apprentissage hors politique, ou *off-policy learning*, dans lequel l'agent ne peut plus interagir avec l'environnement et doit apprendre à partir d'un jeu de données journalisé

$$\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$$

collecté par une politique de logging π_0. L'objectif est alors d'apprendre une politique $\hat{\pi}$ de grande valeur

$$V(\pi) = \mathbb{E}_{X\sim\nu}\mathbb{E}_{A\sim\pi(\cdot|X)}[r(X, A)],$$

et la performance est mesurée par l'écart de sous-optimalité $V(\pi_*) - V(\hat{\pi})$. Dans ce second cadre, l'apprentissage exige un raisonnement contrefactuel : il faut estimer ce qui se serait passé si une autre action avait été choisie, alors même que les données ne contiennent que les actions effectivement sélectionnées par π_0.

Dans les deux paradigmes, la grande taille de l'espace d'actions amplifie les difficultés statistiques, computationnelles et d'optimisation. En ligne, explorer indépendamment des milliers ou millions d'actions devient prohibitif : chaque action reçoit peu d'observations, ce qui ralentit considérablement l'apprentissage et augmente le regret. Hors politique, la couverture des données se dégrade lorsque K croît : de nombreuses actions sont rarement ou jamais observées dans certaines régions de l'espace des contextes. Les méthodes fondées sur un modèle de récompense souffrent alors d'un fort biais d'extrapolation, tandis que les méthodes par pondération inverse des propensions peuvent avoir une variance très élevée lorsque $\pi_0(a \mid x)$ est faible. Cette thèse montre également qu'un autre obstacle, souvent sous-estimé, devient dominant dans les grands espaces d'actions : l'optimisation des objectifs hors politique standards peut devenir intrinsèquement difficile, indépendamment de la qualité statistique de l'estimateur de valeur.

La première partie de la thèse est consacrée à l'apprentissage en ligne dans les grands espaces d'actions. Les méthodes classiques telles que *Upper Confidence Bound* et *Thompson Sampling* reposent souvent sur des modèles disjoints, où chaque action a possède son propre paramètre θ_a et où la récompense est modélisée sous la forme $r(x, a) = \phi(x)^\top \theta_a$. Ces modèles sont attractifs en pratique, notamment dans les systèmes de recommandation, car ils évitent de construire manuellement des caractéristiques conjointes contexte-action. Cependant, leur faiblesse est statistique : apprendre séparément un paramètre pour chaque action nécessite beaucoup de données par action, ce qui est incompatible avec de très grands catalogues. La contribution principale de cette partie consiste à conserver la flexibilité des modèles disjoints tout en introduisant des structures bayésiennes capables de partager l'information entre actions.

Le premier algorithme proposé est *mixed-effect Thompson Sampling*, ou `meTS`. Il repose sur un modèle bayésien hiérarchique où les paramètres d'actions θ_a sont couplés par des effets latents partagés $\Psi = (\psi_\ell)_{\ell\in[L]}$. Ces effets peuvent représenter, par exemple, des catégories ou des facteurs communs entre items. Le modèle suppose que les paramètres d'actions sont conditionnellement indépendants sachant les effets latents, mais qu'ils partagent de l'information à travers ces effets. À chaque tour, `meTS` échantillonne d'abord les effets

latents depuis leur postérieur, puis échantillonne les paramètres d'actions conditionnellement à ces effets, avant de choisir l'action maximisant la récompense échantillonnée. Cette procédure conserve le principe de Thompson Sampling tout en rendant l'exploration statistiquement plus efficace.

Dans le cas linéaire-gaussien, la thèse dérive des mises à jour exactes en forme fermée, et propose des approximations de Laplace tractables pour les modèles linéaires généralisés. L'analyse théorique établit une borne de regret bayésien de l'ordre

$$\widetilde{\mathcal{O}}\left(\sqrt{TdK_{\text{eff}}(\sigma_0^2+\sigma_\Psi^2)}\right),$$

où K_{eff} est un nombre effectif d'actions. Lorsque la structure latente est informative et que $L \ll K$, on a $K_{\text{eff}} \ll K$, ce qui conduit à une amélioration multiplicative par rapport à Thompson Sampling standard. Sur le plan computationnel, la factorisation conditionnelle réduit fortement les coûts mémoire et temps par rapport à une modélisation bayésienne dense de toutes les actions. Les expériences montrent que les gains de `meTS` augmentent avec la taille de l'espace d'actions, confirmant que le partage d'information est essentiel pour l'exploration à grande échelle.

La seconde contribution en ligne est *diffusion Thompson Sampling*, ou `dTS`. Cette méthode généralise `meTS` en remplaçant la hiérarchie à un niveau par une hiérarchie profonde inspirée des modèles de diffusion. Les paramètres d'actions sont générés au terme d'une chaîne de variables latentes reliées par des transformations non linéaires pré-entraînées. Cette structure permet de représenter des dépendances complexes entre actions, bien au-delà des effets linéaires ou catégoriels. Le défi technique est que le postérieur exact devient intraitable à cause des non-linéarités des fonctions de lien et du modèle de récompense. La thèse introduit donc une procédure d'inférence en ligne fondée sur des mises à jour de type gaussien, où les précisions postérieures combinent précision a priori et précision issue des données, et où les moyennes sont obtenues par combinaison pondérée entre les prédictions du prior et les estimations de maximum de vraisemblance.

L'intérêt de `dTS` est double. D'une part, l'algorithme exploite des priors riches appris hors ligne, par exemple à partir de représentations d'items, pour accélérer l'exploration en ligne. D'autre part, il conserve une structure de diffusion dans le postérieur, plutôt que de l'approximer par une simple gaussienne globale. Dans le cadre linéaire-gaussien, la thèse établit une borne de regret bayésien de l'ordre

$$\widetilde{\mathcal{O}}\left(\sqrt{TdK_{\text{eff}}\sum_{\ell=1}^{L+1}\sigma_\ell^2}\right),$$

et montre que la complexité peut être rendue linéaire en $L+K$. Empiriquement, `dTS` améliore les performances des méthodes de référence, y compris lorsque les priors de diffusion sont imparfaits ou appris à partir de données limitées. Cette contribution montre que les modèles génératifs profonds peuvent être utilisés non seulement pour représenter des actions, mais aussi pour structurer l'incertitude nécessaire à l'exploration.

La seconde partie de la thèse porte sur l'apprentissage hors politique dans les grands espaces d'actions. Les méthodes classiques se divisent principalement en méthodes directes,

qui apprennent un modèle de récompense $\hat{r}(x, a)$, et méthodes par *importance sampling*, qui estiment directement la valeur d'une politique au moyen du ratio $\pi(a \mid x)/\pi_0(a \mid x)$. Les méthodes directes sont sensibles au biais de modèle et à la rareté des observations par action. Les méthodes IPS sont non biaisées sous des hypothèses de support appropriées, mais leur variance peut exploser lorsque la politique cible attribue de la masse à des actions peu probables sous la politique de logging. En outre, la thèse montre que les objectifs IPS induisent souvent des paysages non concaves, plats et riches en maxima locaux lorsque l'espace d'actions est grand.

La première contribution hors politique est la *structured Direct Method*, ou `sDM`. Elle transpose au cadre offline l'idée de structuration bayésienne introduite dans `meTS`. Au lieu d'estimer indépendamment un paramètre par action, `sDM` couple les paramètres θ_a à travers un vecteur latent partagé ψ. Après observation du jeu de données journalisé, l'algorithme calcule le postérieur de ψ, puis les postérieurs conditionnels des paramètres d'actions. La récompense estimée pour chaque action est obtenue en intégrant l'incertitude postérieure, et la politique apprise agit ensuite gloutonnement par rapport à cette récompense moyenne postérieure.

L'analyse introduit une notion de sous-optimalité bayésienne adaptée au cadre hors politique. Elle montre que `sDM` atteint une convergence en $\mathcal{O}(1/\sqrt{n})$ sans imposer l'hypothèse de support uniforme complet $\pi_0(a \mid x) \geq \gamma > 0$ pour toutes les actions. La borne dépend plutôt de l'alignement entre la politique de logging et la politique optimale : plus les actions optimales sont couvertes par les données, plus l'apprentissage est efficace. Cette analyse révèle également un phénomène important : sous le critère bayésien considéré, les politiques gloutonnes sont optimales et peuvent surpasser les politiques pessimistes, contrairement au cadre fréquentiste où le pessimisme est souvent nécessaire pour se protéger contre les pires cas. Les expériences confirment que `sDM` améliore les méthodes directes standards, avec des gains croissants lorsque K augmente.

La contribution suivante remet en question une hypothèse centrale de l'apprentissage hors politique : l'idée selon laquelle l'amélioration des estimateurs de valeur suffit à améliorer l'apprentissage de politiques. La thèse montre que, dans les grands espaces d'actions, l'erreur d'optimisation peut dominer l'erreur d'estimation. Même un estimateur statistiquement sophistiqué peut conduire à une mauvaise politique si l'objectif qu'il induit est difficile à optimiser. L'analyse des paysages d'optimisation montre que les objectifs fondés sur des estimateurs peuvent présenter des plateaux où les méthodes de gradient restent bloquées pendant $\mathcal{O}(K)$ itérations, ainsi qu'un nombre exponentiel de maxima locaux en fonction de K.

Pour comprendre ces échecs, la thèse analyse les politiques oracle associées à différents estimateurs, c'est-à-dire les politiques qui maximiseraient ces estimateurs avec une quantité infinie de données. Cette analyse met en évidence que chaque estimateur impose un biais inductif spécifique : IPS favorise les politiques proches du support de logging, tandis que des méthodes clusterisées opèrent au niveau de groupes d'actions. Ces observations motivent des paramétrisations de politiques adaptées à l'objectif, qui réduisent l'espace de recherche effectif de K à une taille beaucoup plus petite, comme la taille du support de logging ou le nombre de clusters.

La conclusion principale de cette partie est toutefois plus radicale : il peut être préférable

d'abandonner l'estimation explicite de valeur pour optimiser directement des objectifs de vraisemblance pondérée par la politique. La thèse introduit les objectifs *policy-weighted log-likelihood*, de la forme

$$\hat{U}_g(\pi) = \frac{1}{n} \sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i)) \log \pi(A_i \mid X_i),$$

où g est une fonction de pondération positive. Ces objectifs ne sont pas des estimateurs de valeur, mais ils possèdent des paysages d'optimisation bien plus favorables : pour des politiques softmax linéaires, ils sont concaves, et deviennent fortement concaves avec régularisation ℓ_2. Ils admettent donc un optimum global unique accessible par optimisation stochastique standard. Les expériences à très grande échelle, incluant des espaces allant jusqu'à un million d'actions, montrent que ces objectifs simples et stables surpassent des méthodes hors politique fondées sur des estimateurs de valeur plus complexes. Cette contribution établit que, pour l'apprentissage de politiques dans les grands espaces d'actions, l'optimisabilité de l'objectif est un critère aussi fondamental que sa précision statistique.

La dernière contribution principale de la thèse améliore les méthodes IPS régularisées en introduisant un pessimisme praticable et différentiable. Les ratios d'importance peuvent être très grands, ce qui augmente fortement la variance. Pour y remédier, la thèse étudie des estimateurs par *exponential smoothing*, notamment

$$\hat{V}^{\alpha}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)^{\alpha}} R_i,$$

et

$$\tilde{V}^{\beta}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} \right)^{\beta} R_i.$$

Ces estimateurs interpolent entre absence de régularisation et forte réduction de variance, tout en restant différentiables et donc compatibles avec l'optimisation par gradient.

Pour apprendre de manière sûre avec ces estimateurs biaisés mais moins variables, la thèse dérive une borne PAC-bayésienne bilatérale contrôlant l'écart entre la valeur vraie et l'estimateur régularisé. Cette borne décompose l'erreur en plusieurs termes interprétables : une divergence entre la politique apprise et la politique de logging, un biais dû à la régularisation des poids, et une variance résiduelle. Elle conduit à un objectif pessimiste qui maximise une borne inférieure empirique de la valeur :

$$\hat{V}_n^{\alpha}(\pi_\theta) - \text{pénalités de divergence} - \text{biais de régularisation} - \text{variance résiduelle}.$$

L'intérêt crucial de cette formulation est que tous les termes sont empiriques et différentiables. Contrairement à de nombreuses approches pessimistes dont les constantes théoriques sont inexploitables en pratique, cet objectif peut être optimisé à grande échelle par ascension de gradient stochastique. La thèse propose également un cadre PAC-bayésien unifié couvrant plusieurs familles de régularisation des poids d'importance, comme le clipping, l'exponential smoothing et l'implicit exploration. Ce cadre permet une comparaison cohérente de différentes formes de pessimisme et fournit des objectifs pratiques pour l'apprentissage offline sécurisé.

Enfin, la thèse présente plusieurs contributions additionnelles liées aux thèmes principaux. Dans le cadre en ligne, les idées de structuration bayésienne sont étendues au problème de *best-arm identification* à budget fixé. L'algorithme `PI-BAI` utilise l'information a priori pour répartir efficacement le budget d'exploration dans des bandits structurés. L'analyse fournit des garanties bayésiennes dépendant du prior sur la probabilité d'erreur, et montre que des allocations non adaptatives bien informées peuvent surpasser des stratégies adaptatives classiques. Dans le cadre hors politique, la thèse contribue également au développement du *logarithmic smoothing*, un estimateur pessimiste de la forme

$$\hat{V}_{\mathrm{LS}}^{\lambda}(\pi) = \frac{1}{n\lambda} \sum_{i=1}^{n} \log \left(1 + \lambda \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} R_i \right).$$

Cet estimateur agit comme une alternative douce et différentiable au clipping, bénéficie de garanties de concentration serrées, et permet d'obtenir des bornes de sous-optimalité plus fines que les approches précédentes. Enfin, cette thèse CIFRE maintient un lien constant avec les systèmes de recommandation industriels à grande échelle. Les applications développées autour de la recommandation optimisant la récompense, de l'évaluation offline, de la simulation contrefactuelle et des modèles de recommandation par ardoise ont fourni à la fois un terrain expérimental et une source de questions théoriques.

Dans son ensemble, cette thèse défend une idée centrale : pour apprendre efficacement dans de grands espaces d'actions, il ne suffit pas d'appliquer directement les algorithmes classiques de bandits contextuels ou d'apprentissage hors politique. Il faut exploiter la structure entre actions, contrôler explicitement l'incertitude et la couverture des données, et concevoir des objectifs dont le paysage d'optimisation reste favorable à grande échelle. Les contributions proposées répondent à ces exigences selon deux axes complémentaires. En apprentissage en ligne, des modèles bayésiens hiérarchiques et diffusionnels permettent de partager l'information entre actions et de réduire le regret. En apprentissage hors politique, des méthodes directes structurées, des objectifs de vraisemblance pondérée et des principes de pessimisme différentiable rendent l'apprentissage statistiquement robuste et computationnellement réalisable. Ces résultats contribuent à rapprocher la théorie des bandits contextuels des contraintes réelles des systèmes interactifs modernes, où les décisions doivent être prises parmi des catalogues massifs, à partir de signaux partiels, bruités et parfois fortement biaisés.

Chapter 1

Overview

1.1 Context and Scope

Interactive machine learning systems are a cornerstone of modern technology, optimizing decision-making in applications ranging from recommender systems and financial markets to robotics. These systems operate in a sequential loop: they process contextual information, select an action from a wide range of possibilities, and receive feedback that depends on both the context and the chosen action. A fundamental challenge in designing these systems is the scale of the decision space; in many real-world settings, the number of potential actions can be huge.

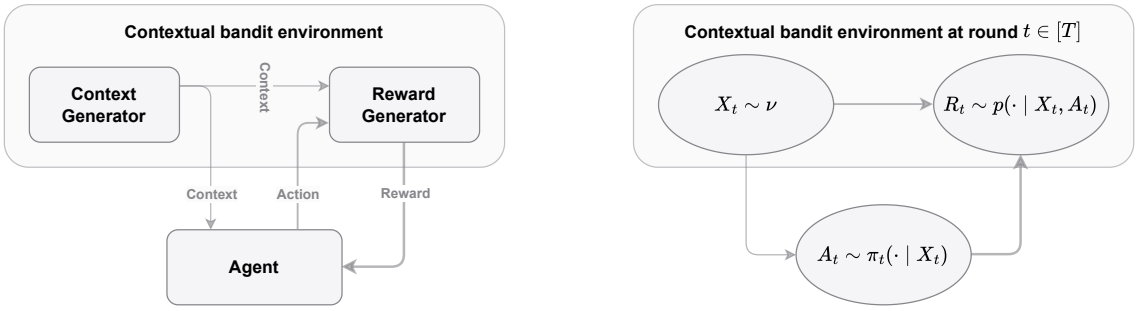
Consider *recommender systems*, where streaming platforms or e-commerce sites must select an item to present to a user. The *context* comprises rich data, such as user preferences and history. The *action* is the selection of a specific item from a catalog containing thousands or millions of options. The system's objective is to learn a policy that maximizes user engagement (the *reward*), measured by metrics such as watch time or clicks.

Similarly, in *computational advertising*, a platform selects which ad to display via real-time bidding. The context includes user attributes and page details. The action is composite: selecting an ad from a large inventory and determining a bid price. The feedback (reward) arrives in stages, from winning the auction to subsequent user clicks or conversions. The system must maximize advertiser value while adhering to budget constraints.

We model these interactive systems using the *contextual bandit* framework. This framework captures the essential characteristics of interactive learning while maintaining the tractability required for theoretical analysis and practical implementation.

1.1.1 Contextual Bandits

Figure 1.1a visualizes the interaction loop. The environment consists of a *Context Generator* and a *Reward Generator*, both of which are *assumed to be fixed but unknown to the agent*. At each round, the environment emits a context. The agent observes this con-



(a) Single interaction loop.

(b) Graphical representation in round $t \in [T]$.

Figure 1.1: Contextual bandit framework.

text and selects an action from a finite set[1]. Finally, the environment generates a scalar reward based on the context-action pair. By repeating this loop, the agent accumulates experience to refine its policy.

Formally, let $\mathcal{X} \subset \mathbb{R}^d$ denote the context space and $\mathcal{A} = [K]$ the finite action set. A *stochastic policy* $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ maps each context $x \in \mathcal{X}$ to a probability distribution $\pi(\cdot \mid x)$ over actions. The environment is specified by:

- A context distribution ν over $\mathcal{X}$;
- A family of conditional reward distributions $\{p(\cdot \mid x, a)\}_{(x,a)\in\mathcal{X}\times\mathcal{A}}$.

The expected reward function for any context-action pair (x, a) is defined as:

$$r(x, a) = \mathbb{E}_{R\sim p(\cdot|x,a)}[R]. \tag{1.1}$$

The interaction unfolds over T rounds. In each round $t \in [T]$:

1. The environment draws a context $X_t \sim \nu$ and reveals it to the agent.
2. The agent selects an action $A_t \sim \pi_t(\cdot \mid X_t)$ according to its current policy π_t.
3. The environment samples a reward $R_t \sim p(\cdot \mid X_t, A_t)$ and returns it to the agent.

A graphical representation of this interaction in round $t \in [T]$ is visualized in Figure 1.1b.

1.1.2 Learning Paradigms

We address learning in this framework via two complementary paradigms: *on-policy* (online) and *off-policy* (offline). With a slight abuse of terminology, we use *learning* loosely to encompass the full agent behavior, including both reward estimation and action selection.

On-Policy (Online) Learning

In this setting, the agent *updates* its policy π_t sequentially. Let $H_t = \{(X_\ell, A_\ell, R_\ell)\}_{\ell<t}$ denote the history available at the start of round t. The agent uses H_t to construct

[1]The action space can technically be infinite, but this thesis focuses on large finite action spaces.

the policy π_t. Following the interaction, the agent augments the history with the new observation (X_t, A_t, R_t) to form H_{t+1} and repeats the process.

Performance is measured by the *cumulative regret*:

$$\mathcal{R}(T) = \sum_{t=1}^{T} \left(r(X_t, A_{t,*}) - r(X_t, A_t) \right), \tag{1.2}$$

where $A_{t,*} = \arg\max_{a \in \mathcal{A}} r(X_t, a)$ is the optimal action in round t. While minimizing regret is equivalent to maximizing cumulative reward, the literature prioritizes regret as it normalizes performance against the optimal oracle, facilitating theoretical comparisons across environments.

The agent faces the *exploration-exploitation dilemma*: it must balance exploration of poorly understood actions with exploitation of actions believed to yield high rewards. Feedback is partial (only the reward for the chosen action is observed) and noisy, and exploration may be constrained by safety or budget requirements.

Remark 1 (Beyond regret minimization)**.** *While this thesis focuses on regret minimization, other objectives exist, most notably* Best-Arm Identification (BAI)*. BAI aims to identify the optimal action rather than maximize cumulative reward. This is important for applications like A/B testing and clinical trials. Although we focus on regret, our core modeling contributions for scaling Thompson sampling extend to the BAI setting, as demonstrated in our related work (Nguyen et al., 2025) (not included in this manuscript).*

Off-Policy (Offline) Learning

In this setting, the agent learns a policy $\hat{\pi}$ from a static logged dataset $\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$ collected by a logging policy π_0 as $X_i \sim \nu$, $A_i \sim \pi_0(\cdot \mid X_i)$ and $R_i \sim p(\cdot \mid X_i, A_i)$. No additional interactions with the environment are allowed. The objective is to find a policy maximizing the expected value:

$$V(\pi) = \mathbb{E}_{X \sim \nu} \mathbb{E}_{A \sim \pi(\cdot|X)}[r(X, A)]. \tag{1.3}$$

Performance is measured by the *suboptimality gap* of the learned policy $\hat{\pi}$:

$$\text{SO}(\hat{\pi}) = V(\pi_*) - V(\hat{\pi}), \tag{1.4}$$

where $\pi_* = \text{argmax}_{\pi \in \Pi} V(\pi)$ is the *unknown* optimal policy in a class of policies Π.

Since the agent learns solely from data generated by π_0, it must perform *counterfactual reasoning*. This introduces several challenges. First, *support mismatch* occurs when specific actions are rarely or never selected by π_0 in certain regions of the context space; consequently, the dataset provides little to no information about the rewards in these regions. Second, *re-weighting instability* arises since standard techniques, such as *inverse propensity scoring* (importance sampling), re-weight logged samples using the density ratio $\pi(a \mid x)/\pi_0(a \mid x)$. When $\pi_0(a \mid x)$ is small, this ratio

can explode, leading to high-variance estimates. Third, *high bias* can affect methods that rely on parametric models to estimate the reward. Extrapolating rewards for unobserved context-action pairs introduces errors if the model is misspecified.

The difficulties in both paradigms are amplified when the action space is large (K in the thousands or millions). In on-policy settings, independent exploration of every action becomes infeasible, yielding high regret. In off-policy settings, data sparsity worsens: reward models risk significant extrapolation error, and importance weights suffer from extreme variance as the probability of observing any specific action vanishes. Developing methods that scale gracefully (both statistically and computationally) with the size of the action space is the central theme of this thesis.

1.2 Background

Most on-policy and off-policy learning algorithms[2] are fundamentally constructed from two core components: *reward estimation*, which approximates the expected reward function $r(x, a)$ for any context-action pair from data, and *decision-making*, which leverages these estimates and their associated uncertainty to select actions.

1.2.1 Reward Estimation

The central task in reward estimation is to learn an approximate function $\hat{r}(x, a)$ that estimates the true expected reward $r(x, a)$ defined in Equation (1.1). This function is trained on a dataset, denoted by DATA, whose structure depends on the learning paradigm:

On-policy data. The agent collects data sequentially. The dataset at round t is the history of interaction up to round t:

$$\text{DATA} = H_t = \{(X_i, A_i, R_i)\}_{i=1}^{t-1}. \tag{1.5}$$

Off-policy data. The agent learns from a static dataset logged by a logging policy π_0:

$$\text{DATA} = \mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^{n}, \quad \text{with } A_i \sim \pi_0(\cdot \mid X_i). \tag{1.6}$$

We denote the learned model as $\hat{r}(x, a) = r(x, a; \hat{\theta})$, where $\hat{\theta}$ are parameters obtained via a statistical objective. Common approaches include:

Maximum Likelihood Estimation (MLE)

MLE seeks parameters θ that maximize the probability of observing the collected rewards. The objective is:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{(X_i, A_i, R_i) \in \text{DATA}} \log p(R_i \mid X_i, A_i; \theta).$$

For Gaussian rewards $p(\cdot \mid x, a; \theta) = \mathcal{N}(r(x, a; \theta), \sigma^2)$, this simplifies to minimizing the sum of squared errors (*ordinary least squares*). For Bernoulli rewards, it corresponds to minimizing binary cross-entropy (e.g., logistic regression).

[2]Recall that we use *learning* loosely to encompass both reward estimation and action selection (Section 1.1.2).

Maximum A Posteriori (MAP)

MAP estimation incorporates a prior distribution $p_0(\theta)$ to regularize the objective:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \left(\sum_{(X_i, A_i, R_i) \in \text{DATA}} \log p(R_i \mid X_i, A_i; \theta) + \log p_0(\theta) \right).$$

For example, combining a Gaussian likelihood with a zero-mean Gaussian prior $p_0(\cdot) = \mathcal{N}(0, \lambda I_d)$ is equivalent to *ridge regression* (ℓ_2 regularization).

Full Bayesian Inference

Rather than a point estimate, Bayesian inference characterizes uncertainty by computing the full posterior distribution $p(\theta \mid \text{DATA})$:

$$p(\theta \mid \text{DATA}) \propto p(\theta) \prod_{(X_i, A_i, R_i) \in \text{DATA}} p(R_i \mid X_i, A_i; \theta).$$

This approach is powerful for guiding exploration. A common implementation assumes conjugate Gaussian distributions: $p_0(\cdot) = \mathcal{N}(\mu_0, \Sigma_0)$ and $p(\cdot \mid x, a; \theta) = \mathcal{N}(r(x, a; \theta), \sigma^2)$. When $r(x, a; \theta)$ is linear in θ (*Bayesian linear regression*), the posterior is Gaussian and analytically tractable. For non-linear models, posterior approximation methods are required.

The functional form of the reward model, $r(x, a; \theta)$, is an important design choice that dictates the balance between computational tractability, data efficiency, and expressive power. While numerous function classes exist, linear models remain a cornerstone of the field due to their simplicity and strong theoretical guarantees. Even within this linear family, there is an important distinction between *joint* and *disjoint* formulations.

The *joint linear model* defines $r(x, a; \theta) = \phi(x, a)^\top \theta$, sharing a single parameter $\theta \in \mathbb{R}^d$ across all actions. While data-efficient, this approach relies on designing a feature map $\phi(x, a)$ capable of capturing complex context-action interactions. Designing $\phi(x, a)$ is very hard in practice, and inadequate feature engineering leads to poor performance. This is why in practice, it is often more common to adopt a *disjoint linear model* that learns an independent parameter $\theta_a \in \mathbb{R}^d$ for each action, yielding $r(x, a; \theta) = \phi(x)^\top \theta_a$. This is common in recommender systems, where predictions are inner products of user and item embeddings. This formulation is robust and avoids complex feature engineering. However, its primary drawback is *poor statistical scalability*: while the computational overhead of maintaining K independent embeddings can be addressed with sufficient compute, the statistical challenge remains. This is because learning each embedding independently requires substantial data per action, which becomes prohibitive as K grows.

Other non-linear models exist for capturing complex reward functions. Generalized linear models (GLMs), for instance, employ a link function to model non-Gaussian rewards (e.g., logistic regression for binary outcomes: $p(R = 1 \mid x, a) = \sigma(\phi(x)^\top \theta_a)$). These models align closely with the linear setting, and the algorithms proposed in this thesis are suitable for them. A significant portion of this thesis focuses on scaling these disjoint reward models to large action spaces.

Remark 2 (Scope of parametric reward modeling)**.** *The reward estimation framework presented above assumes parametric reward models $r(x, a; \theta)$. This assumption underlies Chapters 3, 4 and 6, where we develop structured parametric models that share information across actions to improve statistical efficiency. The remaining two chapters (Chapters 7 and 8) take a slightly different approach: rather than explicitly estimating rewards, they optimize policies directly using inverse propensity scoring and policy-weighted objectives, thereby making them agnostic to the choice of reward model.*

1.2.2 Decision-Making

We now turn to the second component: decision-making. This stage (often) relies on the reward model $\hat{r}$ derived using the estimation techniques discussed previously. While the difference in reward estimation between on-policy and off-policy settings is primarily driven by how data is accumulated, the principles guiding action selection in these two paradigms are fundamentally distinct.

Decision-Making in On-Policy Learning

Recall that in the on-policy setting, the agent must balance exploration and exploitation to minimize regret. The two dominant paradigms are *Upper Confidence Bound (UCB)* and *Thompson Sampling (TS)*.

Upper Confidence Bound (UCB)

UCB drives exploration via a bonus term added to the reward estimate, selecting:

$$A_t = \arg\max_{a \in \mathcal{A}} \left(\hat{r}(X_t, a) + \text{bonus}_t(X_t, a) \right).$$

For linear models (e.g., `LinUCB`), the bonus scales with $\sqrt{\phi(x)^\top V_{t,a}^{-1} \phi(x)}$, where $V_{t,a}$ is the design matrix for action a, encouraging the selection of less-certain actions.

Thompson Sampling (TS)

TS (or posterior sampling) implements randomized exploration. The agent samples a reward function $\tilde{r}_t$ from the posterior and acts greedily with respect to it:

$$A_t = \arg\max_{a \in \mathcal{A}} \tilde{r}_t(X_t, a), \quad \text{where } \tilde{r}_t(x, a) = r(x, a; \theta_t),\ \theta_t \sim p(\theta \mid H_t).$$

Exploration is implicit: high posterior uncertainty yields diverse samples θ_t, leading to varied actions. As data accumulates, the posterior contracts, and behavior naturally becomes exploitative[3].

In large action spaces, standard UCB and TS struggle. Treating actions independently gathers information too slowly, leading to prohibitive regret. This necessitates structured models that share information across actions.

[3]While we describe sampling parameters θ_t, the general principle involves sampling from the posterior predictive distribution of rewards.

Decision-Making in Off-Policy Learning

Recall that the off-policy setting requires identifying an optimal policy from a static dataset collected under a distinct logging policy. The two dominant philosophies for tackling this are *Greedy Policies* and *Pessimistic Policies*.

Greedy Policies

Greedy methods select the policy $\hat{\pi}_{\text{G}}$ that maximizes a point estimate of value, $\hat{V}(\pi)$:

$$\hat{\pi}_{\text{G}} = \arg\max_{\pi \in \Pi} \hat{V}(\pi). \tag{1.7}$$

We distinguish between two primary classes of estimators. The *direct method (DM)* relies on the reward model $\hat{r}$ derived in the previous section. In contrast, *inverse propensity scoring (IPS)* bypasses reward modeling to estimate the policy value directly using importance weighting[4]:

$$\hat{V}_{\text{DM}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \pi(a \mid X_i)\, \hat{r}(X_i, a),$$
$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)}\, R_i.$$

The optimization procedures for these objectives differ significantly. The policy maximizing $\hat{V}_{\text{DM}}(\pi)$ is simply the one that acts greedily with respect to the learned reward model,

$$\hat{\pi}_{\text{G}}^{\text{DM}}(x) = \arg\max_{a \in \mathcal{A}} \hat{r}(x, a)\,. \tag{1.8}$$

For IPS, the optimization is typically performed numerically over a class of parameterized policies π_θ:

$$\hat{\pi}_{\text{G}}^{\text{IPS}} = \arg\max_{\theta \in \mathbb{R}^d} \hat{V}_{\text{IPS}}(\pi_\theta)\,. \tag{1.9}$$

Greedy policies are effective when the underlying estimator is accurate. Specifically, when the reward model $\hat{r}$ is well-specified (for DM), or the importance weights have low variance (for IPS). However, the arg max operator can amplify estimation errors, leading to a policy that over-exploits optimistic model inaccuracies or high-variance weight estimates.

Pessimistic Policies

Pessimistic methods mitigate error amplification by penalizing the objective with a quantified uncertainty term:

$$\hat{\pi}_{\text{P}} = \arg\max_{\pi \in \Pi} \left[\hat{V}(\pi) - \text{pen}(\pi) \right]. \tag{1.10}$$

[4]Importance weighting allows IPS to be unbiased under the *common support* assumption (i.e., $\pi_0(a|x) = 0 \implies \pi(a|x) = 0$)

This principle applies to both DM and IPS as:

$$\texttt{Pess-DM:} \qquad \hat{\pi}_{\mathrm{P}}^{\mathrm{DM}}(x) = \arg\max_{a \in \mathcal{A}} \Big[\hat{r}(x,a) - \beta\, \hat{\sigma}_r(x,a) \Big], \tag{1.11}$$

and

$$\texttt{Pess-IPS:} \qquad \hat{\pi}_{\mathrm{P}}^{\mathrm{IPS}} = \arg\max_{\theta} \Big[\hat{V}_{\mathrm{IPS}}(\pi_\theta) - \beta\, \hat{\sigma}_{\mathrm{IPS}}(\pi_\theta) \Big]. \tag{1.12}$$

Here, $\hat{\sigma}_r$ captures the uncertainty of the reward model, while $\hat{\sigma}_{\mathrm{IPS}}$ captures the uncertainty of the IPS estimator itself (e.g., its variance). The penalty term $\mathrm{pen}(\pi)$ prevents the maximization operator from selecting overestimated policies. It also regulates *distributional shift* by penalizing policies that place probability mass on context-action pairs with low coverage under π_0 (support mismatch).

Standard methods, whether relying on DM or IPS, and whether adopting greedy or pessimistic policies, face severe limitations in large action spaces. For DM, the prevailing practice of modeling action parameters independently prevents information sharing, making learning statistically inefficient. For IPS, the well-recognized issue is *variance*: importance weights can be large. However, we also demonstrate in this thesis that *optimization* can be an even greater bottleneck in large action spaces. This is because standard IPS-based objectives (whether Equation (1.9) or Equation (1.12)) induce highly non-concave landscapes with flat plateaus that trap gradient-based optimizers. Finally, for pessimistic methods, existing formulations often rely on intractable bounds that are incompatible with modern stochastic optimization techniques. This thesis addresses these specific pathologies: we introduce structured models for DM to enforce information sharing; we propose new policy-weighted log-likelihood objectives that yield superior optimization landscapes compared with IPS-based objectives; and we develop variance-reduced, tractable pessimistic objectives that are amenable to stochastic optimization at scale.

1.3 Contributions

Part I: On-Policy Learning in Large Action Spaces

In the *on-policy setting*, exploration strategies that adopt disjoint reward models[5] and learn each action parameter independently gather information slowly, resulting in high regret and failure to converge to an optimal policy within practical time horizons. Part I addresses this challenge by scaling Thompson Sampling to large action spaces while retaining the disjoint reward model parameterization. Our primary contribution is the introduction of structured Bayesian models with informative priors that share statistical strength across actions, enabling efficient exploration without sacrificing the robustness and flexibility of disjoint reward models.

[5]Recall that disjoint models offer greater flexibility and are widely used in industrial settings such as large-scale recommender systems that use a separate embedding for each item, whereas joint reward models require careful feature engineering and are not, to the best of our knowledge, widely deployed in practical recommendation systems.

(Chapter 3) Scaling Thompson Sampling with Mixed-Effects

We propose a hierarchical Bayesian framework that couples action parameters through L shared latent effect parameters $\Psi = (\psi_\ell)_{\ell \in [L]} \in \mathbb{R}^{dL}$, where each effect ψ_ℓ can represent, for example, a category of items:

$$\begin{aligned} \Psi &\sim q_0\,, & \\ \theta_a \mid \Psi &\sim p_{0,a}(\cdot \mid \Psi)\,, & \forall a \in [K]\,, \\ R_t \mid \theta, \Psi, X_t, A_t &\sim p(\cdot \mid X_t; \theta_{A_t})\,, & \forall t \in [T]\,. \end{aligned}$$

Upon this model, we build *mixed-effect Thompson sampling* (`meTS`). `meTS` maintains a posterior over effects $q_t(\Psi) = p(\Psi \mid H_t)$ and K conditional posteriors over actions $p_{t,a}(\theta_a \mid \Psi) = p(\theta_a \mid \Psi, H_t)$. In round $t \in [T]$, parameters are sampled hierarchically:

$$\begin{aligned} \Psi_t &\sim q_t(\cdot)\,, & \\ \theta_{t,a} &\sim p_{t,a}(\cdot \mid \Psi_t)\,, & \forall a \in [K]\,. \end{aligned}$$

Actions are then selected via the standard TS rule: $A_t = \arg\max_{a \in [K]} r(X_t; \theta_{t,a})$. We derive exact closed-form updates for linear models and tractable Laplace approximations for generalized linear models.

Theoretically, we establish a Bayesian regret bound in the linear-Gaussian case:

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}\left(\sqrt{TdK_{\text{eff}}(\sigma_0^2 + \sigma_\Psi^2)}\right)\,,$$

where σ_0^2 and σ_Ψ^2 are the prior variances of the action and effect parameters, respectively, and K_{eff} is the *effective number of actions*. When $L \ll K$, we have $K_{\text{eff}} \ll K$, yielding a multiplicative Bayesian regret improvement of $\sqrt{K/K_{\text{eff}}}$ over standard TS. Computationally, `meTS` exploits the conditional independence of action parameters given the latent effects, reducing memory complexity from $\mathcal{O}(K^2d^2)$ to $\mathcal{O}((L^2 + K)d^2)$ and runtime from $\mathcal{O}(K^3d^3)$ to $\mathcal{O}((L^3 + K)d^3)$. Empirically, `meTS` consistently outperforms baselines, with gains increasing with K.

AISTATS 2023 (Poster) - Aouali et al. (2023b):

- I. Aouali, B. Kveton, and S. Katariya. Mixed-effect Thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2087–2115. PMLR, 2023b.

(Chapter 4) Scaling Thompson Sampling with Diffusion Models

This chapter extends the hierarchical framework of `meTS` by introducing *diffusion Thompson Sampling* (`dTS`), which replaces the single-layer prior with a deep hierarchy of latent variables governed by a *diffusion model*:

$$\begin{aligned} \psi_L &\sim \mathcal{N}(0, \Sigma_{L+1})\,, & \\ \psi_{\ell-1} \mid \psi_\ell &\sim \mathcal{N}(f_\ell(\psi_\ell), \Sigma_\ell)\,, & \forall \ell \in [L] \setminus \{1\}\,, \\ \theta_a \mid \psi_1 &\sim \mathcal{N}(f_1(\psi_1), \Sigma_1)\,, & \forall a \in [K]\,, \\ R_t \mid \theta, (\psi_\ell)_{\ell \in [L]}, X_t, A_t &\sim p(\cdot \mid X_t; \theta_{A_t})\,, & \forall t \in [T]\,. \end{aligned}$$

The link functions f_ℓ are pre-trained non-linear transformations (e.g., neural networks), enabling rich representations of inter-action structure. As in `meTS`, exploration proceeds by sampling parameters top-down through the hierarchy and selecting the reward-maximizing action.

The key technical challenge is that the exact posterior is intractable due to non-linearities in both the reward and link functions. To enable fast online updates without expensive MCMC, we derive a tractable inference procedure based on Gaussian-like updates[6]. The resulting posterior preserves the diffusion structure: it remains a hierarchy of conditional Gaussians, but with *fine-tuned link functions and precisions*:

$$\bar{\Sigma}_{t,\ell-1}^{-1} = \underbrace{\Sigma_\ell^{-1}}_{\text{prior precision}} + \underbrace{\bar{G}_{t,\ell-1}}_{\text{data precision}} ,$$
$$\hat{f}_{t,\ell}(\psi_\ell) = \bar{\Sigma}_{t,\ell-1}\Big(\underbrace{\Sigma_\ell^{-1} f_\ell(\psi_\ell)}_{\text{prior contribution}} + \underbrace{\bar{B}_{t,\ell-1}}_{\text{data contribution}} \Big) .$$

Here, $\bar{G}_{t,\ell-1}$ and $\bar{B}_{t,\ell-1}$ are sufficient statistics propagated upward through the hierarchy. As data accumulates, covariances contract and means shift from prior toward MLE. Crucially, this formulation preserves the expressiveness of diffusion models since the posterior is not a single Gaussian but a *posterior diffusion model* that retains the generative structure of the prior.

Theoretically, we analyze `dTS` in the fully linear-Gaussian setting to gain analytical insight, deriving a Bayes regret bound:

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}\left(\sqrt{TdK_{\text{eff}} \sum_{\ell=1}^{L+1} \sigma_\ell^2} \right) ,$$

where $\Sigma_\ell = \sigma_\ell^2 I_d$ and $K_{\text{eff}} \ll K$ is the effective number of actions. Computationally, `dTS` exploits hierarchical conditional independence to reduce memory and time complexity further from $\mathcal{O}(K^2d^2)$ and $\mathcal{O}(K^3d^3)$ to $\mathcal{O}((L+K)d^2)$ and $\mathcal{O}((L+K)d^3)$, respectively: linear scaling in L that improves upon `meTS`. Empirically, `dTS` consistently outperforms baselines by leveraging pre-trained diffusion priors, even when these priors are imperfect or trained on limited data.

NeurIPS 2025 (Poster) - (Aouali, 2025, 2023):

- I. Aouali. Diffusion models meet contextual bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- I. Aouali. Linear diffusion models meet contextual bandits with large action spaces. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

[6]By Gaussian-like updates, we mean that the posterior precision is the sum of prior and evidence precisions, and the posterior mean is the precision-weighted combination of prior mean and maximum likelihood estimate.

Part II: Off-Policy Learning in Large Action Spaces

In the *off-policy setting*, both DM and IPS face severe limitations as the action space grows. DM suffers from high model bias and statistical inefficiency due to sparse data coverage. IPS exhibits high variance and potential bias due to insufficient support; moreover, as we demonstrate in this thesis, optimizing IPS-based objectives becomes intractable in large action spaces. Part II addresses these failure modes through three complementary approaches.

(Chapter 6) Scaling Direct Methods with Latent Parameters

Standard DMs estimate independent d-dimensional parameters for each action, which becomes statistically inefficient when actions are rarely observed. We introduce the *structured direct method* (`sDM`), which couples action parameters through a shared latent vector ψ (analogous to Chapter 3):

$$\begin{aligned} \psi &\sim q\,, \\ \theta_a \mid \psi &\sim p_a(\cdot; f_a(\psi)), \\ R \mid X, A, \theta, \psi &\sim p(\cdot \mid X; \theta_A). \end{aligned}$$

`sDM` computes the posterior over latent effects $p(\psi \mid \mathcal{D}_n)$ and conditional posteriors $p(\theta_a \mid \psi, \mathcal{D}_n)$. The marginal posterior $p(\theta_a \mid \mathcal{D}_n)$ is obtained by integrating out ψ, yielding the reward estimate $\hat{r}(x, a) = \mathbb{E}[r(x, a; \theta) \mid \mathcal{D}_n]$, which is used in a greedy policy as:

$$\hat{\pi}_{\text{G}}(a \mid x) = \mathbb{1}\{a = \arg\max_{b \in \mathcal{A}} \hat{r}(x, b)\}\,.$$

To analyze performance, we introduce *Bayesian suboptimality* (BSO) and prove that `sDM` achieves $\mathcal{O}(1/\sqrt{n})$ convergence. The result avoids the restrictive full support assumption, which requires $\pi_0(a \mid x) \geq \gamma > 0$ for all actions. Instead, the bound depends on the alignment between the logging policy π_0 and the optimal policy π_*: performance improves smoothly as coverage of optimal actions increases. We also prove that under BSO, greedy policies are optimal and outperform pessimistic ones. This contrasts with the frequentist setting, where pessimism hedges against worst-case scenarios and is generally preferred. Experiments on synthetic and real-world data confirm that `sDM` outperforms existing methods, with gains increasing with K.

AISTATS 2025 (Poster) - Aouali et al. (2025):

- I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Bayesian off-policy evaluation and learning for large action spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR, 2025.

(Chapter 7) Optimization Matters More Than Estimation

This chapter challenges a common paradigm in off-policy learning. The field has traditionally focused on developing sophisticated value estimators with improved statistical properties, assuming that maximizing a more accurate estimator yields a better policy. We demonstrate that this emphasis is misplaced in large action spaces, where *optimization error* dominates estimation error, making even advanced estimators ineffective for policy learning.

Our key insight is that estimator-based objectives, despite their statistical appeal, induce highly non-concave landscapes when paired with standard policy classes. We show that gradient-based optimization can remain trapped in suboptimal plateaus for $\mathcal{O}(K)$ iterations, and that the landscape contains exponentially many local maxima in K. These pathologies make global optimization intractable for large K.

To characterize these failures, we analyze the *oracle policies* of various estimators: the policies that maximize the estimators with infinite data. This analysis reveals that each estimator induces a distinct inductive bias. For instance, standard IPS searches within the logging policy's support, while cluster-based methods such as `MIPS` (Saito and Joachims, 2022) operate at the cluster level. These insights motivate *objective-aware policy parametrizations*: by aligning the policy parametrization with the estimator's bias, we reduce the effective search space from K to the significantly smaller logging support size k_0 or cluster count C, partially alleviating the optimization challenges.

Ultimately, we advocate for a fundamental shift: abandoning value estimation in favor of *policy-weighted log-likelihood (PWLL)* objectives:

$$\hat{U}_g(\pi) = \frac{1}{n} \sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i)) \log \pi(A_i \mid X_i) \,,$$

where g is a positive weighting function. Although PWLL objectives are not value estimators, we prove they are concave (and strongly concave with ℓ_2 regularization) for linear softmax policies, while achieving oracle policies comparable to estimator-based objectives. This guarantees efficient convergence to a unique global maximum, while eliminating the optimization pathologies.

Large-scale experiments on datasets with up to one million actions validate this approach. Simple PWLL methods consistently outperform state-of-the-art estimator-based objectives, with the performance gap widening as action spaces grow. Moreover, PWLL objectives exhibit remarkable robustness to optimization hyperparameters, whereas estimator-based methods require careful tuning and often fail under minor configuration changes.

CONSEQUENCES, RecSys 2025 (Poster) - Submitted to ICLR 2026 - (Aouali and Sakhi, 2025):

- I. Aouali and O. Sakhi. Off-policy learning in large action spaces: Optimization matters more than estimation. *Under review at ICLR*, 2026.

(Chapter 8) Principled Pessimism for Exponential Smoothing and Beyond

While `sDM` and PWLL offer alternative paradigms, this chapter improves the widely used family of IPS-based methods by combining variance-reducing regularization with principled pessimism for safe policy learning.

The variance of IPS scales with importance weights, which can explode in large action spaces. To control this, we introduce differentiable *exponential smoothing*

(ES) estimators that regularize these weights:

$$\texttt{IPS-}\alpha: \quad \hat{V}^{\alpha}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)^{\alpha}} R_i \,,$$

$$\texttt{IPS-}\beta: \quad \tilde{V}^{\beta}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)}\right)^{\beta} R_i \,.$$

These estimators smoothly trade variance for bias while preserving differentiability.

To learn safely with these regularized estimators, we derive a *two-sided PAC-Bayes generalization bound* that will be used in a pessimistic objective:

$$\left|V(\pi_\theta) - \hat{V}_n^{\alpha}(\pi_\theta)\right| \leq \sqrt{\frac{\mathrm{KL}_1(\theta,\theta_0)}{2n}} + \underbrace{B_n^{\alpha}(\pi_\theta)}_{\text{Regularization bias}} + \frac{\mathrm{KL}_2(\theta,\theta_0)}{n\lambda} + \frac{\lambda}{2}\underbrace{\mathrm{Var}_n^{\alpha}(\pi_\theta)}_{\text{Remaining variance}} \,,$$

where $\mathrm{KL}_{1,2}$ measure divergence between the current policy π_θ and the logging policy π_{θ_0}, B_n^{α} captures the regularization bias, and Var_n^{α} captures the remaining variance. The exact expressions of these quantities are given in Chapter 8. The pessimistic learning objective maximizes the lower bound as:

$$\hat{\pi} = \arg\max_{\pi_\theta} \left[\hat{V}_n^{\alpha}(\pi_\theta) - \sqrt{\frac{\mathrm{KL}_1(\theta,\theta_0)}{2n}} - B_n^{\alpha}(\pi_\theta) - \frac{\mathrm{KL}_2(\theta,\theta_0)}{n\lambda} - \frac{\lambda}{2}\mathrm{Var}_n^{\alpha}(\pi_\theta)\right] .$$

This objective penalizes policies with high bias or variance, steering optimization toward reliable regions. What is important for scalability is that all terms in the objective are empirical and differentiable, enabling end-to-end optimization via standard stochastic gradient ascent. This contrasts with prior pessimistic objectives that relied on intractable theoretical constants or were incompatible with stochastic optimization.

We further present a *unified PAC-Bayes framework* that generalizes this approach to all importance-weight regularizers in the literature (clipping, ES, implicit exploration), enabling fair comparison through a universal set of practical pessimistic objectives. This work also laid the foundation for *logarithmic smoothing* (see Additional Contributions below), which refines the analysis to achieve significantly tighter bounds and sharp suboptimality guarantees.

ICML 2023 (Oral) - UAI 2024 (Poster) - (Aouali et al., 2023a, 2024)

- I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Exponential smoothing for off-policy learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 984–1017. PMLR, 2023a.
- I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Unified PAC-Bayesian study of pessimism for offline policy learning with regularized importance sampling. In *Uncertainty in Artificial Intelligence*, pages 88–109. PMLR, 2024.

Additional Contributions

This section outlines additional research conducted during this thesis. While these contributions are not included in the main manuscript, they correspond to published works involving equal or significant contributions from the author.

On-Policy: Extension to Best-Arm Identification

We extend hierarchical and structured modeling from regret minimization to fixed-budget best-arm identification (BAI), introducing *prior-informed best-arm identification* (`PI-BAI`): a non-adaptive algorithm that leverages prior knowledge for efficient budget allocation.

We provide a fully Bayesian analysis for structured settings (e.g., linear and hierarchical bandits), departing from classical frequentist approaches. This yields the first prior-dependent guarantees on Bayesian error probability in fixed-budget BAI. `PI-BAI` is robust to prior misspecification and consistently outperforms baselines, including adaptive strategies, challenging the prevailing assumption that adaptivity is essential for fixed-budget exploration.

AISTATS 2025 (Poster) - (Nguyen et al., 2025):

- N. Nguyen, I. Aouali, A. György, and C. Vernade. Prior-dependent allocations for Bayesian fixed-budget best-arm identification in structured bandits. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

Off-Policy: Extension to Logarithmic Smoothing

We refine the theoretical framework for regularized IPS estimators from Chapter 8. While the bounds developed there were useful for learning in practice, they can be loose for suboptimality guarantees in certain cases. To address this, we derive a general high-order moment concentration bound for regularized estimators and identify the estimator that minimizes this bound. This analysis yields a novel pessimistic estimator, *logarithmic smoothing (LS)*:

$$\hat{V}_{\text{LS}}^{\lambda}(\pi) = \frac{1}{n\lambda} \sum_{i=1}^{n} \log\left(1 + \lambda \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} R_i\right).$$

Similar to ES, LS acts as a soft, differentiable alternative to clipping, concentrates at a sub-Gaussian rate, and achieves finite variance without requiring bounded importance weights. The resulting high-probability risk bound is provably tighter than state-of-the-art alternatives, enabling sharp suboptimality guarantees.

NeurIPS 2024 (Spotlight) - (Sakhi et al., 2024):

- O. Sakhi, I. Aouali, P. Alquier, and N. Chopin. Logarithmic Smoothing for Pessimistic Off-Policy Evaluation, Selection and Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Off-Policy: Applications to Large-Scale Recommender Systems

This CIFRE thesis maintained a continuous feedback loop between theory and practice. Our work on large-scale industrial recommender systems served a dual purpose: it provided a testing ground for the off-policy methods developed in this thesis, while the real-world challenges encountered in these systems directly motivated the theoretical questions addressed in the main chapters. This resulted in several workshop publications and tutorials shared with the community (Gilotte et al., 2025; Aouali et al., 2022a,b,c, 2021).

- A. Gilotte, O. Sakhi, I. Aouali, and B. Heymann. Offline contextual bandit with counterfactual sample identification. *arXiv preprint arXiv:2509.10520*, 2025.
- I. Aouali, A. Benhalloum, M. Bompaire, A. Ait Sidi Hammou, S. Ivanov, B. Heymann, D. Rohde, O. Sakhi, F. Vasile, and M. Vono. Reward optimizing recommendation using deep learning and fast maximum inner product search. In *proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4772–4773, 2022a..
- I. Aouali, A. Benhalloum, M. Bompaire, B. Heymann, O. Jeunen, D. Rohde, O. Sakhi, and F. Vasile. Offline evaluation of reward-optimizing recommender systems: The case of simulation. *arXiv preprint arXiv:2209.08642*, 2022b.
- I. Aouali, A. A. S. Hammou, O. Sakhi, D. Rohde, and F. Vasile. Probabilistic rank and reward: A scalable model for slate recommendation. *arXiv preprint arXiv:2208.06263*, 2022c.
- I. Aouali, S. Ivanov, M. Gartrell, D. Rohde, F. Vasile, V. Zaytsev, and D. Legrand. Combining reward and rank signals for slate recommendation. *arXiv preprint arXiv:2107.12455*, 2021.

1.4 Related Work

1.4.1 On-Policy Learning in Contextual Bandits

In the on-policy (online) setting (Slivkins, 2019; Lattimore and Szepesvari, 2019; Bubeck et al., 2012; Li et al., 2010; Chu et al., 2011), the agent must balance choosing actions that maximize current reward estimates (*exploitation*) with exploring other actions to improve these estimates (*exploration*). This trade-off is often addressed using upper confidence bounds (UCBs) (Auer et al., 2002) or Thompson sampling (TS) (Thompson, 1933).

Upper confidence bound (UCB) algorithms handle the exploration-exploitation trade-off by constructing high-probability confidence intervals around reward estimates and selecting the action with the largest upper bound (Auer et al., 2002; Chu et al., 2011; Abbasi-Yadkori et al., 2011; Dani et al., 2008). The intuition is *optimism under uncertainty*: poorly explored actions have wide confidence intervals and thus large upper bounds, which encourages exploration. Despite strong theoretical guarantees, UCB methods are often less practical than TS due to sensitivity to confidence-parameter tuning and lack of inherent randomization. Nevertheless, UCB remains a cornerstone of bandit theory and continues to inspire new exploration strategies. Part I focuses on TS, but the

hierarchical principles developed there naturally extend to UCB-based exploration.

Thompson sampling (TS) operates in a Bayesian framework: a prior and likelihood are specified, the agent samples rewards from the posterior at each round, and then chooses the action with the highest sampled reward. TS is randomized by construction, easy to implement, and exhibits strong empirical performance in both simulated and real-world problems (Russo and Van Roy, 2014; Chapelle and Li, 2012; Russo et al., 2018). It also enjoys strong theoretical guarantees, including optimal or near-optimal regret in a variety of models (Kaufmann et al., 2012; Agrawal and Goyal, 2013b; Korda et al., 2013; Russo and Van Roy, 2014; Agrawal and Goyal, 2017; Abeille and Lazaric, 2017; Russo and Van Roy, 2016; Lu and Van Roy, 2019). Part I advances TS by integrating informative hierarchical priors that enable efficient learning in large action spaces.

Hierarchical Bayesian bandits (Bastani et al., 2019; Kveton et al., 2021; Basu et al., 2021; Simchowitz et al., 2021; Wan et al., 2021; Hong et al., 2022b; Peleg et al., 2022; Wan et al., 2022; Tomkins et al., 2021; Urteaga and Wiggins, 2018) apply TS to simple graphical models in which action parameters are typically drawn from Gaussian distributions centered at a small number of latent parameters. These works primarily address meta- and multi-task learning in multi-armed bandits, transferring information across tasks or arms. Our mixed-effect Thompson sampling (Chapter 3) extends this line of work by introducing a hierarchical structure with multiple latent effect parameters in the contextual bandit setting. It also provides Bayes regret bounds and computational guarantees in the large action space regime. Our diffusion Thompson sampling (Chapter 4) further generalizes these approaches by replacing simple Gaussian hierarchies with deep, non-linear diffusion models that capture complex inter-action dependencies through flexible link functions f_ℓ.

Approximate Thompson sampling is a central challenge in Bayesian bandits because most posteriors are intractable and require approximate inference. Prior work (Riquelme et al., 2018; Chapelle and Li, 2012; Kveton et al., 2020) highlights the strong empirical performance of approximate TS in complex models. For mixed-effect TS (Chapter 3), we exploit the Gaussian structure of the hierarchy. We first apply a Laplace-like approximation to the reward likelihood at the action level, obtaining a Gaussian *pseudo-observation* on each parameter θ_a. Because both the priors on latent effects and on action parameters are Gaussian and the hierarchy is linear, these pseudo-observations can then be propagated exactly in closed form through the hierarchy, yielding an approximate posterior that preserves the original mixed-effect structure. For diffusion TS (Chapter 4), the prior hierarchy is defined by non-linear link functions f_ℓ, so Gaussian propagation is no longer exact. To retain a hierarchical diffusion model, we make an additional approximation: at each update we locally linearize the link-function updates. Combined with the Laplace-like approximation on the likelihood, this yields a chain of conditional Gaussians with updated means and precisions, i.e., a posterior diffusion model that preserves the prior hierarchy while remaining computationally tractable.

Bandits with underlying structure are closely related to our setting, where we assume structured relationships among actions. In latent bandits (Maillard and Mannor, 2014; Hong et al., 2020), a single latent variable indexes multiple candidate models. In structured finite-armed bandits (Lattimore and Munos, 2014; Gupta et al., 2018), each action

is linked to a known mean function parameterized by a common latent parameter that is learned online. TS has also been applied to more complex structures such as graphical and combinatorial bandits (Gopalan et al., 2014; Yu et al., 2020). However, these methods do not simultaneously guarantee computational and statistical efficiency in large action spaces. Meta- and multi-task learning with UCB-style methods also has a long history (Azar et al., 2013; Gentile et al., 2014; Deshmukh et al., 2017; Cella et al., 2020; Hu et al., 2021; Cella et al., 2022; Yang et al., 2020), but these works typically adopt a frequentist perspective, analyze stronger notions of regret, and often yield conservative algorithms. In contrast, our mixed-effect and diffusion TS algorithms are Bayesian, come with Bayes regret guarantees, and are explicitly designed to exploit pre-learned structure to achieve both statistical efficiency and scalable online inference.

Large action spaces. Our work directly addresses the challenge of learning with per-action parameters θ_a (disjoint reward models) rather than a single shared parameter θ (joint reward models). The disjoint formulation, while more expressive and widely used in practice, faces severe scalability issues that we address through hierarchical structure. Our analysis shows that both mixed-effect TS (Chapter 3) and diffusion TS (Chapter 4) achieve regret bounds that scale with an effective number of actions $K_{\text{eff}} \ll K$. The expression of K_{eff} depends however on K. Some prior works (Foster et al., 2020; Xu and Zeevi, 2020; Zhu et al., 2022) propose bandit algorithms whose regret is independent of K. However, their setting differs substantially from ours: they assume a reward function $r(x, a) = \phi(x, a)^\top \theta$ with a single shared parameter $\theta \in \mathbb{R}^d$ and a known mapping ϕ, whereas we consider $r(x, a) = \phi(x)^\top \theta_a$ (or simply $r(x, a) = x^\top \theta_a$) with K separate d-dimensional action parameters. The dependence on K in our setting reflects the inherent complexity of learning individual action parameters, which is the price paid for expressiveness. Obtaining a rich, known mapping ϕ that captures complex context-action dependencies can be challenging in practice, whereas our setting mirrors common scenarios such as recommender systems where each product has its own embedding learned from data. Note that both algorithms can be applied to the joint reward-model case; in that setting, our analysis would yield a K-independent regret bound.

1.4.2 Off-Policy Learning in Contextual Bandits

The challenges of large action spaces extend beyond on-policy learning to the equally important off-policy setting (Li et al., 2011; Bottou et al., 2013; Swaminathan and Joachims, 2015a), where decisions must be made using historical data collected under different policies. This section surveys the off-policy contextual bandit literature and positions the methods developed in Part II.

Off-policy learning fundamentally relies on *off-policy evaluation*, which estimates the value of a target policy π using data collected under a logging policy π_0. Given logged data $\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$ with $A_i \sim \pi_0(\cdot \mid X_i)$, off-policy evaluation seeks to estimate $V(\pi)$ without deploying π. Off-policy learning then optimizes over a policy class using this estimated value. Consequently, the prevailing paradigm is *estimator-centric*: first design an estimator $\hat{V}(\pi)$ with good statistical properties, then maximize it. As we show in Chapter 7, this estimate-then-optimize approach breaks down in large action spaces because optimization error, rather than estimation error, becomes the bottleneck.

Inverse propensity scoring (IPS) (Horvitz and Thompson, 1952; Dudík et al., 2012) corrects for distribution shift via importance weights:

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} R_i.$$

While unbiased under the *common support* assumption, IPS suffers from extremely high variance. It can also incur substantial bias when the logging policy has deficient support (Sachdeva et al., 2020), especially in large action spaces where the logging policy can only cover a small fraction of the actions. To mitigate variance, numerous importance-weight regularization techniques have been proposed, such as weight clipping (Ionides, 2008; Bottou et al., 2013) and others (Su et al., 2020; Metelli et al., 2021; Gabbianelli et al., 2024; Swaminathan and Joachims, 2015b; Gilotte et al., 2018). Chapter 8 introduces differentiable exponential-smoothing (ES) estimators that smoothly trade bias for variance and enable gradient-based optimization.

Direct methods (DM) (Jeunen and Goethals, 2021; Aouali et al., 2025) avoid importance weighting by modeling the expected reward for any context–action pair and evaluating policies using:

$$\hat{V}_{\text{DM}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi(a \mid X_i)\,\hat{r}(X_i, a).$$

DM is particularly attractive in large-scale recommender systems where IPS struggles due to its high variance (Sakhi et al., 2020; Jeunen and Goethals, 2021; Aouali et al., 2022c). However, standard implementations typically rely on a *disjoint model* that estimates one parameter vector $\theta_a \in \mathbb{R}^d$ per action. In large action spaces with sparse logging, this leads to severe statistical inefficiency: many actions are rarely observed and thus poorly estimated. Chapter 6 addresses this via the structured direct method (`sDM`), which leverages hierarchical Bayesian modeling to share statistical strength across actions.

Direct methods (DM) (Jeunen and Goethals, 2021; Aouali et al., 2025) build a model of the expected reward for any context–action pairs and evaluate policies using

$$\hat{V}_{\text{DM}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi(a \mid X_i)\,\hat{r}(X_i, a).$$

DM is particularly attractive in large-scale recommender systems, where IPS struggles (Sakhi et al., 2020; Jeunen and Goethals, 2021; Aouali et al., 2022c). Standard implementations, however, typically use a disjoint model that estimates one parameter vector θ_a per action. In large action spaces with sparse logging, this leads to severe statistical inefficiency: many actions are rarely observed and thus poorly estimated. Chapter 6 introduces the structured direct method (`sDM`), which addresses this limitation via hierarchical Bayesian modeling.

Doubly robust (DR) estimators (Robins and Rotnitzky, 1995; Bang and Robins, 2005; Dudík et al., 2011; Dudik et al., 2014; Farajtabar et al., 2018) combine DM and IPS to achieve robustness:

$$\hat{V}_{\text{DR}}(\pi) = \hat{V}_{\text{DM}}(\pi) + \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)}\big(R_i - \hat{r}(X_i, A_i)\big).$$

DR has become a default choice in many off-policy evaluation studies (Dudík et al., 2011; Dudik et al., 2014; Farajtabar et al., 2018; Su et al., 2020). Both of our contributions in Chapter 8 (regularized importance weights) and Chapter 6 (structured reward models) can be used to enhance the components of DR.

Large-scale IPS variants. Importance-weight regularization alone is often insufficient when the action space is very large. Structural assumptions can dramatically reduce the variance. For instance, marginalized IPS (MIPS) (Saito and Joachims, 2022) clusters actions via a mapping $h(a)$ and works with cluster-level importance weights:

$$\hat{V}_{\text{MIPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(h(A_i) \mid X_i)}{\pi_0(h(A_i) \mid X_i)} R_i.$$

This reduces variance by operating over a smaller cluster space instead of the full action set. This dimensionality reduction principle has inspired numerous extensions (Peng et al., 2023; Sachdeva et al., 2024; Cief et al., 2024; Taufiq et al., 2024; Saito et al., 2023). Our `sDM` (Chapter 6) provides a complementary structural approach designed for DM instead of IPS; it can be viewed as a Bayesian latent-structure counterpart to MIPS, replacing hard clustering with soft probabilistic coupling.

Pessimistic off-policy learning. Maximizing a point estimator (IPS, DM, or DR) can be unsafe when the estimator deviates from the true value. Pessimistic approaches instead construct lower confidence bounds on $V(\pi)$ and optimize those, following the principle of *pessimism in the face of uncertainty* (Jin et al., 2021). Asymptotic and finite-sample lower bounds have been developed for various estimators (Bottou et al., 2013; Kuzborskij et al., 2021; Gabbianelli et al., 2024), providing worst-case guarantees on policy performance. Many pessimistic learning methods are directly motivated by such bounds (Swaminathan and Joachims, 2015a; London and Sandler, 2019; Kuzborskij et al., 2021; Aouali et al., 2023a; Wang et al., 2023). For example, Swaminathan and Joachims (2015a) combine empirical-Bernstein inequalities with clipped IPS, leading to variance-penalized learning objectives.

Recently, the PAC-Bayesian paradigm (McAllester, 1998; Catoni, 2007; Alquier, 2021) has been increasingly applied to off-policy learning, offering a flexible toolkit for deriving data-dependent generalization bounds. London and Sandler (2019) introduced a scalable PAC-Bayesian perspective, which has been further developed by Flynn et al. (2023); Sakhi et al. (2022); Aouali et al. (2023a, 2024); Gabbianelli et al. (2024) to yield tight, directly optimizable bounds. Chapter 8 advances this direction by deriving a two-sided PAC-Bayes bound for exponentially smoothed IPS estimators, resulting in a fully differentiable pessimistic learning objective that jointly controls reward, bias, variance, and divergence from the logging policy. Moreover, this framework generalizes to encompass other estimators in the literature, establishing a unified set of principles for pessimistic learning. While our subsequent work on logarithmic smoothing (LS) (Sakhi et al., 2024) further tightens these guarantees, Chapter 8 lays the foundational theoretical groundwork on which LS builds.

Optimization-centric learning. All methods above follow an estimator-centric philosophy: find a good value estimator and maximize it. In large action spaces, these estimator-based objectives typically induce highly non-concave landscapes with flat plateaus and

many local maxima (Chapter 7), ensuring that optimization error dominates estimation error. Chapter 7 proposes a paradigm shift toward *optimization-centric* off-policy learning. Rather than insisting on high-quality estimators of the value, we advocate for *policy-weighted log-likelihood* objectives whose optimization landscape is benign, ensuring convergence to effective policies even in massive action spaces.

Part I

On-Policy Learning in Large Action Spaces

Chapter 2

Introduction to Part I

This first part of the thesis addresses the following fundamental question:

How can we design exploration-exploitation algorithms that remain both statistically efficient and computationally feasible when the number of actions is large?

2.1 Setting and Background

In this part, we consider the on-policy (online) contextual bandit setting where an agent interacts with an environment over T rounds. At each round $t \in [T]$:

1. The agent observes a context $X_t \in \mathcal{X} \subseteq \mathbb{R}^d$ drawn from a distribution ν;
2. The agent selects an action $A_t \in \mathcal{A} = [K]$ based on the history $H_t = \{(X_s, A_s, R_s)\}_{s=1}^{t-1}$;
3. The agent receives a stochastic reward $R_t \sim p(\cdot \mid X_t; \theta_{*,A_t})$.

Each action $a \in [K]$ is associated with an unknown true parameter $\theta_{*,a} \in \mathbb{R}^d$. The true expected reward is given by $r(x, a; \theta_*) = r(x; \theta_{*,a})$, where $\theta_* = (\theta_{*,a})_{a\in[K]} \in \mathbb{R}^{Kd}$ denotes the concatenation of all true action parameters.

Throughout this part, we assume the reward distribution is a generalized linear model (GLM) (McCullagh and Nelder, 1989). For any context $x \in \mathcal{X}$ and action $a \in \mathcal{A}$, $p(\cdot \mid x; \theta_{*,a})$ is an exponential-family distribution with mean $g(\phi(x)^\top \theta_{*,a})$, where g is the mean function. This formulation recovers linear bandits (Auer, 2002) when $p(\cdot \mid x; \theta_{*,a}) = \mathcal{N}(\cdot; \phi(x)^\top \theta_{*,a}, \sigma^2)$ (with identity link $g(u) = u$), and logistic bandits (Filippi et al., 2010) when $p(\cdot \mid x; \theta_{*,a}) = \mathrm{Ber}(g(\phi(x)^\top \theta_{*,a}))$ with the sigmoid link $g(u) = (1 + \exp(-u))^{-1}$.

We adopt a Bayesian perspective where the unknown true parameters θ_* are assumed to be drawn from a prior distribution p_0. Our objective is to minimize the *Bayes regret*:

$$\mathcal{BR}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \Big(r(X_t, A_{t,*}; \theta_*) - r(X_t, A_t; \theta_*)\Big)\right], \tag{2.1}$$

where $A_{t,*} = \arg\max_{a\in[K]} r(X_t, a; \theta_*)$ is the optimal action at round t. The expectation is taken over the prior p_0, the stochastic rewards, contexts, and the agent's policy.

2.1.1 Scalability Challenges

Standard *Thompson sampling (TS)* maintains a posterior distribution $p(\theta \mid H_t)$ over the parameter space and sampling $\theta_t \sim p(\cdot \mid H_t)$ at each round to select the action $A_t = \arg\max_{a \in [K]} r(X_t, a; \theta_t)$. However, when the number of actions K is large, there is trade-off between statistical and computational efficiency:

Statistical inefficiency of disjoint priors. A common simplification is to learn each action's parameter θ_a independently using a factorized prior $p_0(\theta) = \prod_{a=1}^{K} p_{0,a}(\theta_a)$. While this makes posterior updates computationally cheap, it prevents information sharing. The agent must learn about every action from scratch, which is prohibitive in large action spaces.

Computational intractability of joint priors. Conversely, modeling dependencies via a full joint posterior over $\mathbb{R}^{dK}$ allows for information sharing but is computationally intractable. Storing the covariance requires $\mathcal{O}(K^2 d^2)$ memory, and a single update requires $\mathcal{O}(K^3 d^3)$ time, making it intractable for online interaction.

2.2 Hierarchical Models

To address this dilemma, we propose a general hierarchical Bayesian framework where action parameters are coupled through a set of *latent parameters* $\Psi = (\psi_\ell)_{\ell \in [L]} \in \mathbb{R}^{Ld}$, with $L \ll K$. The generative process is defined as:

$$\Psi \sim q_0(\cdot) \qquad \text{(Prior over latent structure)}, \tag{2.2}$$

$$\theta_a \mid \Psi \sim p_{0,a}(\cdot \mid \Psi)\,, \qquad \forall a \in [K] \qquad \text{(Conditional prior per action)}, \tag{2.3}$$

$$R_t \mid \theta, \Psi, X_t, A_t \sim p(\cdot \mid X_t; \theta_{A_t})\,, \qquad \forall t \in [T] \qquad \text{(Reward observation)}. \tag{2.4}$$

Here, q_0 encodes global uncertainty, while $p_{0,i}$ specifies how individual actions deviate from the shared structure. The resulting marginal prior $p_0(\theta)$ naturally couples all action parameters. The corresponding posterior preserves this hierarchy:

$$p(\theta, \Psi \mid H_t) = p(\Psi \mid H_t) \prod_{a=1}^{K} p(\theta_a \mid \Psi, H_{t,a}). \tag{2.5}$$

The latent posterior $p(\Psi \mid H_t)$ aggregates evidence from all actions, enabling *global information sharing*, while the action posteriors $p(\theta_a \mid \Psi, H_{t,a})$ allow for efficient sampling of action parameters θ_a independently given Ψ. Thompson sampling then proceeds by first sampling the global structure Ψ_t, then sampling action parameters $\theta_{t,a}$ conditioned on Ψ_t.

2.3 Roadmap of Part I

The following chapters present two concrete instantiations of this hierarchical framework.

Chapter 3: Mixed-Effects Thompson Sampling. We begin by investigating a linear instantiation of the hierarchy where each action parameter is modeled as a linear combination of L shared effects, $\theta_a \mid \Psi \sim \mathcal{N}(\sum_{\ell=1}^{L} b_{a,\ell}\psi_\ell, \Sigma_{0,a})$, with known weights $b_{a,\ell}$.

We derive *mixed-effect Thompson sampling* (`meTS`), an algorithm that exploits the conjugacy of linear-Gaussian models to perform exact, closed-form posterior updates. For non-linear GLM rewards, we introduce a tractable Laplace approximation that we propagate through the hierarchy. We provide theoretical guarantees showing that `meTS` achieves a Bayes regret bound scaling with the effective number of actions $K_{\text{EFF}} \ll K$.

Chapter 4: Diffusion Thompson Sampling. We then extend the framework to support deep, non-linear hierarchical structures using *diffusion models*. In this setting, the priors form a Markov chain of latent variables $\psi_L \to \cdots \to \psi_1 \to \theta_a$, connected by potentially non-linear link functions f_ℓ (e.g., neural networks) learned from offline data. We propose *diffusion Thompson sampling* (`dTS`) and develop a posterior approximation that updates the link functions and covariances to match observed data while preserving the generative diffusion. This chapter demonstrates how to leverage powerful generative priors for exploration while remaining computationally feasible for online deployment.

Chapter 3

Scaling Thompson Sampling with Mixed Effects

Contents

This chapter begins with the fundamental observation that the expected rewards of actions in real-world problems are often correlated. To model this phenomenon, we study a structured mixed-effect bandit environment in which each *action parameter* depends on one or more *effect parameters* that are shared across actions. Therefore, taking an action teaches the agent about its effect parameters, thereby informing it about other actions that share the same effect parameters. We present three motivating examples for this.

Movie recommendation. Here, we want to recommend a movie to a user with the highest expected rating. User j and movie a are represented by vectors x_j (context) and θ_a (action parameter), respectively. The expected rating that user j gives to movie a is $x_j^\top \theta_a$. We assume that the vector x_j is observed. Then the most natural idea is to learn all θ_a individually using standard bandit methods (Li et al., 2010; Chu et al., 2011). This is statistically inefficient when the number of movies is high. Fortunately, the movies could be organized into L categories and such information can be leveraged to explore efficiently. We present three approaches **(A)**, **(B)** and **(C)** that do this next.

(A) For each category $\ell \in [L]$, a parameter ψ_ℓ is learned online using all interactions with the movies in category ℓ. The parameter ψ_ℓ represents all the movies in category ℓ and is used instead of their individual θ_a. Therefore, this approach has a high bias, as all movies in the same category are assumed to have the same expected rating. This issue can be addressed by a better model. **(B)** We model each movie parameter θ_a as a random variable centered in its category parameter ψ_ℓ. Now movies in the same category no longer have

the same expected rating due to the additional uncertainty. Both the category parameters ψ_ℓ and movie parameters θ_a are learned online. The former is learned using all interactions with the movies in category ℓ, while the latter is learned using all interactions with movie a conditioned on ψ_ℓ. The category parameter ψ_ℓ is learned using more data, which helps to learn θ_a more efficiently. This is a special case of our setting. **(C)** The shortcoming of **(B)** is that each movie belongs to a *single category*, which is unrealistic. To address this issue, we allow movies to belong to *multiple categories* and then proceed as in **(B)**. To connect with our terminology, the categories $\ell \in [L]$ denote effects; their parameters ψ_ℓ are the effect parameters; and the movie parameters θ_a are the action parameters.

Ad placement: Here, the agent selects a list (or *slate*) of M items from a catalog of L items with the objective of maximizing the click-through-rate. We assume that the agent receives only binary bandit feedback indicating whether the user clicked *one of the items in the slate* (Dimakopoulou et al., 2019; Rejwan and Mansour, 2020). Again, user j and slate a are represented by x_j (context) and θ_a (action parameter), respectively. The corresponding click-through-rate is $g(x_j^\top \theta_a)$, where g is the sigmoid function. The set of slates (of size $K \approx L^M$) is exponentially large, which makes learning θ_a individually difficult. Fortunately, the slates are related through a much smaller set of items (of size L). Therefore, slates containing common items can teach the agent about one another, enabling efficient exploration.

Efficient exploration is achieved by decomposing slate a's parameter as $\theta_a = \sum_{\ell \in [L]} b_{a,\ell} \psi_\ell + \epsilon_a$. Here $\psi_\ell \in \mathbb{R}^d$ is the parameter of item ℓ and $b_{a,\ell} \in \mathbb{R}$ is a mixing weight that captures position biases. That is, $b_{a,\ell} = 0$ if item ℓ is not in slate a, and $b_{a,\ell}$ is high if item ℓ is ranked high in slate a. This captures the fact that the probability of a click on an item is influenced by its position on the slate, and this bias can be estimated offline. Finally, ϵ_a is a random noise that can incorporate uncertainty due to *model misspecification*, for instance due to an estimation error of $b_{a,\ell}$. The benefit of this decomposition is that the parameter of item ℓ, ψ_ℓ, is learned using all interactions with the slates with item ℓ. The slate parameter θ_a is learned using all interactions with slate a conditioned on ψ_ℓ. This is more statistically efficient than learning θ_a individually, which only uses the interactions with slate a.

Drug design: Here, the goal is to find the optimal drug design in clinical trials (Durand et al., 2018). Subject j and drug a are represented by vectors x_j and θ_a, respectively, and the expected efficacy of drug a for subject j is $x_j^\top \theta_a$. Again, the most natural idea is to learn all drug parameters θ_a individually. This leads to statistical inefficiency when the number of candidate drugs is high. Fortunately, we can leverage the fact that drug candidates in the same trial often share components to explore efficiently. Precisely, a drug is a combination of multiple components, each with a specific dosage. Each component ℓ is represented by a parameter ψ_ℓ, and the drug parameter θ_a is a *known* combination of the component parameters ψ_ℓ weighted by their dosage. That is, $\theta_a = \sum_{\ell \in [L]} b_{a,\ell} \psi_\ell + \epsilon_a$, where $b_{a,\ell}$ is the dosage of component ℓ in drug a and ϵ_a is a random noise to incorporate uncertainty due to model misspecification. The efficacy of each component has an *effect* on the overall efficacy of the drug and is boosted by the dosage.

In all examples, we assume an underlying structure among the actions, that they are affected by multiple effects. In some problems, it is known how the effect arises. For

instance, in the drug design, the actions are the drugs and the effects are their components. The mixing weight that relates an action (drug) to an effect (component) is the dosage of that component in the drug. In other problems, it may not be apparent how the effect arises and this has to be learned. We discuss this in detail in Section 3.1.3.

We make the following contributions. **1)** We formalize a general mixed-effect bandit framework represented by a two-level graphical model where each action is associated with a d-dimensional parameter that depends on *one or multiple* effect parameters. **2)** We design mixed-effect Thompson sampling (`meTS`), which leverages this structure to be both statistically and computationally tractable. We show that closed-form posteriors can be derived for Gaussian instances and efficient approximations exist in more general cases. **3)** We prove that the Bayes regret of `meTS` is bounded by a sum of two terms: one is associated with learning the action parameters and the other quantifies the cost of learning the effect parameters. Both terms reflect the structure of the environment and the quality of priors. **4)** We show empirically that `meTS` and its variants perform extremely well, and are computationally efficient in both synthetic and real-world problems.

3.1 Setting

We consider the contextual bandit setting in Section 2.1. Each action $a \in \mathcal{A} = [K]$ is associated with an *unknown d-dimensional action parameter* $\theta_a \in \mathbb{R}^d$. The correlations between the action parameters arise because they are derived from L shared *unknown d-dimensional effect parameters*, $\psi_\ell \in \mathbb{R}^d$ for $\ell \in [L]$. Specifically, we assume that the action parameter θ_a is sampled from the *action prior distribution* $p_{0,a}$ as $\theta_a \mid \Psi \sim p_{0,a}(\cdot \mid \Psi)$, where $\Psi = (\psi_\ell)_{\ell \in [L]} \in \mathbb{R}^{Ld}$ is a concatenation of the effect parameters. The distribution $p_{0,a}$ can capture sparsity, when θ_a depends only on a subset of Ψ; and also incorporate uncertainty due to model misspecification, when θ_a is not a deterministic function of Ψ. Finally, the effect parameters Ψ are sampled from a *joint effect prior* q_0, which is known by the agent and represents its initial uncertainty about Ψ. In summary, all variables in our environment are generated as

$$\begin{aligned} \Psi &\sim q_0\,, && \\ \theta_a \mid \Psi &\sim p_{0,a}(\cdot \mid \Psi)\,, && \forall a \in \mathcal{A}\,, \\ R_t \mid X_t, A_t, \theta, \Psi &\sim p(\cdot \mid X_t; \theta_{A_t})\,, && \forall t \in [T]\,, \end{aligned} \tag{3.1}$$

where $p(\cdot \mid x; \theta_a)$ is the *reward distribution* of action a in context x, *which only depends on parameter θ_a and the context x*. The terminology of effect parameters arises from the fact that ψ_ℓ affect the model parameters θ_a, which in turn define R_t. The effects are mixed through the action prior $p_{0,a}$ and hence the name *mixed-effect*.

Our setting can be viewed as a two-level graphical model, where $\psi_1, \ldots, \psi_L$ are parent nodes and $\theta_1, \ldots, \theta_K$ are child nodes (Figure 3.1). The *structure* is represented by missing arrows from parent (effect parameters) to child (action parameters) nodes. A missing arrow from parent ψ_ℓ to child θ_a means that action a is independent of the ℓ-th effect.

Our model can capture all examples provided in the introduction of this chapter. For instance, in movie recommendation, the categories $\ell \in [L]$ and movies $a \in \mathcal{A}$ would

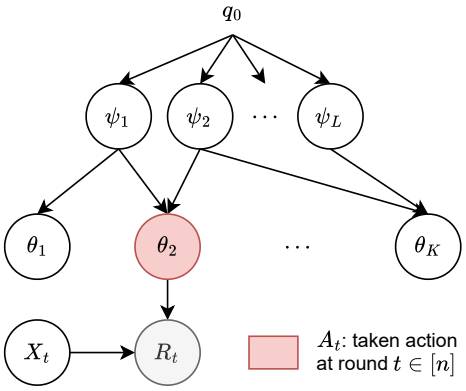


Figure 3.1: Example of a graphical model induced by Equation (3.1).

be represented by the effect parameters ψ_ℓ and action parameters θ_a, respectively. The weight $b_{a,\ell}$ is the relevance of movie a to category ℓ.

Linearity in effects. A simple yet powerful assumption is that the action prior $p_{0,a}$ is parametrized by a weighted sum of effect parameters

$$\theta_a \mid \Psi \sim p_{0,a}\Big(\cdot \,\Big|\, \sum_{\ell=1}^{L} b_{a,\ell}\psi_\ell\Big)\,, \qquad\qquad \forall a \in \mathcal{A}\,,$$

where $b_a = (b_{a,\ell})_{\ell\in[L]} \in \mathbb{R}^L$ are L *known mixing weights* for action a. The effect ℓ on action a is determined by $b_{a,\ell}$. As an example, $b_{a,\ell} = 0$ when action a is independent of effect ℓ. This is an important special case of our setting, since additive models are widely used in both theory and practice, as they often yield closed-form posteriors that are computationally tractable. Next we present two instances of this setting, where $p_{0,a}$ is a multivariate Gaussian with mean $\sum_{\ell=1}^{L} b_{a,\ell}\psi_\ell$ and covariance $\Sigma_{0,a}$.

3.1.1 Mixed-Effect Linear Bandit

A natural joint effect prior q_0 for d-dimensional effect parameters ψ_ℓ is a multivariate Gaussian with mean $\mu_\Psi \in \mathbb{R}^{Ld}$ and covariance $\Sigma_\Psi \in \mathbb{R}^{Ld\times Ld}$. The action prior $p_{0,a}$ is a Gaussian with mean $\sum_{\ell=1}^{L} b_{a,\ell}\psi_\ell \in \mathbb{R}^d$ and covariance $\Sigma_{0,a} \in \mathbb{R}^{d\times d}$:

$$\begin{aligned}
\Psi &\sim \mathcal{N}(\mu_\Psi, \Sigma_\Psi)\,, && (3.2)\\
\theta_a \mid \Psi &\sim \mathcal{N}\Big(\sum_{\ell=1}^{L} b_{a,\ell}\psi_\ell,\, \Sigma_{0,a}\Big)\,, && \forall a \in \mathcal{A}\,,\\
R_t \mid X_t, A_t, \theta, \Psi &\sim \mathcal{N}(X_t^\top \theta_{A_t}, \sigma^2)\,, && \forall t \in [T]\,,
\end{aligned}$$

where $\sigma^2 > 0$ is the variance of the observation noise.

3.1.2 Mixed-Effect Generalized Linear Bandit

Here the effect and action parameters are generated as in Equation (3.2) but the reward R_t is sampled from a *generalized linear model (GLM)* (McCullagh and Nelder, 1989), which is non-linear. In particular, $p(\cdot \mid X_t; \theta_a)$ is an exponential-family distribution with mean $g(X_t^\top \theta_a)$ and the whole model is

$$\begin{aligned} \Psi &\sim \mathcal{N}(\mu_\Psi, \Sigma_\Psi)\,, && (3.3)\\ \theta_a \mid \Psi &\sim \mathcal{N}\Big(\sum_{\ell=1}^{L} b_{a,\ell}\psi_\ell,\ \Sigma_{0,a}\Big)\,, && \forall a \in \mathcal{A}\,,\\ R_t \mid X_t, A_t, \theta, \Psi &\sim p(\cdot \mid X_t; \theta_{A_t})\,, && \forall t \in [T]\,. \end{aligned}$$

Let $\mathrm{Ber}(p)$ be a Bernoulli distribution with mean p. One particular choice of a GLM is $g(u) = 1/(1+\exp(-u))$ and $p(\cdot \mid X_t; \theta) = \mathrm{Ber}(g(X_t^\top \theta))$, which corresponds to a logistic bandit (Filippi et al., 2010).

Remark 3. *Note that in both settings, we use $x^\top \theta$ instead of $\phi(x)^\top \theta$ for some feature-map ϕ, but this is just for ease of exposition, and everything generalizes smoothly to when using ϕ.*

3.1.3 Structure Learning

The structures in Equations (3.2) and (3.3) may be intrinsic in some problems, such as drug design. When this is not the case, we propose the following approach to learning a *proxy structure*. For any $a \in \mathcal{A}$, let $\hat{\theta}_a$ represent an offline estimate of action parameter θ_a (e.g., learned offline using interactions from previous bandit tasks). To learn, we fit a Gaussian mixture model (GMM) (Reynolds et al., 2009) with L clusters to $\hat{\theta}_a$. Each cluster $\ell \in [L]$ is represented by its center $\mu_{\psi_\ell} \in \mathbb{R}^d$ and covariance $\Sigma_{\psi_\ell} \in \mathbb{R}^{d\times d}$. These correspond to the mean of the effect parameter ψ_ℓ and its uncertainty. The GMM also outputs the probability that $\hat{\theta}_a$ belongs to cluster ℓ, for all combinations of $a \in \mathcal{A}$ and $\ell \in [L]$. This probability is the mixing weight $b_{a,\ell}$.

The proposed procedure is general and adaptable to a wide range of use cases. The primary challenge lies in deriving the offline estimates $\hat{\theta}_a$. A straightforward approach involves learning these parameters from historical data collected in previous bandit tasks. Broadly, this can be formulated as an offline representation-learning problem (Tripuraneni et al., 2021), for which numerous techniques exist. For instance, in our MovieLens experiments (Section 3.4.2), we employ a low-rank factorization of the rating matrix to obtain these estimates. A key strength of our approach is its flexibility; it integrates seamlessly with standard offline learning tools, thereby taking a step toward bridging the gap between offline and online learning.

3.2 Algorithm

We propose a Thompson sampling algorithm (Thompson, 1933; Russo and Van Roy, 2014; Scott, 2010), which is a natural Bayesian solution to our problem. The algorithm

Algorithm 1 `meTS`: **M**ixed-**E**ffect **T**hompson **S**ampling.

Input: Joint effect prior q_0, action priors $p_{0,\cdot}$.
Initialize $q_1 \leftarrow q_0$ and $p_{1,\cdot} \leftarrow p_{0,\cdot}$.
for $t = 1, \ldots, T$ **do**
  Sample $\Psi_t \sim q_t$
  for $a = 1, \ldots, K$ **do**
    Sample $\theta_{t,a} \sim p_{t,a}(\cdot \mid \Psi_t)$
  $\theta_t \leftarrow (\theta_{t,a})_{a \in \mathcal{A}}$
  $A_t \leftarrow \operatorname{argmax}_{a \in \mathcal{A}} r(X_t, a; \theta_t)$
  Receive reward $R_t \sim p(\cdot \mid X_t; \theta_{*,A_t})$
  Compute new posteriors q_{t+1} and $p_{t+1,\cdot}$.

is based on hierarchical sampling (Lindley and Smith, 1972), which reflects the structure in our model. Before we present it, we need to introduce additional notation. We denote by $H_t = (X_i, A_i, R_i)_{i \in [t-1]}$ the *history* of all interactions of the agent up to round t, by $S_{t,a} = \{i \in [t-1] : A_i = a\}$ the rounds where the agent takes action a up to round t, and by $H_{t,a} = (X_i, A_i, R_i)_{i \in S_{t,a}}$ the corresponding history.

Our algorithm `meTS` is presented in Algorithm 1. Since effect parameters are shared across actions, their posteriors exhibit dependencies. To handle this, we maintain two types of posterior densities:

- A *joint effect posterior* $q_t(\Psi) = p(\Psi \mid H_t)$ for all effect parameters Ψ in round t;
- An *action posterior* $p_{t,a}(\theta \mid \Psi) = p(\theta_a \mid H_{t,a}, \Psi)$ for each action $a \in \mathcal{A}$, conditioned on the effect parameters.

`meTS` employs hierarchical sampling in each round t:

1. Sample effect parameters: $\Psi_t \sim q_t(\cdot)$
2. Sample action parameters: $\theta_{t,a} \sim p_{t,a}(\cdot \mid \Psi_t)$ for each $a \in \mathcal{A}$
3. Select action: $A_t = \operatorname{argmax}_{a \in \mathcal{A}} r(X_t, a; \theta_t)$ where $\theta_t = (\theta_{t,a})_{a \in \mathcal{A}}$.

This hierarchical sampling scheme is equivalent to sampling from the exact marginal posterior $p(\theta_a \mid H_t)$. To see this, observe that marginalizing over Ψ yields:

$$\begin{aligned} p(\theta_a \mid H_t) &= \int_\Psi p(\theta_a, \Psi \mid H_t) \, \mathrm{d}\Psi \,, \\ &= \int_\Psi p(\theta_a \mid \Psi, H_t) \, p(\Psi \mid H_t) \, \mathrm{d}\Psi \,, \\ &= \int_\Psi p_{t,a}(\theta_a \mid \Psi) \, q_t(\Psi) \, \mathrm{d}\Psi \,. \end{aligned} \tag{3.4}$$

3.2.1 Posterior Derivations

The posteriors are computed as follows. We first express the joint effect posterior q_t as

$$q_t(\Psi) \propto \prod_{a=1}^{K} \int_{\theta_a} \mathcal{L}_{t,a}(\theta_a) p_{0,a}\left(\theta_a \mid \Psi\right) \mathrm{d}\theta_a \; q_0(\Psi)\,, \tag{3.5}$$

where $\mathcal{L}_{t,a}(\theta_a) = \prod_{(x,a,r)\in H_{t,a}} p(r \mid x; \theta_a)$ is the likelihood of all observations of action a up to round t given θ_a. Next, for any action $a \in \mathcal{A}$, the action posterior $p_{t,a}$ is expressed as

$$p_{t,a}(\theta_a \mid \Psi) \propto \mathcal{L}_{t,a}(\theta_a) p_{0,a}(\theta_a \mid \Psi)\,. \tag{3.6}$$

$p_{t,a}$ is similarly sparse to $p_{0,a}$. Specifically, in any round t, $p_{t,a}$ and $p_{0,a}$ are parameterized by the same subset of effect parameters Ψ, since $\mathcal{L}_{t,a}(\theta_a)$ does not depend on Ψ.

The joint effect posterior q_t and action posteriors $p_{t,a}$ have closed forms in Gaussian models, which allows efficient sampling and theoretical analysis. Beyond these, MCMC and variational inference can be used to approximate q_t and $p_{t,a}$. Next we derive closed-form posteriors for the mixed-effect model with linear rewards in Equation (3.2) and provide an efficient approximation for the mixed-effect model with non-linear rewards in Equation (3.3).

3.2.2 Mixed-Effect Linear Bandit

Let

$$G_{t,a} = \sigma^{-2} \sum_{i \in S_{t,a}} X_i X_i^\top\,, \qquad\qquad B_{t,a} = \sigma^{-2} \sum_{i \in S_{t,a}} R_i X_i\,. \tag{3.7}$$

be the outer product of contexts corresponding to action a up to round t, and their sum weighted by rewards, respectively. Both are scaled by the observation noise variance σ^2. Using these quantities, the effect posterior is defined as follows.

Proposition 1. *For any round $t \in [T]$, the joint effect posterior is a multivariate Gaussian $q_t = \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t)$, where*

$$\begin{aligned} \bar{\Sigma}_t^{-1} &= \Sigma_\Psi^{-1} + \sum_{a \in \mathcal{A}} b_a b_a^\top \otimes \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right)\,, \\ \bar{\mu}_t &= \bar{\Sigma}_t\Big(\Sigma_\Psi^{-1}\mu_\Psi + \sum_{a \in \mathcal{A}} b_a \otimes (\Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1} B_{t,a})\Big)\,. \end{aligned} \tag{3.8}$$

The effect posterior is additive in individual actions and each action contributes to the effect posterior mean and covariance proportionally to $b_{a,\ell}$, which is the mixture weight for θ_a in Equation (3.2). Proposition 1 is proved in Section A.2.1.

Now we present the action posterior.

Proposition 2. *For any round $t \in [T]$, action $a \in \mathcal{A}$, and effect parameters Ψ_t, the action posterior is a multivariate Gaussian $p_{t,a}(\cdot \mid \Psi_t) = \mathcal{N}(\cdot; \tilde{\mu}_{t,a}, \tilde{\Sigma}_{t,a})$, where*

$$\begin{aligned} \tilde{\Sigma}_{t,a}^{-1} &= \Sigma_{0,a}^{-1} + G_{t,a}\,, \\ \tilde{\mu}_{t,a} &= \tilde{\Sigma}_{t,a}\Big(\Sigma_{0,a}^{-1} \sum_{\ell=1}^{L} b_{a,\ell}\psi_{t,\ell} + B_{t,a}\Big)\,. \end{aligned} \tag{3.9}$$

The action posterior in Equation (3.9) is a standard multivariate Gaussian posterior whose prior depends on Ψ_t, which is sampled by `meTS`. Proposition 2 is proved in Section A.2.2.

3.2.3 Mixed-Effect Generalized Linear Bandit

Closed-form posteriors are unavailable in this setting, so approximations are required. We use a Laplace-style scheme that approximates the *likelihood* $\mathcal{L}_{t,a}(\cdot)$ by a Gaussian, rather than applying Laplace to the full posterior. This choice preserves a Gaussian form that can be propagated analytically through the hierarchical updates.

Let $\mu_{t,a}^{\text{LAP}}$ denote the MLE (see remark below for a discussion about the computation of the MLE in practice), and let $G_{t,a}^{\text{LAP}}$ be the Hessian[1] of $-\log \mathcal{L}_{t,a}(\cdot)$:

$$\mu_{t,a}^{\text{LAP}} = \operatorname*{argmax}_{\theta_a} \log \mathcal{L}_{t,a}(\theta_a), \qquad G_{t,a}^{\text{LAP}} = \sum_{i \in S_{t,a}} \dot{g}(X_i^\top \mu_{t,a}^{\text{LAP}})\, X_i X_i^\top.$$

We then approximate the *likelihood* (not the posterior) by

$$\mathcal{L}_{t,a}(\theta_a) \;\propto\; \exp\!\big(-\tfrac{1}{2}(\theta_a - \mu_{t,a}^{\text{LAP}})^\top G_{t,a}^{\text{LAP}}(\theta_a - \mu_{t,a}^{\text{LAP}})\big)\,, \tag{3.10}$$

Substituting Equation (3.10) into Equation (3.5) yields

$$q_t(\cdot) \approx \mathcal{N}(\cdot; \bar{\mu}_t, \bar{\Sigma}_t),$$

where $\bar{\mu}_t$ and $\bar{\Sigma}_t$ are computed as in Proposition 1, except for the replacements

$$G_{t,a} \leftarrow G_{t,a}^{\text{LAP}}, \qquad B_{t,a} \leftarrow G_{t,a}^{\text{LAP}} \mu_{t,a}^{\text{LAP}}.$$

Similarly, substituting Equation (3.10) into Equation (3.6) gives

$$p_{t,a}(\cdot \mid \Psi) \approx \mathcal{N}(\cdot; \tilde{\mu}_{t,a}, \tilde{\Sigma}_{t,a}),$$

with

$$G_{t,a} \leftarrow G_{t,a}^{\text{LAP}}, \qquad B_{t,a} \leftarrow G_{t,a}^{\text{LAP}} \mu_{t,a}^{\text{LAP}}.$$

[1] Note that we are assuming a generalized linear model where the log-likelihood of the data associated with action a can be written as $\log \mathcal{L}_{t,a}(\theta_a) = \sum_{i \in S_{t,a}} \big[R_i X_i^\top \theta_a - A(X_i^\top \theta_a) + C(R_i)\big]$ where C is a real-valued function and A is twice continuously differentiable, with derivative $\dot{A} = g$ representing the mean function.

Although these expressions follow mechanically from substituting the Gaussian likelihood approximation, the intuition is straightforward. The replacement $G_{t,a} \leftarrow G_{t,a}^{\text{LAP}}$ reflects the curvature induced by the nonlinear mean function g, while $B_{t,a} \leftarrow G_{t,a}^{\text{LAP}} \mu_{t,a}^{\text{LAP}}$ mirrors the linear-Gaussian case, where the MLE $\hat{\theta}_{t,a}^{\text{MLE}}$ is characterized by the normal equations $G_{t,a}\, \hat{\theta}_{t,a}^{\text{MLE}} = B_{t,a}$, and its generalized-linear counterpart is $\mu_{t,a}^{\text{LAP}}$.

Remark 4. *The MLE* $\mu_{t,a}^{LAP} = \operatorname{argmax}_{\theta_a \in \mathbb{R}^d} \log \mathcal{L}_{t,a}(\theta_a)$ *may be ill-posed. In practice, we use a small* ℓ_2*-regularized estimator:* $\mu_{t,a}^{LAP} \in \operatorname{argmax}_{\theta_a \in \mathbb{R}^d} \log \mathcal{L}_{t,a}(\theta_a) - \frac{\lambda}{2}\|\theta_a\|_2^2$*, where* $\lambda > 0$ *to fix this.*

3.2.4 Computational Complexity

The benefit of modeling the effect parameters is not immediately clear. Thus, it is tempting to marginalize them out, and only maintain a single joint posterior of all action parameters $\theta \in \mathbb{R}^{Kd}$. Posterior updates in this case would be complex and computationally inefficient when $K \gg L$, which is common in practice.

The main advantage of `meTS` is that the sampling of effect parameters $\Psi_t \sim q_t$ allows us to use the conditional independence of actions given Ψ, and model $\theta_a \mid H_{t,a}, \Psi_t$ independently. This is more computationally efficient than modeling the joint $\theta \mid H_t$ when $K \gg L$. To see this, suppose that all posteriors are multivariate Gaussians (Section 3.2.2). In this case, $\theta \mid H_t$ requires $\mathcal{O}(K^2 d^2)$ space, due to storing a $Kd \times Kd$ covariance matrix; while `meTS` requires only $\mathcal{O}((L^2 + K)d^2)$ space, due to storing the covariances of q_t and $p_{t,a}$. Since the sampling relies on covariance inverses, the time complexity also improves. For the joint posterior, it is $\mathcal{O}(K^3 d^3)$, while it is only $\mathcal{O}((L^3 + K)d^3)$ for `meTS`.

One can also marginalize out the effect parameters Ψ and have K separate posteriors, one for each action parameter θ_a. While this improves computational efficiency, it does not model that the actions are correlated, since $\theta_a \mid H_{t,a}$ is modeled instead of $\theta_a \mid H_t$. This leads to a statistical inefficiency due to the loss of information as the histories of other actions $H_{t,a'}$ are discarded. We validate this through theory (Section 3.3.2) and experiments (Section 3.4).

3.3 Analysis

This section is organized as follows. First, we state our regret bound. Then, we discuss how it captures the structure of our problem. We use $\tilde{\mathcal{O}}$ for the big O notation up to polylogarithmic factors.

3.3.1 Main Result

We analyze `meTS` in the linear setting in Section 3.1.1. Throughout, we assume that the *true* action parameters and rewards are generated according to the same hierarchical model used by `meTS` (Equation (3.2)), i.e., we operate in the fully well-specified setting. For ease of exposition, we further assume the existence of constants $\sigma_0, \sigma_\Psi, \kappa_x > 0$ such

that

$$\Sigma_{0,a} = \sigma_0^2 I_d \quad \text{for all } a \in \mathcal{A}, \qquad \Sigma_\Psi = \sigma_\Psi^2 I_{Ld}, \qquad \|X_t\|_2^2 \le \kappa_x \quad \text{for all } t \in [T].$$

The bound on $\|X_t\|_2$ is standard, and we relax the other two assumptions in Section A.3.

Theorem 1. *For any $\delta \in (0,1)$, the Bayes regret of* `meTS` *in the mixed-effect model in Section 3.1.1 is bounded as*

$$\mathcal{BR}(T) \le \sqrt{2T\left(\mathcal{R}^{\text{A}}(T) + \mathcal{R}^{\text{E}}(T)\right)\log(1/\delta)} + cT\delta\,, \tag{3.11}$$

where $c = \sqrt{\frac{2}{\pi}\kappa_x(\sigma_0^2 + \kappa_b\sigma_\Psi^2)}K$, $\kappa_b = \max_{a\in\mathcal{A}}\|b_a\|_2^2$,

$$\mathcal{R}^{\text{A}}(T) = dKc_{\text{A}}\log\left(1 + \frac{T\kappa_x\sigma_0^2}{d\sigma^2}\right),\ c_{\text{A}} = \frac{\kappa_x\sigma_0^2}{\log\left(1+\frac{\kappa_x\sigma_0^2}{\sigma^2}\right)},$$
$$\mathcal{R}^{\text{E}}(T) = dLc_{\text{E}}\log\left(1 + \frac{K\kappa_b\sigma_\Psi^2}{\sigma_0^2 + \frac{\sigma^2}{T\kappa_x}}\right),\ c_{\text{E}} = \frac{\kappa_x\kappa_b\sigma_\Psi^2\left(1+\frac{\kappa_x\sigma_0^2}{\sigma^2}\right)}{\log\left(1+\frac{\kappa_x\kappa_b\sigma_\Psi^2}{\sigma^2}\right)}.$$

The second term in Equation (3.11) is constant for $\delta = 1/T$, in which case the above bound is $\tilde{\mathcal{O}}(\sqrt{T})$. The main quantities of interest are $\mathcal{R}^{\text{A}}(T)$ and $\mathcal{R}^{\text{E}}(T)$, and they have natural interpretations. $\mathcal{R}^{\text{A}}(T)$ corresponds to the action regression problem: with K parameters of dimension d, prior width σ_0, maximum context length $\sqrt{\kappa_x}$, and T observations with noise σ. The dependence of $\mathcal{R}^{\text{A}}(T)$ on these quantities is identical to a corresponding linear bandit (Lu and Van Roy, 2019). On the other hand, $\mathcal{R}^{\text{E}}(T)$ corresponds to the effect regression problem: with L parameters of dimension d, prior width σ_Ψ, maximum mixing-weight length $\sqrt{\kappa_b}$, and K actions that can be viewed as observations with noise σ_0 (Section 3.2.2). The dependence of $\mathcal{R}^{\text{E}}(T)$ on these quantities mimics those in $\mathcal{R}^{\text{A}}(T)$.

To simplify exposition, let $\kappa_x = \kappa_b = \sigma = 1$. Then

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}\left(\sqrt{Td\big(K\sigma_0^2 + L\sigma_\Psi^2(1+\sigma_0^2)\big)}\right). \tag{3.12}$$

This can be re-written as $\mathcal{BR}(T) = \tilde{\mathcal{O}}\left(\sqrt{TdK_{\text{eff}}(\sigma_0^2+\sigma_\Psi^2)}\right)$, where $K_{\text{eff}} = \frac{K\sigma_0^2 + L\sigma_\Psi^2(1+\sigma_0^2)}{\sigma_0^2+\sigma_\Psi^2}$ is the *effective number of actions*. When $L \ll K$ and $\sigma_0^2 \ll \sigma_\Psi^2$, we have $K_{\text{eff}} \ll K$, yielding significant regret reduction over standard Thompson Sampling.

The dependence on σ_0^2 and σ_Ψ^2 is natural: since Bayesian regret measures performance under the prior, smaller prior variances correspond to more informative beliefs about the true parameters, which makes learning easier and reduces regret. Conversely, larger variances reflect greater prior uncertainty and increase the difficulty of identifying the optimal action. The scaling with K, L, and d is also intuitive: fewer parameters to estimate lead to lower regret. These trends are consistent with our empirical observations in Section A.4.

3.3.2 Benefits of Structure

Note that we do not provide a matching lower bound. To argue that our upper bound reflects the intrinsic structure of the problem, we compare `meTS` to agents that either have access to more information or exploit less structure. We start with the former. Consider `meTS` with known effect parameters Ψ. Setting $\sigma_\Psi = 0$ in Equation (3.12) yields the reduced regret

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}(\sqrt{TdK\sigma_0^2}),$$

which no longer depends on L. Likewise, consider `meTS` under a perfectly specified linear model, in which $\theta_a = \sum_{\ell \in [L]} b_{a,\ell}\psi_\ell$ for all $a \in \mathcal{A}$. This corresponds to $\sigma_0 = 0$, giving

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}(\sqrt{TdL\sigma_\Psi^2}),$$

which is independent of K. In particular, the K-dependence in our regret bound arises precisely from modeling the variability of action parameters around the effect parameters via $\Sigma_{0,a}$. Without it, the regret of `sDM` is independent of K

We now turn to an agent that neither knows Ψ nor models it explicitly. This agent learns only θ (Section 3.2.4) by marginalizing out Ψ in Equation (3.2):

$$\theta_a \sim \mathcal{N}\Big(\sum_{\ell=1}^{L} b_{a,\ell}\mu_{\psi_\ell},\ \breve{\Sigma}_{0,a}\Big), \qquad \forall a \in \mathcal{A},$$

where $\breve{\Sigma}_{0,a} = (\sigma_0^2 + \|b_a\|_2^2\sigma_\Psi^2)I_d$ is the marginal prior covariance and $(\mu_{\psi_\ell})_{\ell\in[L]}$ is the prior mean of the effects, so that $\mu_\Psi = (\mu_{\psi_\ell})_{\ell\in[L]}$ (Section 3.1.1). Importantly, marginalizing out Ψ and treating each action parameter independently discards action correlations, even though the true generative model induces correlations via the shared effects. This agent therefore uses a less structured and less informative prior than `meTS`.

Using the definition of $\breve{\Sigma}_{0,a}$ and $\kappa_b = \max_{a\in\mathcal{A}} \|b_a\|_2^2 = 1$, the regret of this agent scales as in Equation (3.12) with $\sigma_\Psi = 0$, except that the maximum prior variance σ_0^2 is replaced by $\sigma_0^2 + \sigma_\Psi^2$. Hence,

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}\big(\sqrt{TdK(\sigma_0^2 + \sigma_\Psi^2)}\big).$$

When $K > L$ (up to constants), this regret can be substantially larger than the bound for `meTS` in Equation (3.12). The improvement is on the order of $\sqrt{K/L}$ in regimes where the effects are far more uncertain than the actions, i.e., $\sigma_\Psi \gg \sigma_0$. For example, in our ad-placement setting, L is the number of catalog items, while $K \approx L^M$ is the number of slates of size M. Thus $K/L \approx L^{M-1}$, with typical scales such as $L \approx 10^6$ and $M \approx 10$. Our empirical results in Sections A.4 and 3.4.1 support this: `meTS` significantly outperforms classical methods when the effect parameters are more uncertain than the action parameters.

3.4 Experiments

We evaluate `meTS` on both synthetic and real-world problems. In each plot, we report the average values and their standard errors. Additional experiments are conducted in

Section A.4. The code is provided in this Github repository.

3.4.1 Synthetic Experiments

We start with two synthetic problems: the linear and logistic bandit settings in Equations (3.2) and (3.3), respectively. The effect prior is parameterized by $\mu_\Psi = \mathbf{0}_{Ld}$ and $\Sigma_\Psi = 3I_{Ld}$, the action covariance is $\Sigma_{0,a} = I_d$ for all $a \in \mathcal{A}$, and the observation noise is $\sigma = 1$. We use this setting since modeling of the effect parameters is the most beneficial when they are more uncertain than the action ones (Section 3.3.2). The context X_t is sampled uniformly from $[-1, 1]^d$. We run 50 simulations and sample the mixing weights $b_{a,\ell}$ from $[-1, 1]$ in each run.

We consider the following baselines. For the linear setting, we compare `meTS-Lin` (Section 3.2.2), `LinUCB` (Abbasi-Yadkori et al., 2011), `LinTS` (Agrawal and Goyal, 2013a) and `HierTS` (Hong et al., 2022b). For the logistic setting, we compare `meTS-GLM` (Section 3.2.3), `meTS-Lin` (Section 3.2.2), `UCB-GLM` (Li et al., 2017), `GLM-TS` (Chapelle and Li, 2012) and `HierTS` (Hong et al., 2022b). `GLM-UCB` (Filippi et al., 2010) is excluded because it exhibits very high regret. We also include variational mean-field approximations of `meTS` (`meTS-Lin-Fa` and `meTS-GLM-Fa`), where the full Gaussian effect posterior q_t is approximated as $q_t(\Psi) \approx \prod_{\ell=1}^{L} q_{t,\ell}(\psi_\ell)$. This factorization enables sampling each $\psi_\ell \in \mathbb{R}^d$ independently and replaces operations on a full $Ld \times Ld$ covariance with blockwise updates. This improves the time and space complexities of `meTS` by L^2 and L, respectively.

All baselines but `HierTS` ignore the structure. `HierTS` incorporates the structure similarly to `meTS-Lin` but only has a single effect parameter with prior $\mathcal{N}(\mathbf{0}_d, 3I_d)$, with the same mean and covariance as the effect parameters of `meTS`. To compare fairly with `LinTS` and `GLM-TS`, their marginal prior mean and covariance are chosen as $\mathbf{0}_d$ and $\breve{\Sigma}_{0,a} = \Sigma_{0,a} + \Gamma_a \Sigma_\Psi \Gamma_a^\top$, where $\Gamma_a = b_a^\top \otimes I_d$. This is to account for the uncertainty of the effect parameters despite marginalizing them out.

In Figure 3.2, we plot the regret in both problems for $T = 5000, K = 100, L = 3$, and $d = 2$ (higher values of K up to $100,000$ are tested in our additional experiment in Figure 3.3 below). `meTS` and its factored variant outperform all baselines that ignore the structure or incorporate it partially. Moreover, `meTS-GLM` outperforms `meTS-Lin` in the logistic bandit, which shows the benefit of the approximation in Section 3.2.3. This attests to the generality and flexibility of `meTS` and the posterior derivations in Section 3.2. We also show in Section A.4.1 that a higher K, L, or d leads to a higher regret due to learning more parameters, which is captured by our regret bounds.

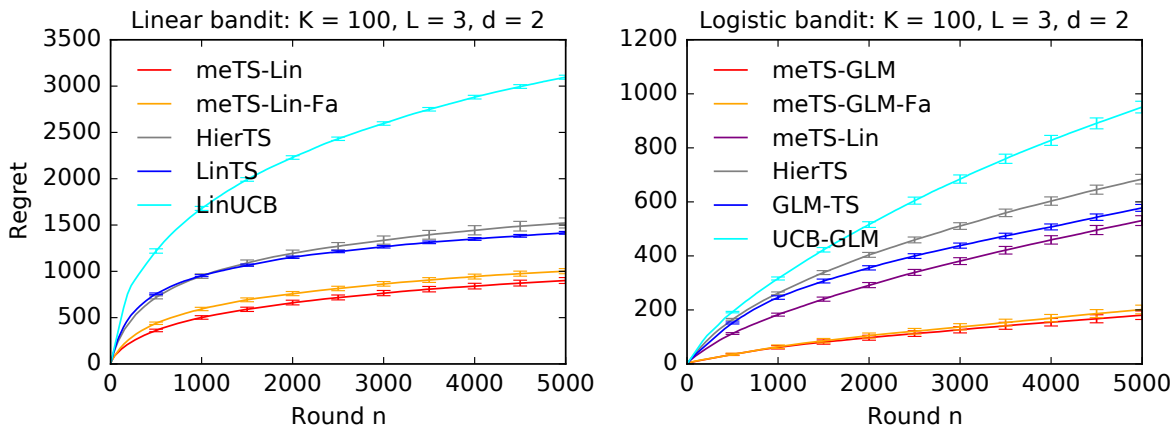


Figure 3.2: Evaluation on synthetic problems.

In Figure 3.3, we examine how the final cumulative regret scales with the number of actions K in the linear bandit setting, fixing $L = 3$ and $d = 2$. As K increases from 100 to $100,000$, `meTS-Lin` consistently achieves substantially lower regret than `LinTS`, with the gap widening as K grows. This demonstrates that `meTS-Lin` scales more favorably with the action space size by leveraging the shared effect structure, which aligns with our theoretical analysis. When $K \gg L$, this structural advantage becomes increasingly pronounced.

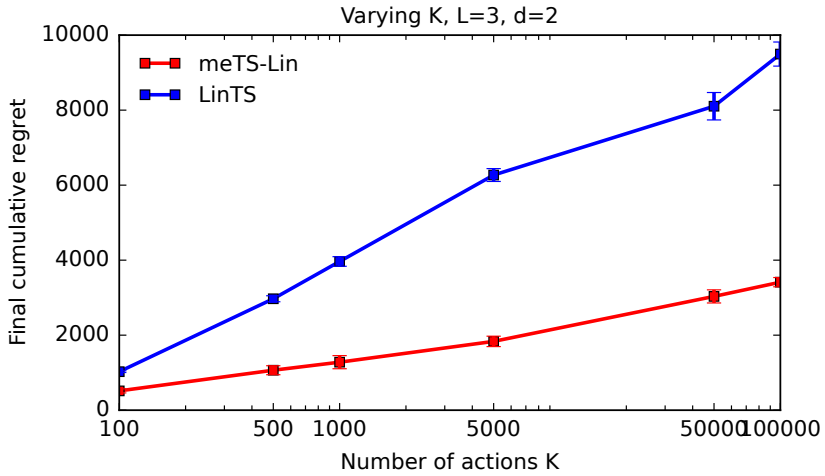


Figure 3.3: Final cumulative regret as a function of the number of actions K in the linear bandit setting with $L = 3$ and $d = 2$.

3.4.2 MovieLens Experiments

We study the problem of movie recommendation using the MovieLens 1M dataset (Lam and Herlocker, 2016). This dataset contains one million ratings given by 6040 users to 3952 movies. We apply low-rank factorization to the rating matrix to obtain 5-dimensional representations: $x_j \in \mathbb{R}^5$ for user $j \in [6040]$ and $\theta_a \in \mathbb{R}^5$ for movie $a \in [3952]$. We use

the movies as actions and the context X_t is sampled uniformly from user vectors x_j. We consider both linear and logistic rewards. Given a user x_j, the linear reward for movie θ_a is sampled from $\mathcal{N}(x_j^\top \theta_a, \sigma^2)$ while the logistic reward is sampled from $\text{Ber}(g(x_j^\top \theta_a))$, where g is the sigmoid function. We run 50 simulations with $K = 100$ randomly sampled movies in each run. We compare `meTS` to most baselines in Section 3.4.1. We do not include `UCB-GLM` and `GLM-UCB` because their regret is very high. In `LinTS` and `GLM-TS`, the prior mean of action a is μ and its covariance is $\breve{\Sigma}_0 = \text{diag}(v) \in \mathbb{R}^{d \times d}$, where $\mu \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$ are the mean and variance of the movie vectors along all dimensions, respectively.

The mixed-effect structure in Equations (3.2) and (3.3) is not available in this problem. Therefore, we use the approach in Section 3.1.3 to learn it. More precisely, we cluster the movies into $L = 5$ mixture components by training a GMM on the offline action vectors θ_a (Section 3.1.3). Each cluster center corresponds to an effect parameter mean $\mu_{\psi_\ell} \in \mathbb{R}^d$ and the mixing weight $b_{a,\ell}$ is the probability that movie a belongs to cluster ℓ, as given by the GMM. We set the effect prior covariance as $\Sigma_\Psi = 0.75\,\text{diag}((\breve{\Sigma}_0)_{\ell \in [L]}) \in \mathbb{R}^{Ld \times Ld}$ and the prior covariance of action a as $\Sigma_{0,a} = 0.25\,\breve{\Sigma}_0 \in \mathbb{R}^{d \times d}$, where $\breve{\Sigma}_0$ is the same as in both `LinTS` and `GLM-TS`. This means that the marginal covariance of action a in `meTS` is $0.25\,\breve{\Sigma}_0 + 0.75\,\Gamma_a \Sigma_\Psi \Gamma_a^\top$, where $\Gamma_a = b_a^\top \otimes I_d$. Therefore, it is on the same order as $\breve{\Sigma}_{0,a}$ when $\|b_a\|_2^2 \approx 1$, and `meTS` is parameterized comparably to `LinTS` and `GLM-TS`. At the same time, we also model that the effect parameters are more uncertain than the action ones, since $\Sigma_{0,a} = 0.25\,\breve{\Sigma}_0$ while $\Sigma_\Psi = 0.75\,\text{diag}((\breve{\Sigma}_0)_{\ell \in [L]})$.

In Figure 3.4, we plot the regret for $T = 5000$ rounds. We observe that `meTS` has the lowest regret, even if the true rewards are not generated from a mixed-effect model. This shows the robustness of `meTS` to model misspecification, which we further validate in Section A.4.3. It also highlights the flexibility of our framework, where a proxy structure is learned from offline data.

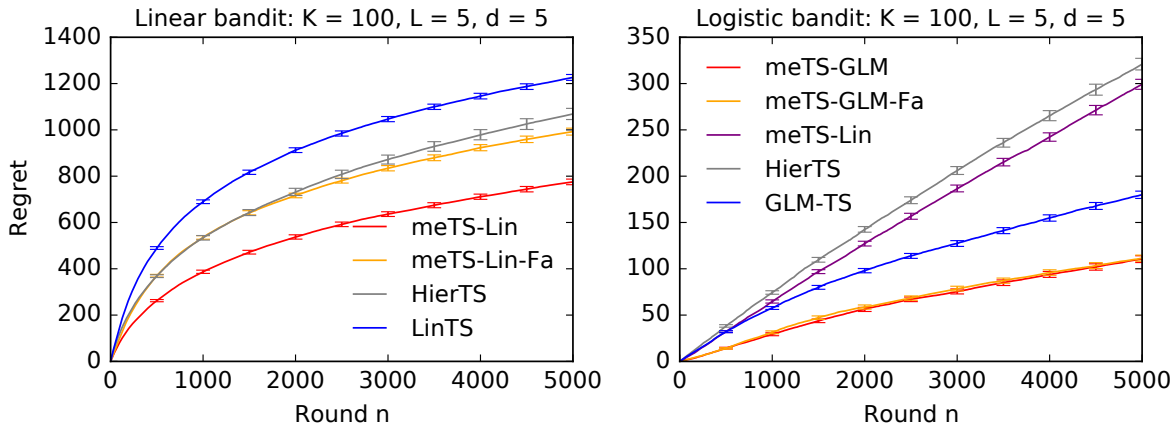


Figure 3.4: Evaluation on MovieLens problems.

3.5 Conclusion

In this chapter, we introduced a mixed-effect bandit framework based on a two-level graphical model in which each action may depend on multiple underlying effects. This

structure enables more efficient exploration, and we designed `meTS` to leverage it effectively. When implemented as analyzed, `meTS` performs strongly on both synthetic and real-world benchmarks. Although our presentation focused on the concrete models in Equations (3.2) and (3.3), the underlying algorithmic ideas extend seamlessly to the general mixed-effect model in Equation (3.1). The methodological and theoretical tools developed here lay the groundwork for richer formulations, one of which we explore in detail in the next chapter.

Our work has several limitations. First, the regret analysis assumes a well-specified prior: the true parameters must be generated from the same hierarchical model used by `meTS`. While some experiments suggest robustness to misspecification, formal guarantees under prior mismatch remain open. Second, closed-form posteriors are available only for linear-Gaussian rewards; generalized linear models require Laplace approximations (Section 3.2.3), which are not analyzed theoretically. Third, the mixed-effect structure must be known or learned offline, adding an additional modeling step. Finally, `meTS` models only two-level hierarchies; deeper latent structures, which may better capture complex action correlations, require the diffusion-based approach developed in Chapter 4.

Chapter 4

Scaling Thompson Sampling with Diffusion Models

Contents

In the previous chapter, we explored how action correlations can be captured using mixed-effects models, in which actions share a set of effect parameters. This approach proved effective when the underlying structure, such as categories in movie recommendation or components in drug design, can be learned through clustering. However, real-world action correlations might exhibit more complex patterns. Thus, this chapter presents an alternative approach inspired by the remarkable success of diffusion models in approximating complex distributions (Sohl-Dickstein et al., 2015; Ho et al., 2020; Dhariwal and Nichol,

2021; Rombach et al., 2022). Rather than explicitly modeling shared effects, we leverage pre-trained diffusion models to capture the rich structure of action parameters and use them as priors in contextual Thompson sampling.

We make the following contributions. **1)** We introduce a framework for contextual bandits with *diffusion-derived priors* and develop diffusion Thompson sampling (`dTS`), which is both statistically efficient and computationally tractable. `dTS` enables fast posterior updates and sampling via an efficient approximation inspired by exact Gaussian posteriors. **2)** Beyond applying pre-trained diffusion models to contextual bandits, a key contribution is enabling efficient *posterior computation* and *sampling* for a d-dimensional parameter $\theta \mid \mathcal{D}$ under a diffusion model prior, without updating the diffusion model parameters (i.e., without backpropagating through the neural network). This is relevant not only to bandits and RL but also to broader applications (Chung et al., 2022). Our approximations are motivated by exact closed-form solutions available when the diffusion model is fully linear; these solutions form the basis for our nonlinear approximations, which achieve strong empirical performance while avoiding the computational burden of standard approximate posterior sampling techniques.

4.1 Setting

We consider the contextual bandit setting in Section 2.1. Then, we define the prior distribution using a diffusion model, with a set of L consecutive *unknown latent parameters* $\psi_\ell \in \mathbb{R}^d$ for $\ell \in [L]$. Precisely, the action parameter θ_a depends on the 1-st latent parameter ψ_L as $\theta_a \mid \psi_1 \sim \mathcal{N}(f_1(\psi_1), \Sigma_1)$, where the *link function* f_1 and covariance Σ_1 are *known*. Also, the $\ell-1$-th latent parameter $\psi_{\ell-1}$ depends on the ℓ-th latent parameter ψ_ℓ as $\psi_{\ell-1} \mid \psi_\ell \sim \mathcal{N}(f_\ell(\psi_\ell), \Sigma_\ell)$, where f_ℓ and Σ_ℓ are *known*. Finally, the L-th latent parameter ψ_L is sampled as $\psi_L \sim \mathcal{N}(0, \Sigma_{L+1})$, where Σ_{L+1} is *known*. We summarize this model in Equation (4.1) below:

$$\begin{aligned} \psi_L &\sim \mathcal{N}(0, \Sigma_{L+1}) \,, && \\ \psi_{\ell-1} \mid \psi_\ell &\sim \mathcal{N}(f_\ell(\psi_\ell), \Sigma_\ell) \,, && \forall \ell \in [L]/\{1\} \,, \\ \theta_a \mid \psi_1 &\sim \mathcal{N}(f_1(\psi_1), \Sigma_1) \,, && \forall a \in \mathcal{A} \,, \\ R_t \mid \theta, (\psi_\ell)_{\ell \in [L]}, X_t, A_t &\sim p(\cdot \mid X_t; \theta_{A_t}) \,, && \forall t \in [T] \,. \end{aligned} \tag{4.1}$$

In practice, this model can be built by pre-training a diffusion model on offline estimates of the action parameters θ_a.

Remark 5 (**Joint models**). *Our algorithm and analysis also apply to the case where all actions share a single unknown parameter $\theta \in \mathbb{R}^d$. Let $\phi : \mathcal{X} \times [K] \to \mathbb{R}^d$ be a known feature map, and assume the reward distribution mean is $g\left(\phi(x,a)^\top \theta\right)$. Then, the diffusion prior in Equation (4.1) specializes by replacing the per-action parameters $(\theta_a)_{a \in [K]}$ with a single shared parameter θ:*

$$\begin{aligned} \psi_L &\sim \mathcal{N}(0, \Sigma_{L+1}), && \\ \psi_{\ell-1} \mid \psi_\ell &\sim \mathcal{N}\left(f_\ell(\psi_\ell), \Sigma_\ell\right), && \forall\, \ell \in [L] \setminus \{1\}, \\ \theta \mid \psi_1 &\sim \mathcal{N}\left(f_1(\psi_1), \Sigma_1\right), && \\ R_t \mid \theta, (\psi_\ell)_{\ell \in [L]}, X_t, A_t &\sim p\left(\cdot \,\middle|\, \phi(X_t, A_t)^\top \theta\right), && \forall\, t \in [T]. \end{aligned} \tag{4.2}$$

This formulation is useful when a shared feature map ϕ is available. In that case, the diffusion model can be pre-trained on parameters $\{\theta_s\}_{s=1}^S$ from previous tasks, and `dTS` *can then be applied to a new task $S+1$ using the pre-trained prior. To avoid clutter, our main exposition focuses on the model in Equation (4.1), but all theoretical results and algorithmic components extend naturally to this shared-parameter case, which we also include in some experiments (explicitly noted when applicable).*

4.2 Algorithm

We design a Thompson sampling algorithm that samples the latent and action parameters hierarchically (Lindley and Smith, 1972). Let $H_t = (X_i, A_i, R_i)_{i\in[t-1]}$ denote the history of all interactions up to round t, and let $H_{t,a} = (X_i, A_i, R_i)_{\{i\in[t-1];A_i=a\}}$ be the history of interactions *with action* a up to round t. To motivate our algorithm, we decompose the posterior density $p(\theta_a \mid H_t)$ recursively as

$$p(\theta_a \mid H_t) = \int_{\psi_{1:L}} p(\psi_L \mid H_t) \prod_{\ell=2}^{L} p(\psi_{\ell-1} \mid \psi_\ell, H_t) p(\theta_a \mid \psi_1, H_{t,a}) \, \mathrm{d}\psi_{1:L}. \tag{4.3}$$

Hierarchical sampling. This decomposition induces the following sampling procedure. First, draw a sample $\psi_{t,L}$ according to the posterior density $p(\psi_L \mid H_t)$. Then, for each $\ell \in [L] \setminus \{1\}$, draw $\psi_{t,\ell-1}$ from the conditional posterior $p(\psi_{\ell-1} \mid \psi_{t,\ell}, H_t)$. Finally, given $\psi_{t,1}$, draw each action parameter independently from $p(\theta_a \mid \psi_{t,1}, H_{t,a})$ (the θ_a are conditionally independent given ψ_1). This defines Algorithm 2, **d**iffusion **T**hompson **S**ampling (`dTS`).

Posterior components via recursion. To implement `dTS`, we provide a recursive scheme to express the required posteriors using known quantities. These expressions may not always admit closed forms and often require approximation. The conditional action-posterior can be written as

$$p(\theta_a \mid \psi_1, H_{t,a}) \propto \prod_{i\in S_{t,a}} p(R_i \mid X_i; \theta_a) \, \mathcal{N}(\theta_a; f_1(\psi_1), \Sigma_1), \tag{4.4}$$

where $S_{t,a} = \{\ell \in [t-1] : A_\ell = a\}$ is the set of rounds in which action a was selected.

Now, we characterize the conditional latent-posteriors. Before we do so, we make the following notation clarification. With slight abuse of notation, $p(H_t \mid \psi_\ell)$ denotes the likelihood of the observations up to round t given ψ_ℓ:

$$p(H_t \mid \psi_\ell) = p((R_i)_{i<t} \mid (X_i)_{i<t}, (A_i)_{i<t}, \psi_\ell)$$

With this notation in mind, for any $\ell \in [L] \setminus \{1\}$, the conditional latent-posterior is

$$p(\psi_{\ell-1} \mid \psi_\ell, H_t) \propto p(H_t \mid \psi_{\ell-1}) \, \mathcal{N}(\psi_{\ell-1}; f_\ell(\psi_\ell), \Sigma_\ell),$$

and the top-layer posterior is

$$p(\psi_L \mid H_t) \propto p(H_t \mid \psi_L) \, \mathcal{N}(\psi_L; 0, \Sigma_{L+1}).$$

All terms above are known except the likelihoods $p(H_t \mid \psi_\ell)$, which are computed recursively. The recursion starts with

$$p(H_t \mid \psi_1) = \prod_{a=1}^{K} \int_{\theta_a} \left[\prod_{i \in S_{t,a}} p(R_i \mid X_i; \theta_a) \right] \mathcal{N}(\theta_a; f_1(\psi_1), \Sigma_1) \, \mathrm{d}\theta_a, \tag{4.5}$$

and for $\ell \in [L] \setminus \{1\}$, proceeds as

$$p(H_t \mid \psi_\ell) = \int_{\psi_{\ell-1}} p(H_t \mid \psi_{\ell-1}) \, \mathcal{N}(\psi_{\ell-1}; f_\ell(\psi_\ell), \Sigma_\ell) \, \mathrm{d}\psi_{\ell-1}. \tag{4.6}$$

Algorithm 2 `dTS`: **d**iffusion **T**hompson **S**ampling

Input: Prior components $\{f_\ell, \Sigma_\ell\}_{\ell=1}^{L+1}$ and reward model p.
for $t = 1, \ldots, T$ **do**
  Draw $\psi_{t,L}$ according to the posterior density $p(\psi_L \mid H_t)$
  for $\ell = L, \ldots, 2$ **do**
    Draw $\psi_{t,\ell-1}$ according to $p(\psi_{\ell-1} \mid \psi_{t,\ell}, H_t)$
  for $a = 1, \ldots, K$ **do**
    Draw $\theta_{t,a}$ according to $p(\theta_a \mid \psi_{t,1}, H_{t,a})$
  Select action $A_t = \mathrm{argmax}_{a \in [K]} \, r(X_t, a; \theta_t)$, where $\theta_t = (\theta_{t,a})_{a \in [K]}$
  Observe reward $R_t \sim p(\cdot \mid X_t; \theta_{*,A_t})$ and update the posteriors.

All posterior expressions above use known quantities $(f_\ell, \Sigma_\ell, p(r \mid x; \theta))$. However, these expressions typically need to be approximated, except when the link functions f_ℓ are linear and the reward distribution $p(\cdot \mid x; \theta)$ is linear-Gaussian, where closed-form solutions can be obtained with careful derivations. These approximations are not trivial, and prior studies often rely on computationally intensive approximate sampling algorithms. In the following sections, we explain how we derive our efficient approximations which are motivated by the closed-form solutions of linear instances.

4.2.1 Posterior Approximation

The reward distribution is parameterized as a generalized linear model (GLM) (McCullagh and Nelder, 1989), which allows for non-linear rewards. In addition, the diffusion model itself is highly non-linear due to the link functions f_ℓ. These two sources of non-linearity make the posterior intractable, so we apply two layers of approximation: (i) a likelihood approximation to linearize the reward model, and (ii) a diffusion approximation to handle the non-linear hierarchy induced by the diffusion model prior.

(i) Likelihood approximation. We use an approach similar to the Laplace approximation, but instead of approximating the entire posterior, we approximate only the likelihood by a Gaussian. Precisely, the reward distribution $p(\cdot \mid x; \theta_a)$ belongs to the exponential family with mean function g. Thus

$$\prod_{i \in S_{t,a}} p(R_i \mid X_i; \theta_a) \;\approx\; \mathcal{N}\big(\theta_a; \hat{B}_{t,a}, \hat{G}_{t,a}^{-1}\big), \tag{4.7}$$

where $\hat{B}_{t,a}$ is the maximum likelihood estimate and $\hat{G}_{t,a}$ is the Hessian of the negative log-likelihood:

$$\hat{B}_{t,a} = \operatorname*{argmax}_{\theta_a \in \mathbb{R}^d} \sum_{i \in S_{t,a}} \log p(R_i \mid X_i; \theta_a), \qquad \hat{G}_{t,a} = \sum_{i \in S_{t,a}} \dot{g}\big(X_i^\top \hat{B}_{t,a}\big) X_i X_i^\top, \tag{4.8}$$

and $S_{t,a} = \{\ell \in [t-1] : A_\ell = a\}$ is the set of rounds in which action a was selected. Of course, $\hat{G}_{t,a}$ might not be invertivle and thus we replace it by $\hat{G}_{t,a} + 10^{-3} I_d$ in practice. Unlike Laplace, which fits a global Gaussian to the full posterior, this step linearizes only the likelihood, thereby preserving the hierarchical diffusion structure of the prior.

(ii) Diffusion approximation. Plugging the Gaussian likelihood approximation (4.7) into the posterior expressions $p(\theta_a \mid \psi_1, H_{t,a})$ and $p(\psi_{\ell-1} \mid \psi_\ell, H_t)$ removes the non-linearity of the reward model. However, the diffusion hierarchy remains non-linear through f_ℓ. To handle this, we build on the closed-form posteriors of the *linear diffusion case* (where $f_\ell(\psi_\ell) = \mathrm{W}_\ell \psi_\ell$; see Section B.1) and generalize them by replacing the linear terms $\mathrm{W}_\ell \psi_\ell$ with their non-linear counterparts $f_\ell(\psi_\ell)$. This substitution yields a *posterior diffusion model* that retains the same hierarchical form as the prior but with data-dependent means and covariances for the conditional Gaussians. Details on how we transition from the linear to the general non-linear setting are provided in Sections B.1 and B.2. The resulting approximate posteriors admit the following closed-form expressions.

Approximate action posterior. We approximate the conditional action posterior as

$$p(\theta_a \mid \psi_1, H_{t,a}) \;\approx\; \mathcal{N}\big(\theta_a; \hat{\mu}_{t,a}, \hat{\Sigma}_{t,a}\big),$$

where

$$\hat{\Sigma}_{t,a}^{-1} = \underbrace{\Sigma_1^{-1}}_{\text{prior precision}} + \underbrace{\hat{G}_{t,a}}_{\text{data precision}}, \qquad \hat{\mu}_{t,a} = \hat{\Sigma}_{t,a}\Big(\underbrace{\Sigma_1^{-1} f_1(\psi_1)}_{\text{prior contribution}} + \underbrace{\hat{G}_{t,a}\hat{B}_{t,a}}_{\text{data contribution}} \Big). \tag{4.9}$$

This posterior update has a clear interpretation. The posterior precision $\hat{\Sigma}_{t,a}^{-1}$ is the sum of the prior precision and the *data precision*. The posterior mean $\hat{\mu}_{t,a}$ is the precision-weighted average of the prior mean and the MLE $\hat{B}_{t,a}$. As more data are observed, the covariance shrinks and the mean moves from the prior mean $f_1(\psi_1)$ toward the MLE $\hat{B}_{t,a}$. When no data are available ($\hat{G}_{t,a} = 0$), the posterior reduces to the prior $\mathcal{N}(f_1(\psi_1), \Sigma_1)$; in the limit of infinite data ($\hat{G}_{t,a} \to \infty$), the posterior collapses to the MLE $\hat{B}_{t,a}$, with $\hat{\mu}_{t,a} \to \hat{B}_{t,a}$ and $\hat{\Sigma}_{t,a} \to 0$.

Approximate latent posteriors. For each $\ell \in [L+1] \setminus \{1\}$, we approximate the latent posterior as

$$p(\psi_{\ell-1} \mid \psi_\ell, H_t) \;\approx\; \mathcal{N}\big(\psi_{\ell-1}; \bar{\mu}_{t,\ell-1}, \bar{\Sigma}_{t,\ell-1}\big),$$

with

$$\bar{\Sigma}_{t,\ell-1}^{-1} = \underbrace{\Sigma_\ell^{-1}}_{\text{prior precision}} + \underbrace{\bar{G}_{t,\ell-1}}_{\text{data precision}}, \quad \bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\Big(\underbrace{\Sigma_\ell^{-1} f_\ell(\psi_\ell)}_{\text{prior contribution}} + \underbrace{\bar{B}_{t,\ell-1}}_{\text{data contribution}} \Big), \tag{4.10}$$

where, by convention, $f_{L+1}(\psi_{L+1}) = 0$ since the top layer ψ_L has no parent ψ_{L+1}. The quantities $\bar{G}_{t,\ell}$ and $\bar{B}_{t,\ell}$ are computed recursively. The base recursion is

$$\bar{G}_{t,1} = \sum_{a=1}^{K} \left(\Sigma_1^{-1} - \Sigma_1^{-1}\hat{\Sigma}_{t,a}\Sigma_1^{-1}\right), \qquad \bar{B}_{t,1} = \Sigma_1^{-1} \sum_{a=1}^{K} \hat{\Sigma}_{t,a}\hat{G}_{t,a}\hat{B}_{t,a}, \tag{4.11}$$

and for each $\ell \in [L] \setminus \{1\}$,

$$\bar{G}_{t,\ell} = \Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}, \qquad \bar{B}_{t,\ell} = \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\bar{B}_{t,\ell-1}. \tag{4.12}$$

The latent posterior update in Equation (4.10) has the same structure as the action posterior. The posterior precision $\bar{\Sigma}_{t,\ell-1}^{-1}$ is the sum of the prior and data precisions , and the posterior mean is their precision-weighted combination. The data terms $\bar{G}_{t,\ell-1}$ and $\bar{B}_{t,\ell-1}$ are computed recursively (Equations (4.11) and (4.12)), so information collected at the action level propagates upward through the hierarchy.

Interpretation. The resulting approximate posterior remains a diffusion model whose conditional Gaussians have updated, data-dependent means and covariances. The latent-posterior means can be viewed as *refined link functions*:

$$\hat{f}_{t,\ell}(\psi_\ell) = \bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\left(\Sigma_\ell^{-1} f_\ell(\psi_\ell) + \bar{B}_{t,\ell-1}\right),$$

and $\bar{\Sigma}_{t,\ell}$ represents their updated uncertainty. Both are updated with data: covariances contract as uncertainty decreases, and means move from the prior toward the MLE. Unlike a full Laplace approximation, this formulation preserves the expressiveness of the posterior rather than replacing it globally with a single Gaussian, while also avoiding the heavy computation required by other approximate inference methods.

4.2.2 Extension to Joint Reward Models

For the shared-parameter model in Remark 5, `dTS`'s posterior approximations are similar. The action posterior is $p(\theta \mid \psi_1, H_t) \approx \mathcal{N}(\hat{\mu}_t, \hat{\Sigma}_t)$, where

$$\hat{\Sigma}_t^{-1} = \Sigma_1^{-1} + \hat{G}_t\,, \qquad \hat{\mu}_t = \hat{\Sigma}_t\left(\Sigma_1^{-1} f_1(\psi_1) + \hat{G}_t\hat{B}_t\right). \tag{4.13}$$

where

$$\hat{B}_t = \operatorname*{argmax}_{\theta\in\mathbb{R}^d} \sum_{i<t} \log p\left(R_i \mid \phi(X_i, A_i)^\top\theta\right), \quad \hat{G}_t = \sum_{i<t} \dot{g}\left(\phi(X_i, A_i)^\top\hat{B}_t\right)\phi(X_i, A_i)\phi(X_i, A_i)^\top.$$

Similarly, for $\ell \in [L+1] \setminus \{1\}$, the latent posterior is $p(\psi_{\ell-1} \mid \psi_\ell, H_t) \approx \mathcal{N}(\bar{\mu}_{t,\ell-1}, \bar{\Sigma}_{t,\ell-1})$, where

$$\bar{\Sigma}_{t,\ell-1}^{-1} = \Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}, \qquad \bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\left(\Sigma_\ell^{-1} f_\ell(\psi_\ell) + \bar{B}_{t,\ell-1}\right), \tag{4.14}$$

where, by convention, $f_{L+1}(\psi_{L+1}) = 0$ and the quantities $\bar{G}_{t,\ell}$ and $\bar{B}_{t,\ell}$ are computed recursively as

$$\text{Base case:} \qquad \bar{G}_{t,1} = \Sigma_1^{-1} - \Sigma_1^{-1}\hat{\Sigma}_t\Sigma_1^{-1}, \qquad \bar{B}_{t,1} = \Sigma_1^{-1}\hat{\Sigma}_t\hat{G}_t\hat{B}_t. \tag{4.15}$$

$$\text{Recursive case:} \qquad \bar{G}_{t,\ell} = \Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}, \qquad \bar{B}_{t,\ell} = \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\bar{B}_{t,\ell-1}. \tag{4.16}$$

Again, this shared-parameter variant of `dTS` is presented for completeness and to illustrate the generality of our posterior derivations; the main focus of the chapter remains on the per-action disjoint formulation in Equation (4.1). Unless stated otherwise, all theoretical results and experiments use the main version of `dTS` described in Algorithm 2.

4.3 Analysis

In this section, we present an informal Bayes regret analysis of `dTS` to build intuition around `dTS`'s Bayesian regret scaling with problem parameters d, K, L, etc. This analysis is informal for two reasons. First, we analyze a simplified linear-Gaussian setting rather than the general nonlinear case on which we focus in this chapter: the reward distribution is linear-Gaussian and each link function $f_\ell(\psi_\ell) = \mathrm{W}_\ell \psi_\ell$ is a known linear mapping, inducing a hierarchy of L linear-Gaussian layers from the latent root to the action parameters.

Second, we assume the model is well-specified (similar to Chapter 7): the true action parameters are generated according to the diffusion prior used by `dTS`. Under these assumptions, the posterior becomes exact, enabling an analysis analogous to that used in Chapter 3. However, our recursive hierarchical structure introduces technical differences: posteriors must be derived inductively using total covariance decompositions, and regret bounds require tracking information flow across all latent layers. We emphasize that this regret bound does not extend to the general nonlinear case studied in experiments; it is included here solely to provide theoretical intuition under simplifying assumptions. Formal statements and derivations are provided in Sections B.4 and B.5.

Bayes regret bound. The bound of `dTS` in this case is

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}\Big(\sqrt{T(dK\sigma_1^2 + d\sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \sigma_{\text{MAX}}^{2\ell})}\Big),$$

where $\sigma_{\text{MAX}}^2 = \max_{\ell \in [L+1]} 1 + \frac{\sigma_\ell^2}{\sigma^2}$. This can be re-written as $\mathcal{BR}(T) = \tilde{\mathcal{O}}(\sqrt{TdK_{\text{eff}} \sum_{\ell=1}^{L+1} \sigma_\ell^2})$, where $K_{\text{eff}} = \frac{K\sigma_1^2 + \sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \sigma_{\text{MAX}}^{2\ell}}{\sum_{\ell=1}^{L+1} \sigma_\ell^2}$ is the *effective number of actions*. This dependence on the horizon T aligns with prior Bayes regret bounds scaling with T. However, the bound comprises $L+1$ main terms. First, one relates to action parameters learning, conforming to a standard form (Lu and Van Roy, 2019), while the L remaining terms are associated with learning each of the latent parameters.

Sparsity refinement. If each mixing matrix exhibits column sparsity, that, $\mathrm{W}_\ell = (\bar{\mathrm{W}}_\ell, 0_{d,d-d_\ell})$ with $d_\ell \ll d$ active columns, then the bound becomes

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}\Big(\sqrt{T(dK\sigma_1^2 + \sum_{\ell=1}^{L} d_\ell \sigma_{\ell+1}^2 \sigma_{\text{MAX}}^{2\ell})}\Big).$$

Hence, informative, *sparse* priors can cut the cost of learning deep latent chains down from d to d_ℓ. As in Chapter 3, a less informative prior (such as high variance) leads to a more challenging problem and thus a higher bound. Therefore, smaller values of K, L, d, d_ℓ translate to fewer parameters to learn, leading to lower regret. The regret also decreases when the initial variances σ_ℓ^2 decrease. These dependencies are common in Bayesian analysis, and empirical results match them.

Dependence on K. The reader may question why our bound depends on K. This dependence arises from two modeling choices. First, we study the *disjoint* (per-action) setting

$r(x, a; \theta) = x^\top \theta_a$, where $\theta = (\theta_a)_{a \in [K]} \in \mathbb{R}^{dK}$, requiring the learning of Kd parameters. Second, we model the relationship between θ_a and ψ_1 stochastically as $\mathcal{N}(W_1 \psi_1, \sigma_1^2 I_d)$ to accommodate potential nonlinearity. While this choice confers robustness to model misspecification, it introduces additional uncertainty and requires learning both the action parameters θ_a and the latent parameters ψ_ℓ, resulting in a bound that depends on both K and L.

Despite this dependence, `dTS` enjoys two key advantages. First, the regret scales with $K\sigma_1^2$ rather than $K \sum_\ell \sigma_\ell^2$, which is particularly beneficial when σ_1 is small, as is often the case with diffusion model priors. Second, thanks to informative priors, our bound has significantly smaller constants compared to both the Bayesian and frequentist regret bounds for `LinTS`. We demonstrate this empirically in Section B.6.5 and provide a theoretical comparison in Section 4.3.1. Both analyses confirm that `dTS`'s advantage over `LinTS` increases as the action space grows.

Can regret be independent of K? Prior works (Foster et al., 2020; Xu and Zeevi, 2020; Zhu et al., 2022) have proposed bandit algorithms whose regret does not scale with K. However, these results apply to the *shared-parameter* setting $r(x, a; \theta) = \phi(x, a)^\top \theta$, where only a single d-dimensional parameter must be learned, but this formulation requires access to a suitable feature map ϕ. `dTS` is compatible with this setting (Section 4.2.2), in which case its regret would indeed be independent of K. Alternatively, even in the disjoint per-action case considered in this chapter, setting $\sigma_1 = 0$ would yield a K-independent regret bound. However, we believe this assumption is unrealistic in practice and would compromise the robustness of `dTS` to model misspecification.

4.3.1 Benefits

Computational benefits. Action correlations prompt an intuitive approach: marginalize all latent parameters and maintain a joint posterior of $(\theta_a)_{a \in [K]} \mid H_t$. Unfortunately, this is computationally inefficient for large action spaces. To illustrate, suppose that all posteriors are multivariate Gaussians. Then maintaining the joint posterior $(\theta_a)_{a \in [K]} \mid H_t$ necessitates converting and storing its $dK \times dK$-dimensional covariance matrix, leading to $\mathcal{O}(K^3 d^3)$ and $\mathcal{O}(K^2 d^2)$ time and space complexities. In contrast, the time and space complexities of `dTS` are $\mathcal{O}\big((L+K)d^3\big)$ and $\mathcal{O}\big((L+K)d^2\big)$. This is because `dTS` requires converting and storing $L + K$ covariance matrices, each being $d \times d$-dimensional. The improvement is huge when $K \gg L$, which is common in practice. Certainly, a more straightforward way to enhance computational efficiency is to discard latent parameters and maintain K individual posteriors, each relating to an action parameter $\theta_a \in \mathbb{R}^d$ (`LinTS`). This improves time and space complexity to $\mathcal{O}\big(Kd^3\big)$ and $\mathcal{O}\big(Kd^2\big)$. However, `LinTS` maintains independent posteriors and fails to capture the correlations among actions; it only models $\theta_a \mid H_{t,a}$ rather than $\theta_a \mid H_t$ as done by `dTS`. Consequently, `LinTS` incurs higher regret due to the information loss caused by unused interactions of similar actions. Our regret bound and empirical results reflect this aspect.

Statistical benefits. We argue that our bound reflects the overall structure of the problem by comparing `dTS` to algorithms that only partially use the structure or do not use it at all as follows. Precisely, when the link functions are linear, we can transform the diffusion prior into a Bayesian linear model (`LinTS`) by marginalizing out the latent

parameters; in which case the prior on action parameters becomes $\theta_a \sim \mathcal{N}(0, \Sigma)$, with the θ_a being not necessarily independent, and Σ is the marginal initial covariance of action parameters and it writes $\Sigma = \sigma_1^2 I_d + \sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top$ with $\mathrm{B}_\ell = \prod_{i=1}^{\ell} \mathrm{W}_i$. Then, it is tempting to directly apply `LinTS` to solve our problem. This approach will induce higher regret because the additional uncertainty of the latent parameters is accounted for in Σ despite integrating them. This causes the *marginal* action uncertainty Σ to be much higher than the *conditional* action uncertainty $\sigma_1^2 I_d$, since we have $\Sigma = \sigma_1^2 I_d + \sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top \succcurlyeq \sigma_1^2 I_d$. This discrepancy leads to higher regret, especially when K is large. This is due to `LinTS` needing to learn K independent d-dimensional parameters, each with a considerably higher initial covariance Σ. This is also reflected by our regret bound. To simply comparisons, suppose that $\sigma \geq \max_{\ell \in [L+1]} \sigma_\ell$ so that $\sigma_{\text{MAX}}^2 \leq 2$. Then the regret bounds of `dTS` (where we bound $\sigma_{\text{MAX}}^{2\ell}$ by 2^ℓ) and `LinTS` read

$$\texttt{dTS} : \tilde{\mathcal{O}}\Big(\sqrt{T(dK\sigma_1^2 + \sum_{\ell=1}^{L} d_\ell \sigma_{\ell+1}^2 2^\ell)}\Big), \qquad \texttt{LinTS} : \tilde{\mathcal{O}}\Big(\sqrt{TdK(\sigma_1^2 + \sum_{\ell=1}^{L} \sigma_{\ell+1}^2)}\Big).$$

Then regret improvements are captured by the variances σ_ℓ and the sparsity dimensions d_ℓ, and we proceed to illustrate this through the following scenarios.

(I) Decreasing variances. Assume that $\sigma_\ell = 2^\ell$ for any $\ell \in [L+1]$. Then, the regrets become

$$\texttt{dTS} : \tilde{\mathcal{O}}\Big(\sqrt{T(dK + \sum_{\ell=1}^{L} d_\ell 4^\ell)}\Big), \qquad \texttt{LinTS} : \tilde{\mathcal{O}}\Big(\sqrt{TdK2^L}\Big)$$

Now to see the order of gain, assume the problem is high-dimensional ($d \gg 1$), and set $L = \log_2(d)$ and $d_\ell = \lfloor \frac{d}{2^\ell} \rfloor$. Then the regret of `dTS` becomes $\tilde{\mathcal{O}}\big(\sqrt{nd(K+L)}\big)$, and hence the multiplicative factor 2^L in `LinTS` is removed and replaced with a smaller additive factor L.

(II) Constant variances. Assume that $\sigma_\ell = 1$ for any $\ell \in [L+1]$. Then, the regrets become

$$\texttt{dTS} : \tilde{\mathcal{O}}\Big(\sqrt{T(dK + \sum_{\ell=1}^{L} d_\ell 2^\ell)}\Big), \qquad \texttt{LinTS} : \tilde{\mathcal{O}}\Big(\sqrt{TdKL}\Big)$$

Similarly, let $L = \log_2(d)$, and $d_\ell = \lfloor \frac{d}{2^\ell} \rfloor$. Then `dTS`'s regret is $\tilde{\mathcal{O}}\big(\sqrt{Td(K+L)}\big)$. Thus the multiplicative factor L in `LinTS` is removed and replaced with the additive factor L. By comparing this to **(I)**, the gain with decreasing variances is greater than with constant ones. In general, diffusion models use decreasing variances (Ho et al., 2020) and hence we expect great gains in practice. All observed improvements in this section could become even more pronounced when employing non-linear diffusion models. In our theory, we used linear diffusion models, and yet we can already discern substantial differences. Moreover, under non-linear diffusion Equation (4.1), the latent parameters cannot be analytically marginalized, making `LinTS` with exact marginalization inapplicable.

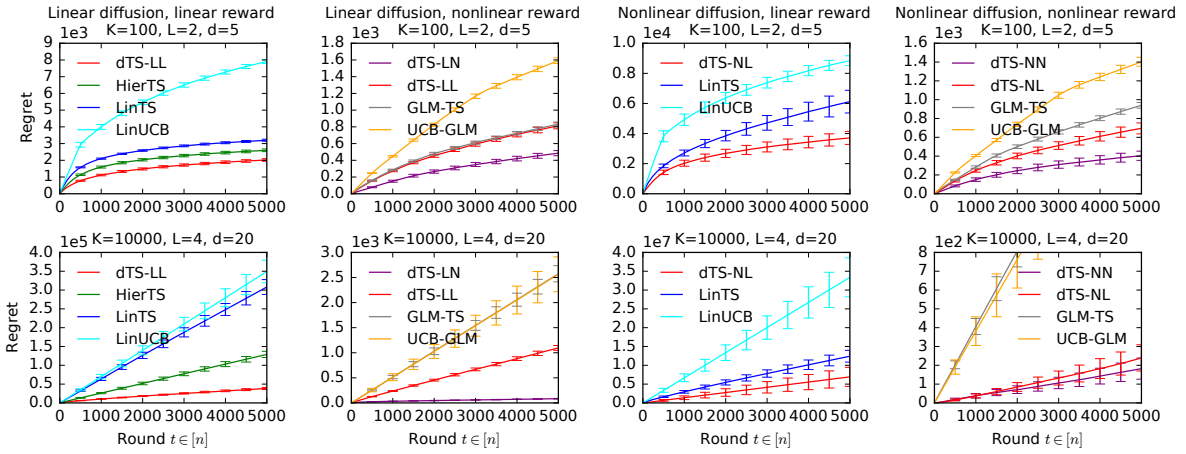


Figure 4.1: Regret of `dTS` with varying diffusion and reward models and varying parameters d, K, L.

4.4 Experiments

Experimental setup. We evaluate `dTS` using both synthetic and MovieLens problems. In our experiments, we run 50 random simulations and plot the average regret with standard error. Our main contribution is to demonstrate that pretraining a diffusion model offline enables the construction of expressive and informative priors that substantially improve exploration efficiency in contextual bandits. We first evaluate `dTS` in a setting where the prior matches the true generative process (Section 4.4.1 to isolate the benefit of informative priors), and then consider a misspecified regime (Section 4.4.2 and Section B.6) where the prior is either trained on out-of-distribution data or intentionally perturbed. These experiments show that even when the prior is imperfect, `dTS` maintains strong performance: highlighting its robustness and practical relevance.

4.4.1 True Prior is a Diffusion Model

Synthetic bandit problems are generated from the diffusion model in Equation (4.1) with both linear and non-linear rewards. Linear rewards follow $p(\cdot \mid x; \theta_a) = \mathcal{N}(x^\top \theta_a, 1)$, while non-linear rewards are binary from $p(\cdot \mid x; \theta_a) = \text{Ber}(g(x^\top \theta_a))$, with g as the sigmoid function. Covariances are $\Sigma_\ell = I_d$, and contexts X_t are uniformly drawn from $[-1, 1]^d$. We vary $d \in \{5, 20\}$, $L \in \{2, 4\}$, $K \in \{10^2, 10^4\}$, and set the horizon to $T = 5000$, considering both linear and non-linear models.

Linear diffusion. We consider Equation (4.1) with $f_\ell(\psi) = \mathrm{W}_\ell \psi$, where W_ℓ uniformly drawn from $[-1, 1]^{d \times d}$. Sparsity is introduced by zeroing the last d_ℓ columns of W_ℓ as $\mathrm{W}_\ell = (\bar{\mathrm{W}}_\ell, 0_{d, d-d_\ell})$. For $d = 5$ and $L = 2$, $(d_1, d_2) = (5, 2)$; for $d = 20$ and $L = 4$, $(d_1, d_2, d_3, d_4) = (20, 10, 5, 2)$.

Non-linear diffusion. We consider Equation (4.1) where f_ℓ are 2-layer neural networks with random weights in $[-1, 1]$, `ReLU` activation, and hidden layers of size $h = 20$ for $d = 5$, and $h = 60$ for $d = 20$.

Baselines. For linear rewards, we use `LinUCB` (Abbasi-Yadkori et al., 2011), `LinTS` (Agrawal and Goyal, 2013a), and `HierTS` (Hong et al., 2022b), marginalizing out all latent parameters except ψ_L, which corresponds to `HierTS-1` in Section B.3. For non-linear rewards, we include `UCB-GLM` (Li et al., 2017) and `GLM-TS` (Chapelle and Li, 2012). We exclude `GLM-UCB` (Filippi et al., 2010) due to high regret and `HierTS` as it's designed for linear rewards. We name `dTS` as `dTS-dr`, where `d` refers to diffusion type (`L` for linear, `N` for non-linear) and `r` indicates reward type (`L` for linear, `N` for non-linear). For example, `dTS-LL` signifies `dTS` in linear diffusion with linear rewards.

Results and interpretations. Results are shown in Figure 4.1 and we make the following observations:

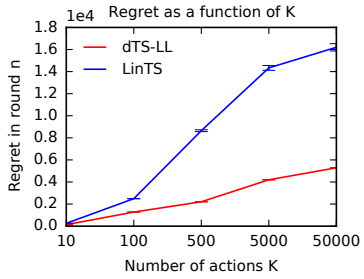
1) `dTS` demonstrates superior performance (Figure 4.1). `dTS` consistently outperforms the baselines across all settings, including the four combinations of linear/non-linear diffusion and reward (columns in Figure 4.1) and both bandit settings with varying K, L, and d (rows in Figure 4.1).

2) Latent diffusion structure may be more important than the reward distribution. When rewards are non-linear (second and fourth columns in Figure 4.1), we include variants of `dTS` that use the correct diffusion prior but the wrong reward distribution, applying linear-Gaussian instead of logistic-Bernoulli (`dTS-LL` in the second column and `dTS-NL` in the fourth). Despite the reward misspecification, these variants outperform models using the correct reward distribution but ignoring the latent diffusion structure, such as `GLM-TS` and `UCB-GLM`. This highlights the importance of accounting for latent structure, which can be more critical than an accurate reward distribution.

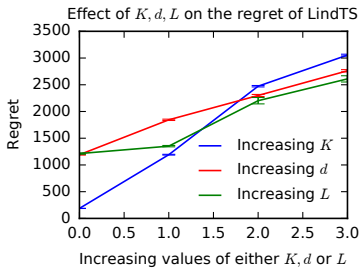
3) Performance gap between `dTS` and `LinTS` widens as K increases (Figure 4.2a). To show `dTS`'s improved scalability, we evaluate its performance with varying values of $K \in [10, 5 \times 10^4]$, in the linear diffusion and rewards setting. Figure 4.2a shows the final cumulative regret for varying K values for both `dTS-LL` and `LinTS`, revealing a widening performance gap as K increases.

4) Regret scaling with K, d and L matches our theory (Figure 4.2b). We assess the effect of the number of actions K, context dimension d, and diffusion depth L on `dTS`'s regret. Using the linear diffusion and rewards setting, for which we have derived a Bayes regret upper bound, we plot `dTS-LL`'s regret across varying values of $K \in \{10, 100, 500, 1000\}$, $d \in \{5, 10, 15, 20\}$, and $L \in \{2, 4, 5, 6\}$ in Figure 4.2b. As predicted by our theory, the empirical regret increases with larger values of K, d, or L, as these make the learning problem more challenging, leading to higher regret.

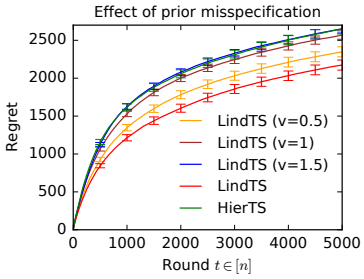
5) Diffusion prior misspecification (Figure 4.2c). Here, `dTS`'s diffusion prior parameters differ from the true diffusion prior. In the linear diffusion and reward setting, we replace the true parameters W_ℓ and Σ_ℓ with misspecified ones, $\mathrm{W}_\ell + \epsilon_1$ and $\Sigma_\ell + \epsilon_2$, where ϵ_1 and ϵ_2 are uniformly sampled from $[v, v+0.5]^{d \times d}$, with v controlling the misspecification level. We vary $v \in \{0.5, 1, 1.5\}$ and assess `dTS`'s performance, comparing it to the well-specified `dTS-LL` and the strongest baseline in this fully-linear setting, `HierTS`. As shown in Figure 4.2c, `dTS`'s performance decreases with increasing misspecification but remains superior to the baseline, except at $v = 1.5$, where their performances are comparable. Additional misspecification experiments are presented in Section 4.4.2, where the bandit



(a) Perf. gap w.r.t. K.

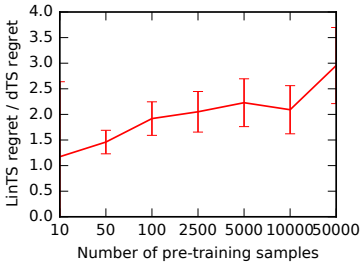


(b) Scaling w.r.t. K, d, L.

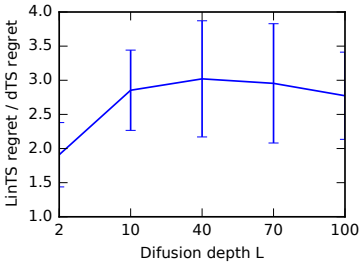


(c) Prior misspecification.

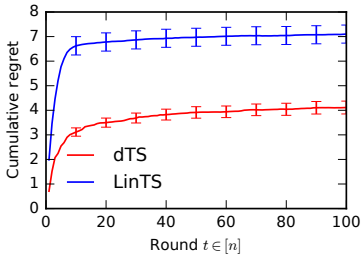
Figure 4.2: Effect of various factors on dTS's performance.



(a) Ratio of LinTS/dTS cumulative regret in the last round with varying pre-training sample size in $[10, 5\times10^4]$. **Higher values mean a bigger performance gap**.



(b) Ratio of LinTS/dTS cumulative regret in the last round with varying diffusion depth L in $[2, 100]$. **Higher values mean a bigger performance gap**.



(c) Regret of dTS in **MovieLens**. The diffusion model with $L = 40$ is pre-trained on embeddings obtained by low-rank factorization of MovieLens rating matrix.

Figure 4.3: **(a) and (b):** Impact of pre-training sample size and diffusion depth L for the **Swiss roll data**. **(c):** Regret of dTS in **MovieLens**.

environment is not sampled from a diffusion model.

4.4.2 True Prior is Not a Diffusion Model

Swiss roll data. Unlike previous experiments, the true action parameters are now sampled from the Swiss roll distribution (see Figure B.1 in Section B.6.1), rather than from a diffusion model. The diffusion model used by dTS is pre-trained on samples from this distribution, with the offline pre-training procedure described in Section B.6.2. Figure 4.3a shows that larger sample sizes increase the performance gap between dTS and LinTS. More samples improve the estimation of the diffusion prior (see Figure B.1 in Section B.6.1), leading to better dTS performance. Notably, comparable performance was achieved with as few as 10 samples, and dTS outperformed LinTS by a factor of 1.5 with just 50 samples. While more samples may be required for more complex problems, LinTS would also struggle in such cases. Therefore, we expect these gains to be even more significant in more challenging settings.

We studied the effect of the pre-trained diffusion model depth L and found that $L \approx 40$

yields the best performance, with a drop beyond that point (Figure 4.3b). While our theory doesn't apply directly here, as it assumes a linear diffusion model, it still offers some intuition on the decreased performance for $L > 40$. The theorem shows `dTS`'s regret bound increases with L when the true distribution is a diffusion model. For small L, the pre-trained model doesn't fully capture the true distribution, making the theorem inapplicable, but at $L \approx 40$, the distribution is nearly captured, and further increases in L lead to higher regret, consistent with our theory.

MovieLens data. We also evaluate `dTS` using the standard MovieLens (Lam and Herlocker, 2016) setting. In this semi-synthetic experiment, a user is sampled from the rating matrix in each interaction round, and the reward is the rating the user gives to a movie (see Clavier et al. (2023, Section 5) for details about this setting). Here, the true distribution of action parameters is unknown and not a diffusion model. The diffusion model is pre-trained on offline estimates of action parameters obtained through low-rank factorization of the rating matrix. Figure 4.3c demonstrates that `dTS` outperforms `LinTS` in this setting. Additional **CIFAR** ablations are provided in Section B.6.4 where similar strong improvements are observed.

4.5 Conclusion

We use a pre-trained diffusion model as a strong and flexible prior for `dTS`. Diffusion pre-training leverages abundant offline data, which is then fine-tuned through online interactions via our tractable posterior approximation. This approximation enables efficient posterior sampling and updates while maintaining strong empirical performance. Moreover, `dTS` admits a simple Bayesian regret bound in the linear–Gaussian setting.

Our work has several limitations. First, our Bayes regret analysis applies only to the linear-Gaussian setting with a well-specified prior; extending formal guarantees to nonlinear diffusion models remains open. Second, our posterior approximation, while motivated by exact solutions in the linear case, lacks theoretical justification for general nonlinear link functions: its strong empirical performance does not come with formal approximation error bounds. Finally, `dTS` requires offline pre-training of the diffusion model, which assumes access to historical estimates of action parameters; in domains where such data is unavailable or expensive to obtain, the benefits of diffusion priors may not be realized.

Part II

Off-Policy Learning in Large Action Spaces

Chapter 5

Introduction to Part II

This second part of the thesis addresses the following fundamental question:

How can we reliably learn high-performing policies from static logged data when the number of actions is large?

5.1 Setting and Background

In this part, we consider the off-policy (offline) setting where an agent is provided with a static logged dataset $\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$ collected by a logging policy π_0. The data collection process proceeds as follows: for each round $i \in [n]$:

1. The environment draws a context $X_i \sim \nu$, where ν is a distribution with support $\mathcal{X}$ forming a compact subset of $\mathbb{R}^d$;
2. The logging policy selects an action $A_i \sim \pi_0(\cdot \mid X_i)$ from the action set $\mathcal{A} = [K]$;
3. The environment generates a stochastic reward $R_i \sim p(\cdot \mid X_i, A_i)$, where $R_i \in [0, 1]$.

Unlike the on-policy setting of Part I, no further interaction with the environment is permitted. The objective is to learn a new policy $\hat{\pi}$ from this static dataset that maximizes the true (but unknown) expected value:

$$V(\pi) = \mathbb{E}_{X\sim\nu}\left[\mathbb{E}_{A\sim\pi(\cdot|X)}[r(X, A)]\right], \tag{5.1}$$

where $r(x, a) = \mathbb{E}_{R\sim p(\cdot|x,a)}[R]$ is the expected reward function. Performance is measured by the *suboptimality gap* of the learned policy:

$$\text{SO}(\hat{\pi}) = V(\pi_*) - V(\hat{\pi}), \tag{5.2}$$

where $\pi_* = \arg\max_{\pi\in\Pi} V(\pi)$ is the optimal policy within a class Π.

Since $V(\pi)$ cannot be computed directly, off-policy learning algorithms often rely on an empirical estimate $\hat{V}(\pi)$ constructed from $\mathcal{D}_n$. This estimation task is known as *off-policy evaluation (OPE)* in the literature. The two dominant estimation approaches are: direct

method (DM) and inverse propensity scoring (IPS). DM employ a learned reward model $\hat{r}(x, a)$ to estimate the value as:

$$\hat{V}_{\text{DM}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \pi(a \mid X_i)\, \hat{r}(X_i, a)\,. \tag{5.3}$$

IPS re-weights observed rewards using importance sampling as:

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} R_i. \tag{5.4}$$

Given an estimator $\hat{V}(\pi)$, the agent must then select a policy. This step distinguishes between *greedy policies*, which directly maximize $\hat{\pi} = \arg\max_{\pi \in \Pi} \hat{V}(\pi)$, and *pessimistic policies*, which incorporate an uncertainty penalty $\hat{\pi} = \arg\max_{\pi \in \Pi} [\hat{V}(\pi) - \text{pen}(\pi)]$.

5.1.1 Scalability Challenges

When the number of actions K is large, both estimation paradigms and their associated optimization procedures encounter fundamental obstacles:

Statistical inefficiency of DM. Standard DM approaches model each action's reward function independently. As K grows, the data available per action diminishes, leading to poorly estimated reward functions and lower performance.

High variance of importance sampling. IPS's variance grows with the importance weights $\pi(a|x)/\pi_0(a|x)$. These weights explode in large action spaces, producing estimates too noisy for reliable optimization.

Intractable optimization landscapes. Beyond estimation challenges, the optimization problem $\arg\max_{\pi \in \Pi} \hat{V}(\pi)$ itself becomes computationally intractable in large action spaces. IPS-based objectives induce highly non-concave landscapes with exponentially many local maxima and flat plateaus that trap gradient-based optimizers. As we show in Chapter 7, this optimization bottleneck often dominates estimation error, making even statistically superior estimators ineffective in practice.

5.2 Methodological Approaches

To address these challenges, the methods developed in this part pursue three complementary strategies: *structured reward modeling* for sample-efficient DM, *surrogate objectives* that prioritize optimization tractability over estimation accuracy, and *principled regularization and pessimism* for importance-weighted estimators.

Structured Bayesian models. Drawing inspiration from the hierarchical framework of Part I, we introduce latent structure into reward modeling. Action parameters are coupled through shared latent variables ψ:

$$\begin{aligned} \psi &\sim q(\cdot), \\ \theta_a \mid \psi &\sim p_a(\cdot; f_a(\psi)), \quad \forall a \in \mathcal{A}, \\ R \mid X, A, \theta, \psi &\sim p(\cdot \mid X; \theta_A). \end{aligned} \tag{5.5}$$

This formulation enables information sharing across actions: observations from frequently selected actions inform the posterior over ψ, which in turn improves reward estimates for rarely observed actions.

Optimization-aware objectives. Rather than designing sophisticated value estimators and then optimizing them, we propose objectives designed primarily for favorable optimization landscapes. The *policy-weighted log-likelihood (PWLL)* family:

$$\hat{U}_g(\pi) = \frac{1}{n} \sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i)) \log \pi(A_i \mid X_i)\,, \tag{5.6}$$

where g is a positive weighting function, yields concave objectives for linear-softmax policies π. This guarantees efficient convergence to a unique global maximum, bypassing the optimization pathologies of value estimation altogether.

Regularized importance weighting. For practitioners committed to IPS-based methods, we develop variance-controlled estimators through importance weight regularization. The exponential smoothing estimator:

$$\hat{V}^{\alpha}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)^{\alpha}} R_i\,, \quad \alpha \in [0, 1]\,, \tag{5.7}$$

smoothly trades variance for bias while preserving differentiability. Combined with pessimistic optimization and PAC-Bayes generalization bounds, this yields principled, tractable objectives that are amenable to stochastic gradient ascent for safe off-policy learning.

5.3 Roadmap of Part II

The following chapters develop these methodological approaches.

Chapter 6: Scaling Direct Methods with Latent Parameters. We begin by addressing the statistical inefficiency of DM through structured Bayesian modeling. Building on the hierarchical framework of Part I, we introduce the *structured direct method* (`sDM`), which couples action parameters through a shared latent vector. The posterior over these latent variables aggregates evidence across all actions, enabling effective generalization to actions with sparse data coverage. We analyze performance through *Bayesian suboptimality* and prove that greedy policies paired with `sDM` achieve $\mathcal{O}(1/\sqrt{n})$ convergence under mild assumptions on the alignment between logging and optimal policies.

Chapter 7: Optimization Matters More Than Estimation. We then challenge the conventional paradigm of off-policy learning. Through theoretical analysis and large-scale experiments, we demonstrate that *optimization error* dominates estimation error in large action spaces. Specifically, we prove that for any IPS-based estimator, gradient ascent can remain trapped in suboptimal regions for $\mathcal{O}(K)$ iterations, and that the optimization landscape contains exponentially many local maxima in K. We then propose *objective-aware policy parametrizations*: by aligning the policy class with the estimator's inductive bias, we can partially mitigate these optimization challenges. However, for a

more complete solution, we propose *policy-weighted log-likelihood (PWLL)* objectives as an alternative to IPS-based objectives. These objectives are provably concave for linear softmax policies, guaranteeing efficient convergence to a global optimum. Experiments on datasets with up to one million actions validate that PWLL consistently outperforms state-of-the-art estimator-based methods.

Chapter 8: Principled Pessimism for Exponential Smoothing and Beyond. Finally, for practitioners committed to importance weighting methods, we develop a theoretically grounded framework for variance control and pessimistic policy learning. We propose *exponential smoothing* estimators that regularize importance weights, trading controlled bias for reduced variance. To leverage these regularized estimators for safe policy learning, we derive two-sided PAC-Bayes generalization bounds where all quantities are empirical and differentiable. The pessimistic learning objective maximizes the lower bound on policy value, penalizing policies with high bias or variance and steering optimization toward reliable regions. This also yields tractable objectives amenable to stochastic gradient optimization. We further present a *unified PAC-Bayes framework* covering the major importance weight regularization techniques in the literature (clipping, exponential smoothing, implicit exploration), enabling principled comparison and demonstrating that the choice of pessimistic objective often matters more than the specific regularizer.

Chapter 6

Scaling Direct Methods with Latent Parameters

Contents

In this chapter, we address the statistical inefficiency of direct methods (DMs) in large action spaces: the first challenge highlighted in Chapter 5. Standard DMs estimate independent d-dimensional parameters for each action. This approach becomes statistically inefficient in large action spaces, where data are collected by a logging policy that explores only a small subset of available actions, leaving many actions rarely or never observed. To overcome this limitation, we analyze DMs through a Bayesian lens and propose making them sample-efficient by incorporating informative priors.

We make the following contributions. **1)** We introduce the *structured direct method* (`sDM`), a Bayesian approach that uses informative priors to share reward information across actions. By updating beliefs about similar actions based on observed data, `sDM` improves statistical efficiency without compromising computational scalability. **2)** To evaluate `sDM`, we propose Bayesian metrics that assess the average performance across problem instances sampled from the prior. This departs from the standard frequentist focus on worst-case scenarios. These metrics formally quantify the benefits of informative priors. **3)** Our theoretical analysis of Bayesian suboptimality (BSO) reveals two key insights: (1) performance degrades gracefully even without the standard assumption of full logging support, and (2) greedy policies are provably optimal under the BSO metric, standing in contrast to the pessimistic policies typically favored in frequentist settings. **4)** We empirically validate `sDM` and our theoretical findings using both synthetic and real-world datasets.

6.1 Setting

We consider the setting described in Section 5.1. The only additional assumption is the existence of *unknown true parameters* $\theta_{*,a} \in \mathbb{R}^d$ for each action a, such that rewards are distributed as $R_i \sim p(\cdot \mid X_i; \theta_{*,A_i})$. Let $\theta_* = (\theta_{*,a})_{a \in \mathcal{A}} \in \mathbb{R}^{dK}$ denote the concatenation of all action parameters. The *reward function* $r(x, a; \theta_*) = \mathbb{E}_{R \sim p(\cdot|x;\theta_{*,a})}[R]$ gives the expected reward of action a in context x. The goal is to find a policy $\pi \in \Pi$ that maximizes:

$$V(\pi; \theta_*) = \mathbb{E}_{X \sim \nu} \mathbb{E}_{A \sim \pi(\cdot|X)}[r(X, A; \theta_*)] \,.$$

This chapter focuses on DM that estimates the value $V(\pi; \theta_*)$ as:

$$\hat{V}_{\mathrm{DM}}(\pi) = \frac{1}{n} \sum_{i \in [n]} \sum_{a \in \mathcal{A}} \pi(a \mid X_i) \hat{r}(X_i, a) \,, \tag{6.1}$$

where $\hat{r}(x, a)$ is an estimation of $r(x, a; \theta_*)$. DM estimators may exhibit modeling bias, but they generally have lower variance than IPS (Saito and Joachims, 2022). Another advantage of DM is its practical utility without assuming access to the logging policy π_0 (Jeunen and Goethals, 2021; Aouali et al., 2022c; Hong et al., 2023). Also, DMs can be incorporated into a Bayesian framework, where informative priors can be used to enhance statistical efficiency. This allows for the development of scalable methods suitable for large action spaces, as shown in our work.

6.2 Structured DM

6.2.1 Structured Priors

Pitfalls of non-structured priors. Before presenting `sDM`, we first describe the pitfalls of using the following widely used standard prior,

$$\begin{aligned} \theta_a &\sim \mathcal{N}(\mu_a, \Sigma_a) \,, && \forall a \in \mathcal{A} \,, \\ R \mid \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2) \,, \end{aligned} \tag{6.2}$$

where $\phi(x)$ provides a d-dimensional representation of the context $x \in \mathcal{X}$, and $\mathcal{N}(\mu_a, \Sigma_a)$ represents the prior density of θ_a, with σ^2 being the reward noise variance. Under this prior, each action a has an associated parameter θ_a. Given the prior in Equation (6.2), the posterior distribution of an action parameter follows a multivariate Gaussian: $\theta_a \mid \mathcal{D}_n \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$, where

$$\hat{\Sigma}_a^{-1} = \Sigma_a^{-1} + G_a \,, \qquad \hat{\Sigma}_a^{-1} \hat{\mu}_a = \Sigma_a^{-1} \mu_a + B_a \,.$$

with

$$G_a = \sigma^{-2} \sum_{i \in [n]} \mathbb{1}_{\{A_i = a\}} \phi(X_i) \phi(X_i)^\top \,, \qquad B_a = \sigma^{-2} \sum_{i \in [n]} \mathbb{1}_{\{A_i = a\}} R_i \phi(X_i)$$

Note that G_a and B_a only use the subset of samples $\mathcal{D}_n$ where action a was observed, meaning data from other actions $b \neq a$ do not contribute to the posterior inference for

action a. This results in statistical inefficiency, especially if the logged data $\mathcal{D}_n$ doesn't cover all actions. In particular, the posterior for an unseen action a, $\theta_a \mid \mathcal{D}_n$, would simply revert to the prior $\mathcal{N}(\mu_a, \Sigma_a)$, since we would have $G_a = 0_{d\times d}$ and $B_a = 0_d$ in such case.

Structured priors. To address the above issue, we assume that action rewards correlate and embed this knowledge into the prior. While one could model these correlations by considering the joint posterior distribution of $(\theta_a)_{a\in\mathcal{A}} \mid \mathcal{D}_n$, this becomes computationally burdensome when the number of actions K is large. Instead, we introduce an *unknown* d'*-dimensional latent parameter* $\psi \in \mathbb{R}^{d'}$, sampled from a *latent prior* $q(\cdot)$, such as $\psi \sim q(\cdot)$. The correlations between actions naturally arise because each action parameter θ_a is derived from the same latent parameter ψ.

Specifically, the action parameters θ_a are conditionally independent given ψ and are sampled from a *conditional prior* p_a as $\theta_a \mid \psi \sim p_a(\cdot; f_a(\psi))$ for all $a \in \mathcal{A}$. Here, p_a is parameterized by $f_a(\psi)$, where $f_a : \mathbb{R}^{d'} \to \mathbb{R}^d$ is a known prior function that encodes the hierarchical relationship between action parameters θ_a and the latent parameter ψ. This structure allows for sparsity, meaning that θ_a may depend only on a subset of ψ's coordinates. Moreover, p_a accounts for model uncertainty, allowing for cases where θ_a is not a deterministic function of ψ, i.e., $\theta_a \neq f_a(\psi)$.

The reward distribution for action a in context x is given by $p(\cdot \mid x; \theta_a)$, which depends only on x and θ_a. To summarize, the structured prior is defined below, and its graphical representation is given in Figure 6.1.

$$\begin{aligned} \psi &\sim q(\cdot)\,, && \\ \theta_a \mid \psi &\sim p_a(\cdot; f_a(\psi))\,, && \forall a \in \mathcal{A}\,, \\ R \mid \psi, \theta, X, A &\sim p(\cdot \mid X; \theta_A)\,. && \end{aligned} \tag{6.3}$$

To derive the posterior under this prior, we assume that: *(i)* (X, A) is independent of ψ, and given ψ, (X, A) is independent of θ; and *(ii)* given ψ, the parameters θ_a for all $a \in \mathcal{A}$ are independent.

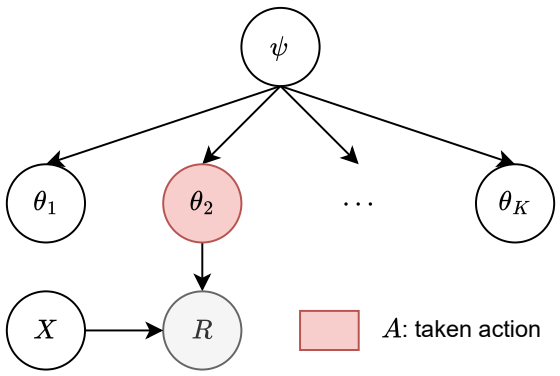


Figure 6.1: Graph representation of the structured prior.

Now, we discuss how to perform off-policy learning under this general structured prior in Equation (6.3), before applying it to linear-Gaussian distributions in Section 6.3.

6.2.2 Off-Policy Learning

Off-policy learning relies on an estimate of the value function $V(\pi; \theta_*)$ obtained using the logged data $\mathcal{D}_n$. In DMs, the estimator $\hat{V}_{\text{DM}}$ in Equation (6.1) requires access to the learned reward $\hat{r}(x, a) \approx r(x, a; \theta_*)$. In our Bayesian setting, this requires access to the action posterior $\theta_a \mid \mathcal{D}_n$ under the prior in Equation (6.3) since the reward is then estimated as $\hat{r}(x, a) = \mathbb{E}[r(x, a; \theta) \mid \mathcal{D}_n]$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, and this estimate is plugged into $\hat{V}_{\text{DM}}$ in Equation (6.1) to estimate $V(\pi; \theta_*)$. Thus, we need to derive the posterior density of the action parameter θ_a, $p(\theta_a \mid \mathcal{D}_n)$, under the structured prior in Equation (6.3), which reads

$$p(\theta_a \mid \mathcal{D}_n) = \int_{\psi} p(\theta_a \mid \psi, \mathcal{D}_n) p(\psi \mid \mathcal{D}_n) \, \mathrm{d}\psi \,, \tag{6.4}$$

where $\psi \mid \mathcal{D}_n$ is the latent posterior and $\theta_a \mid \psi, \mathcal{D}_n$ is the *conditional* action posterior. To compute $p(\theta_a \mid \mathcal{D}_n)$, we first compute $p(\theta_a \mid \psi, \mathcal{D}_n)$ and $p(\psi \mid \mathcal{D}_n)$ and then integrate out ψ following Equation (6.4). First,

$$p(\theta_a \mid \psi, \mathcal{D}_n) \propto \mathcal{L}_a(\theta_a) p_a(\theta_a; f_a(\psi)) \,, \tag{6.5}$$

with $\mathcal{L}_a(\theta_a) = \prod_{(X,A,R) \in S_a} p(R|X; \theta_a)$ is the likelihood of observations of action a ($S_a = (X_i, A_i, R_i)_{i \in [n], A_i = a}$ is the subset of $\mathcal{D}_n$ where $A_i = a$). Similarly,

$$p(\psi \mid \mathcal{D}_n) \propto \prod_{b \in \mathcal{A}} \int_{\theta_b} \mathcal{L}_b(\theta_b) p_b\left(\theta_b; f_b(\psi)\right) \mathrm{d}\theta_b \, q(\psi) \,, \tag{6.6}$$

This allows us to further develop Equation (6.4) as

$$p(\theta_a \mid \mathcal{D}_n) \propto \int_{\psi} \mathcal{L}_a(\theta_a) p_a(\theta_a; f_a(\psi)) \prod_{b \in \mathcal{A}} \int_{\theta_b} \mathcal{L}_b(\theta_b) p_b\left(\theta_b; f_b(\psi)\right) \mathrm{d}\theta_b \, q(\psi) \, \mathrm{d}\psi \,. \tag{6.7}$$

All the quantities inside the integrals in Equation (6.7) are given (the parameters of p_a and q) or tractable (the terms in $\mathcal{L}_a$). Thus, if these integrals can be computed, then the posterior can be fully characterized in closed form, which we will do in Section 6.3 in the fully linear case. Otherwise, the posterior should be approximated.

Finally, we act greedy with respect to our estimator $\hat{V}_{\text{DM}}$ and define the learned policy as the one maximizing it: $\hat{\pi}_{\text{G}} = \operatorname{argmax}_{\pi \in \Pi} \hat{V}_{\text{DM}}(\pi)$. If the set of policies Π contains deterministic policies, then

$$\hat{\pi}_{\text{G}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b \in \mathcal{A}} \hat{r}(x, b)\} \,. \tag{6.8}$$

In particular, we do not adopt the common pessimism approach (Jin et al., 2021). In pessimism, one constructs confidence intervals of the reward estimate $\hat{r}(x, a)$ of the form $|r(x, a; \theta) - \hat{r}(x, a)| \leq u(x, a)$, and then defines the learned policy as $\hat{\pi}_{\text{P}}(a \mid x) = \mathbb{1}\{a = \operatorname{argmax}_{b \in \mathcal{A}} \hat{r}(x, b) - u(x, b)\}$. The advantage of one over another depends on the evaluation metric used. Our metric is the Bayesian suboptimality (BSO), defined in Section 6.4. It assesses the average performance of algorithms across multiple problems rather than the worst-case. The Greedy policy is more suitable for BSO optimization than pessimism (demonstrated theoretically and empirically in Sections C.3.3 and C.4.4).

6.3 Linear-Gaussian Case

In this section, we use linear functions f_a combined with Gaussian distributions for the structured prior Equation (6.3). Precisely, we assume that the latent prior $q(\cdot) = \mathcal{N}(\cdot; \mu, \Sigma)$ is Gaussian with mean $\mu \in \mathbb{R}^{d'}$ and covariance $\Sigma \in \mathbb{R}^{d' \times d'}$. Moreover, let $\mathrm{W}_a \in \mathbb{R}^{d \times d'}$ be the *mixing matrix* for action a, we define $f_a(v) = \mathrm{W}_a v$ for any $v \in \mathbb{R}^{d'}$. We define the conditional prior $p_a(\cdot; f_a(\psi)) = \mathcal{N}(\cdot; \mathrm{W}_a\psi, \Sigma_a)$ is Gaussian with mean $f_a(\psi) = \mathrm{W}_a\psi \in \mathbb{R}^d$ and covariance $\Sigma_a \in \mathbb{R}^{d \times d}$. The reward distribution $p(\cdot \mid x; \theta_a)$ is also linear-Gaussian as $\mathcal{N}(\cdot; \phi(x)^\top \theta_a, \sigma^2)$, where $\phi(\cdot)$ outputs a d-dimensional representation of x and $\sigma > 0$ is the observation noise variance. The whole prior is

$$\begin{aligned} \psi &\sim \mathcal{N}(\mu, \Sigma)\,, && \\ \theta_a \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_a\psi,\, \Sigma_a\Big)\,, && \forall a \in \mathcal{A}\,, \\ R \mid \psi, \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2)\,. && \end{aligned} \tag{6.9}$$

6.3.1 Applications

Mixed-effect modeling. Equation (6.9) allows modeling that action parameters depend on a linear mixture of effect parameters (Chapter 3). Precisely, let J be the number of effects and assume that $d' = dJ$ so that the latent parameter ψ is the concatenation of J, d-dimensional effect parameters, $\psi_j \in \mathbb{R}^d$, such as $\psi = (\psi_j)_{j \in [J]} \in \mathbb{R}^{dJ}$. Moreover, assume that for any $a \in \mathcal{A}$, $\mathrm{W}_a = w_a^\top \otimes I_d \in \mathbb{R}^{d \times dJ}$ where $w_a = (w_{a,j})_{j \in [J]} \in \mathbb{R}^J$ are the *mixing weights* of action a. Then, $\mathrm{W}_a\psi = \sum_{j \in [J]} w_{a,j}\psi_j$ for any $a \in \mathcal{A}$. Sparsity, i.e., when an action a only depends on a subset of effects, is captured through the mixing weights w_a: $w_{a,j} = 0$ when action a is independent of the j-th effect parameter ψ_j and $w_{a,j} \neq 0$ otherwise. Also, the level of dependence between action a and effect j is quantified by the absolute value of $w_{a,j}$. This mixed-effect model can be used in numerous applications (the reader can refer to the first paragraphs of Chapter 3 for examples).

Low-rank modeling. Equation (6.9) can also model the case where the dimension of the latent parameter ψ is much smaller than that of the action parameters θ_a, i.e., when $d' \ll d$. Again, this is captured through the mixing matrices W_a, when W_a is low-rank.

6.3.2 Closed-Form Solutions for `sDM`

The conditional action posterior is known in closed-form as $\theta_a \mid \psi, \mathcal{D}_n \sim \mathcal{N}(\tilde{\mu}_a, \tilde{\Sigma}_a)$, with

$$\tilde{\Sigma}_a^{-1} = \Sigma_a^{-1} + G_a\,, \qquad \tilde{\Sigma}_a^{-1}\tilde{\mu}_a = \Sigma_a^{-1}\mathrm{W}_a\psi + B_a\,, \tag{6.10}$$

where

$$G_a = \sigma^{-2} \sum_{i \in [n]} \mathbb{1}\{A_i = a\}\phi(X_i)\phi(X_i)^\top\,, \qquad B_a = \sigma^{-2} \sum_{i \in [n]} \mathbb{1}\{A_i = a\}R_i\phi(X_i)\,.$$

This posterior has the standard form except that the prior mean $\mathrm{W}_a\psi$ now depends on the latent parameter ψ. Similarly, the effect posterior writes $\psi \mid \mathcal{D}_n \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$, where

$$\begin{aligned}\bar{\Sigma}^{-1} &= \Sigma^{-1} + \sum_{a\in\mathcal{A}} \mathrm{W}_a^\top(\Sigma_a^{-1} - \Sigma_a^{-1}\tilde{\Sigma}_a\Sigma_a^{-1})\mathrm{W}_a\,,\\ \bar{\Sigma}^{-1}\bar{\mu} &= \Sigma^{-1}\mu + \sum_{a\in\mathcal{A}} \mathrm{W}_a^\top\Sigma_a^{-1}\tilde{\Sigma}_a B_a.\end{aligned} \tag{6.11}$$

The latent posterior precision $\bar{\Sigma}^{-1}$ is the sum of the latent prior precision Σ^{-1} and the learned action precisions $\Sigma_a^{-1} - \Sigma_a^{-1}\tilde{\Sigma}_a\Sigma_a^{-1}$, weighted by $\mathrm{W}_a^\top\mathrm{W}_a$. The contribution of each action's learned precision to the latent precision is proportional to $\mathrm{W}_a^\top\mathrm{W}_a$. This intuition similarly applies to interpreting $\bar{\mu}$. Finally, from Equation (6.7), the action posterior is $\theta_a \mid \mathcal{D}_n \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$, where

$$\hat{\Sigma}_a = \tilde{\Sigma}_a + \tilde{\Sigma}_a\Sigma_a^{-1}\mathrm{W}_a\bar{\Sigma}\mathrm{W}_a^\top\Sigma_a^{-1}\tilde{\Sigma}_a\,, \qquad \hat{\mu}_a = \tilde{\Sigma}_a\big(\Sigma_a^{-1}\mathrm{W}_a\bar{\mu} + B_a\big)\,. \tag{6.12}$$

Finally, from Equation (6.9), the reward function is $r(x, a; \theta) = \phi(x)^\top\theta_a$. Thus, the estimated reward is

$$\hat{r}(x, a) = \mathbb{E}[r(x, a; \theta) \mid \mathcal{D}_n] = \phi(x)^\top\hat{\mu}_a\,, \qquad \forall (x, a) \in \mathcal{X}\times\mathcal{A}\,.$$

This can then be plugged in Equation (6.8) for decision-making, leading to

$$\hat{\pi}_{\mathrm{G}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b\in\mathcal{A}} \phi(x)^\top\hat{\mu}_b\}\,.$$

To see why this is more beneficial than the standard prior in Equation (6.2), notice that the mean and covariance of the posterior of action a, $\hat{\mu}_a$ and $\hat{\Sigma}_a$, are now computed using the mean and covariance of the latent posterior, $\bar{\mu}$ and $\bar{\Sigma}$. But $\bar{\mu}$ and $\bar{\Sigma}$ are learned using the interactions with all the actions in $\mathcal{D}_n$. Thus $\hat{\mu}_a$ and $\hat{\Sigma}_a$ are also learned using the interactions with all the actions in $\mathcal{D}_n$, in contrast with the standard prior in Equation (6.2) where they were learned using only the interaction with action a. The additional computational cost of considering the structured prior in Equation (6.9) is small. The computational and space complexities are $\mathcal{O}(K((d^2 + d'^2)(d + d')))$ and $\mathcal{O}(Kd^2)$. For example, when $d' = \mathcal{O}(d)$, these complexities become $\mathcal{O}(Kd^3)$ and $\mathcal{O}(Kd^2)$, respectively. This is exactly the cost of the standard prior in Equation (6.2). In contrast, this strictly improves the computational efficiency of jointly modeling the action parameters, where the complexities are $\mathcal{O}(K^3d^3)$ and $\mathcal{O}(K^2d^2)$ since the joint posterior of $(\theta_a)_{a\in\mathcal{A}} \mid \mathcal{D}_n$ requires converting and storing a $dK \times dK$ covariance matrix.

Remark 6. *`sDM` with linear-Gaussian hierarchies can be used even with data generated from non-linear rewards, and we empirically investigate its robustness to misspecification. We found that this model performs well even if the true rewards are not generated from a linear-Gaussian distribution.*

6.4 Analysis

6.4.1 Bayesian Metrics

The performance of a learned policy $\hat{\pi}$ is evaluated using suboptimality (SO):

$$\mathrm{SO}(\hat{\pi}; \theta_*) = V(\pi_*; \theta_*) - V(\hat{\pi}; \theta_*)\,,$$

where $\pi_* = \text{argmax}_{\pi \in \Pi} V(\pi; \theta_*)$ is the optimal policy. This metric is well-suited when the environment is governed by a unique, fixed ground truth θ_*. It applies to any policy $\hat{\pi}$, whether learned through frequentist approaches (e.g., MLE) or Bayesian ones (e.g., ours). However, when the environment is modeled as a random variable θ_* sampled from some prior distribution, SO becomes less appropriate. Thus, drawing on recent developments in Bayesian analysis for online bandits through Bayes regret (Russo and Van Roy, 2014), we introduce a new metric for offline settings, termed *Bayes suboptimality*, defined as:

$$\text{BSO}(\hat{\pi}) = \mathbb{E}[V(\pi_*; \theta_*) - V(\hat{\pi}; \theta_*)], \tag{6.13}$$

where the expectation is taken over all random variables: the logged data $\mathcal{D}_n$ and θ_*, which is treated as a random variable sampled from the prior. The BSO can be computed in two ways. One method involves taking the expectation under the prior θ_*, followed by taking an expectation under data generated from a fixed environment θ_* as $\mathcal{D}_n \mid \theta_*$. The other method involves taking an expectation under the data $\mathcal{D}_n$, followed by taking an expectation under the posterior $\theta_* \mid \mathcal{D}_n$. The BSO is a reasonable metric for assessing the average performance of algorithms across multiple environments, due to the expectation over θ_*. It is also known that Bayes regret captures the benefits of using informative priors (Chapter 3), and this is similarly achieved by the BSO.

6.4.2 Theoretical Results

Our theory relies on the important well-specified assumption:

Assumption 1 (Well-specified priors)**.** *Action parameters $\theta_{*,a}$ and rewards are drawn from Equation (6.9).*

We also make simplifying assumptions for the sake of exposition.

Assumption 2 (Diagonal covariances for simplicity)**.** *We assume $\Sigma_a = \sigma_0^2 I_d$, $\Sigma = \tau^2 I_{d'}$, $\|\phi(x)\|_2 \leq 1$, and the matrices W_a are normalized such that $\lambda_1(\mathrm{W}_a \mathrm{W}_a^\top) = \lambda_d(\mathrm{W}_a \mathrm{W}_a^\top) = 1$.*

This yields our bound on the BSO of `sDM`.

Theorem 2 (Covariance-Dependent Bound)**.** *Let $\pi_*(x)$ be the optimal action for context x. Then the BSO of* `sDM` *under the structured prior in Equation (6.9) satisfies*

$$\textit{BSO}(\hat{\pi}_G) \leq \alpha_n \, \mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] + \sqrt{\frac{(2\log(2K)+2)(\sigma_0^2+\tau^2)}{n}}, \tag{6.14}$$

where $\alpha_n = \sqrt{d + 2\sqrt{d\log(Kn)} + 2\log(Kn)}$.

Scaling of the bound in Theorem 2 aligns with existing frequentist results (Jin et al., 2021, Theorem 4.4). The main differences lie in the constants and the fact that this rate is achieved using greedy policies in Equation (6.8). This contrasts with the frequentist setting where pessimism is used (Jin et al., 2021) and known to be optimal (Jin et al., 2021,

Theorem 4.7). In fact, greedy policies are optimal when BSO is used as the performance metric. Specifically, $\textsc{Bso}(\hat{\pi}_{\text{G}}) \leq \textsc{Bso}(\pi)$ for any policy π, including pessimistic ones. Therefore, in the Bayesian setting and when BSO is used as a performance metric, greedy policies should always be preferred to pessimistic ones. This fundamental difference is proven in Section C.3.3 and it is of independent interest beyond this work.

Theorem 2 suggests that the BSO primarily depends on the posterior covariance of action $\pi_*(X)$ in the direction of the context $\phi(X)$. That is, when the uncertainty in the posterior distribution of the optimal action $\pi_*(X)$ is low on average across different contexts X and logged data $\mathcal{D}_n$, then the BSO bound is correspondingly small. In particular, the tightness of the bound depends on the degree to which the logged data covers the optimal actions on average.

Theorem 2 can highlight the advantages of using `sDM` over the non-structured prior in Equation (6.2). To see this, notice that the parameters of the non-structured prior in Equation (6.2), μ_a and Σ_a, are obtained by marginalizing out ψ in Equation (6.9). In this case, $\mu_a^{\text{NS}} \leftarrow \text{W}_a\mu$ and $\Sigma_a^{\text{NS}} \leftarrow \Sigma_a + \text{W}_a\Sigma\text{W}_a^\top$. The corresponding posterior covariance is $\hat{\Sigma}_a^{\text{NS}} = ((\Sigma_a + \text{W}_a\Sigma\text{W}_a^\top)^{-1} + G_a)^{-1}$, and is generally larger than the covariance of `sDM`, $\hat{\Sigma}_a$ in Equation (6.12). This is more pronounced when the number of actions K is large and when the latent parameters are more uncertain than the action parameters. Thus, the BSO bound of `sDM` is smaller due to the reduced posterior uncertainty it exhibits. Also, note that even when $\pi_*(X)$ is unobserved in the logged data $\mathcal{D}_n$, `sDM`'s posterior covariance $\hat{\Sigma}_{\pi_*(X)}$ can remain small since we use interactions with all actions to compute it. This contrasts with standard non-structured priors in Equation (6.2), where observing $\pi_*(X)$ is necessary; without such observations, the posterior covariance $\hat{\Sigma}_{\pi_*(X)}$ would simply be the prior covariance $\Sigma_{\pi_*(X)}$.

Next, we provide another bound on the BSO that scales as $\mathcal{O}(1/\sqrt{n})$. To simplify the exposition, we roughly present its scaling with n in Theorem 3 and defer the complete general statement to Section C.3.2. We make the following additional assumptions:

Assumption 3. *Let $G = \mathbb{E}_{X\sim\nu}[XX^\top]$ with $g = \lambda_d(G)$. We assume that $g > 0$.*

Assumption 4 (Context-independent logging policy)**.** *A is independent of X, i.e., $\pi_0(a \mid x) = \pi_0(a) = p_a$ for all x and a. Equivalently, (X_i) are i.i.d. $\sim \nu$ and independent of (A_i), with $\mathbb{P}(A = a) = p_a$.*

Theorem 3 (Scaling with n)**.** *For n large enough, the BSO of `sDM` under the structured prior in Equation (6.9) scales as*

$$\textit{Bso}(\hat{\pi}_{\text{G}}) = \tilde{\mathcal{O}}\left(\sqrt{\mathbb{E}\left[\frac{d}{n\rho_X + 1}\right]} + \sqrt{\frac{\log K}{n}}\right),$$

where $\rho_X = \pi_0(\pi_*(X))$.

The above bound becomes smaller or larger depending on how well the logging policy π_0 covers the optimal actions for each context x.

6.5 Experiments

We evaluate `sDM` using both synthetic and real datasets. We use the average reward of the learned policy relative to the optimal policy as the evaluation metric.

6.5.1 Synthetic Problems

Setting. We simulate synthetic data using the linear-Gaussian model in Equation (6.9) with $\sigma = 1$. The contexts X are sampled uniformly from $[-1, 1]^d$, with $d = 10$. The matrices W_a are sampled uniformly from $[-1, 1]^{d \times d'}$, where we very d' as $d' \in \{5, 10., 20\}$. We set $\Sigma = 3I_{d'}$ and $\Sigma_a = I_d$, meaning the latent parameters are more uncertain than the action parameters. The latent mean μ is randomly sampled from $[-1, 1]^{d'}$. The number of actions is varied as $K \in \{100, 1000\}$, and we use a uniform logging policy to collect data. Additional experiments with different logging policies are presented in Section C.4.

Baselines. First, we use `sDM` under prior in Equation (6.9). Second, we examine `DM (Bayes)`, which uses the standard non-structured prior in Equation (6.2), where parameters μ_a and Σ_a are obtained by marginalizing out the latent parameters ψ in Equation (6.9). Thus `DM (Bayes)` is a standard Bayesian DM that does not capture arm reward correlations. We also include `DM (Freq)`, which estimates $\theta_{*,a}$ by the MLE. We include `IPS` (Horvitz and Thompson, 1952), self-normalized `IPS` (`snIPS`) (Swaminathan and Joachims, 2015b), and doubly robust (`DR`) (Dudik et al., 2014), which we optimize to learn the optimal policy. `MIPS` (Saito and Joachims, 2022) and `PC` (Sachdeva et al., 2024) are also included. Implementation details of baselines is provided in Section C.4.1.

Results. In Figure 6.2, we plot the results and we observe that `sDM` consistently outperforms the baselines across all settings. This performance gap becomes even more significant when sample size n is small. These results highlight `sDM`'s enhanced efficiency in using available logged data, making it particularly beneficial in data-limited situations and scalable to large action spaces.

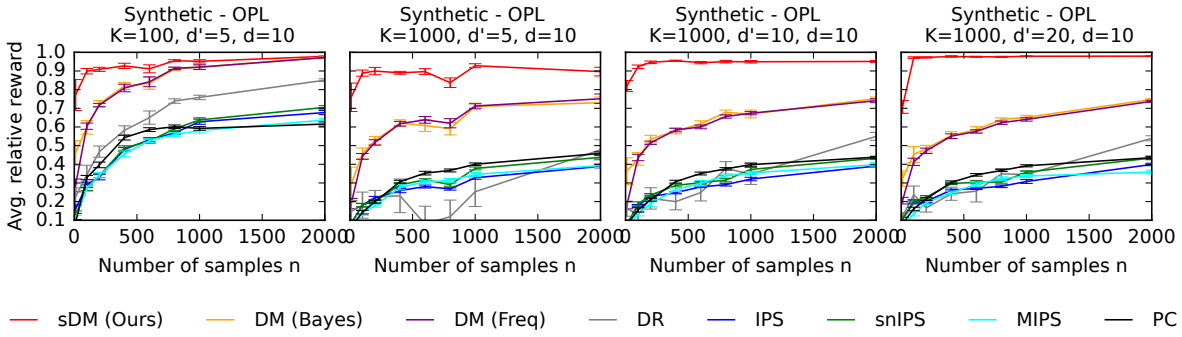


Figure 6.2: The average relative reward of the learned policy using one of the baselines on **synthetic problems** with varying n, K and d'.

Scaling to large action spaces. `sDM` achieves improved scalability compared to standard DM as it leverages data more efficiently. While it still learns a d-dim. parameter for each action a, it does so by considering interactions with all actions in the logged

data $\mathcal{D}_n$, instead of only using interactions with the specific action a. This is crucial, especially given that many actions may not even be observed in $\mathcal{D}_n$. To show `sDM`'s improved scalability, we compare it to the most competitive baseline, `DM (Bayes)`, for varying $K \in [10, 100000]$ with $n = 1000$. The results in Figure 6.3 reveal that the performance gap between `sDM` and `DM (Bayes)` becomes more significant when the number of actions K increases. Hence, despite the necessity for `sDM` to learn distinct parameters for each action, accommodating practical scenarios like recommender systems where unique embeddings are learned for each product, it still enjoys good scalability.

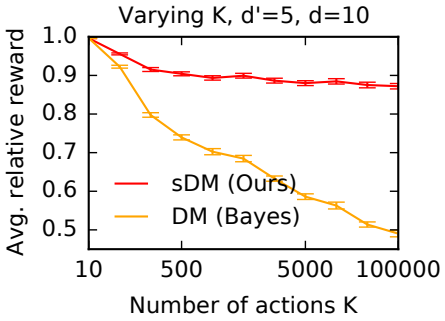


Figure 6.3: `sDM` *vs.* `DM (Bayes)` for varying K.

6.5.2 MovieLens Problems

Setting. We use MovieLens 1M (Lam and Herlocker, 2016), which contains 1 million ratings representing the interactions between 6,040 users and 3,952 movies. To create a semi-synthetic environment, we first apply a low-rank factorization to the rating matrix, producing 5-dim. representations: $x_u \in \mathbb{R}^5$ for user $u \in [6040]$ and $\theta_a \in \mathbb{R}^5$ for movie $a \in [3952]$. Movies are treated as actions, and contexts X are sampled randomly from the user vectors. The reward for movie a and user u is modeled as $\mathcal{N}(x_u^\top \theta_a, 1)$, serving as proxy for ratings. A uniform logging policy is used to collect data.

Baselines. We consider the same baselines as in synthetic data. A prior is not needed for `DM (Freq)`, `IPS`, `snIPS`, and `DR`. However, for `DM (Bayes)`, a standard prior in Equation (6.2) is inferred from data, where we set μ_a to be the mean of movie vectors across all dimensions, and $\Sigma_a = \text{diag}(v)$, where v represents the variance of movie vectors across all dimensions. Unlike the synthetic experiments, the latent structure assumed by `sDM` is not inherently present in MovieLens. But we learn it by training a Gaussian Mixture Model (GMM) to cluster movies into $J = 5$ mixture components. This gives rise to the mixed-effect structure described in Section 6.3, which represents a specific instance of `sDM` with $d' = dJ = 25$. `MIPS` also has access to movie clusters, while we use the knn smoothing implementation of `PC` (see (Sachdeva et al., 2024, Section 3)). Note that `DM (Bayes)`, `sDM`, `MIPS` and `PC` use the same subset of data (of size 1000) to learn their priors/assumed structure and thus we compare them fairly. We conduct experiments with $K \in \{100, 1000\}$ randomly selected movies.

Results. Results are in Section 6.5.2. Even though the latent structure assumed by `sDM` is not inherently present in MovieLens, `sDM` still outperforms the baselines by learning it

offline.

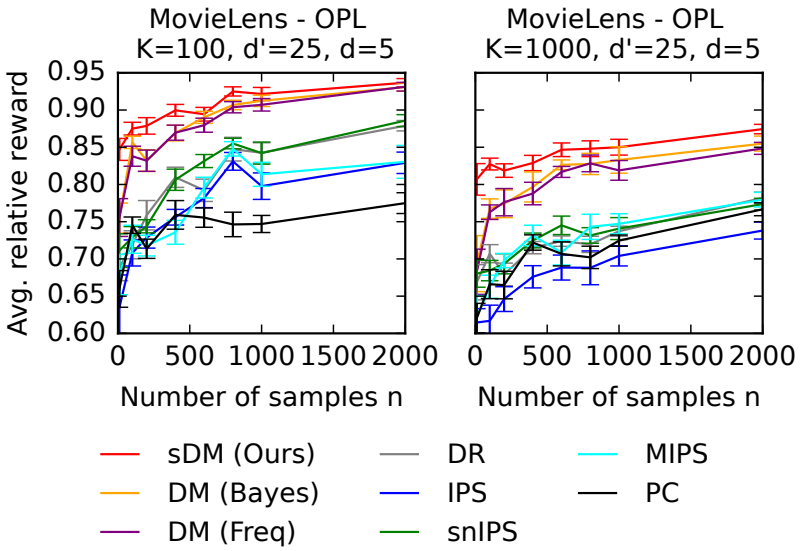


Figure 6.4: The average relative reward of the learned policy using one of the baselines on **MovieLens problems** with varying n, K and d'.

6.6 Conclusion

We introduced `sDM`, a structured approach to off-policy learning that leverages latent structure among actions to enhance statistical efficiency while maintaining computational tractability, particularly in large action spaces with limited data coverage. Within a Bayesian framework, we proved that greedy policies outperform pessimistic ones under Bayesian suboptimality and established $\mathcal{O}(1/\sqrt{n})$ convergence without requiring restrictive full-support assumptions.

Our work has several limitations. First, our theoretical analysis assumes a well-specified prior; while we empirically observed robustness to misspecification, formal guarantees under prior mismatch remain an open question. Second, closed-form posterior updates are available only for linear-Gaussian hierarchies; extending to nonlinear reward models requires approximate inference, which may compromise computational efficiency or statistical accuracy. Third, the latent structure must be specified or learned pre-trained, adding an additional modeling task. Finally, extending `sDM` to handle nonlinear hierarchies, building on the diffusion-based approach of Chapter 4, is a promising avenue for future work.

Chapter 7

Optimization Matters More than Esimation

Contents

This chapter challenges the dominant paradigm in off-policy learning (explored in Chapter 8), which frames the problem as finding a policy $\hat{\pi} = \text{argmax}_\pi \hat{V}(\pi)$ (or, with pessimism, $\hat{\pi} = \text{argmax}_\pi[\hat{V}(\pi) - \text{pen}(\pi)]$), where $\hat{V}$ is an IPS-based[1] estimate of the true policy value $V(\pi)$. The rationale behind these objectives is that maximizing a more accurate value estimate yields a better policy. However, this *estimator-centric view* neglects a crucial factor: the optimization landscape.

[1]Recall that IPS is an importance-weighting estimator of the policy value. We use *IPS-based* to refer to any estimator derived from or inspired by importance weighting.

IPS-based objectives (Dudík et al., 2011; Dudík et al., 2012; Dudik et al., 2014; Wang et al., 2017; Farajtabar et al., 2018; Su et al., 2020; Metelli et al., 2021; Kuzborskij et al., 2021; Saito and Joachims, 2022) are highly non-concave under common policy parameterizations (Chen et al., 2019), prone to suboptimal local maxima and plateaus: issues that are exacerbated in large action spaces. Even sophisticated estimators designed to reduce variance fail to overcome this optimization barrier, as they often induce equally difficult landscapes.

We make the following contributions. **1)** We show that *objective-aware policy parametrization* can partially alleviate these difficulties by structuring the policy class to match the implicit biases of the estimator. Such parametrizations reduce the effective search space and can shorten optimization plateaus and local maxima. However, this strategy does not eliminate the fundamental non-concavity of IPS-based objectives, leaving optimization as the central bottleneck. **2)** Motivated by this limitation, we advocate for an alternative approach based on *policy-weighted log-likelihood (PWLL)* objectives. Unlike traditional estimators, PWLL optimizes an objective $\hat{U}(\pi)$ designed for ease of optimization rather than accuracy in estimating $V(\pi)$. Although PWLL objectives perform poorly as value estimators, their favorable concave landscape makes them significantly more effective for policy learning. **3)** Through theoretical and empirical analysis, we demonstrate that this optimization-centric approach consistently enables simpler PWLL objectives to outperform complex, state-of-the-art IPS-based methods, particularly in large action spaces.

Setting and organization. This chapter considers the general setting of Section 5.1 with $R \in [0, 1]$. The remainder is organized as follows. Section 7.1 employs an asymptotic lens to analyze IPS-based objectives and derives objective-aware policy parametrizations that partially alleviate their optimization challenges. Section 7.2 introduces PWLL objectives and establishes their favorable optimization properties. Section 7.3 presents large-scale experiments. We conclude in Section 7.4.

7.1 Analysis of IPS-Based Objectives

IPS-based objectives optimize an estimator $\hat{V}(\pi)$ of the policy value $V(\pi)$. To understand the policies to which these estimators converge, we study their *oracle policies* $\pi_*^{\text{METHOD}} = \operatorname{argmax}_\pi \mathbb{E}[\hat{V}^{\text{METHOD}}(\pi)]$. Taking the expectation removes sampling fluctuations and isolates the inductive bias of each objective: different estimators yield different oracle policies, even with infinite data. Crucially, oracle policies admit closed-form expressions, enabling precise characterization of each estimator's implicit bias. This analysis motivates *objective-aware parametrizations* that align the policy class with the estimator's bias to ease optimization: the first improvement we propose in this chapter.

7.1.1 Standard IPS-Based Objectives

The foundational `IPS` estimator (Horvitz and Thompson, 1952) re-weights observed rewards by the ratio between the target policy π and the logging policy π_0:

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)} R_i. \tag{7.1}$$

In expectation, IPS selects the best-rewarding action among those in the support of π_0:

$$\pi_*^{\texttt{IPS}}(a \mid x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a' \in \mathcal{A}} r(x, a') \mathbb{1}[\pi_0(a' \mid x) > 0]\right]. \tag{7.2}$$

Clipped IPS (cIPS). To mitigate the high variance of IPS, a widely used variant is cIPS (Bottou et al., 2013) that clips small propensity scores at a threshold $\tau \in (0, 1)$:

$$\hat{V}_{\text{CIPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\max\{\pi_0(A_i \mid X_i), \tau\}} R_i. \tag{7.3}$$

This clipping introduces a bias. The oracle policy down-weights the rewards of rare actions, causing it to favor actions that were frequent under π_0, even if they are suboptimal:

$$\pi_*^{\texttt{cIPS}}(a \mid x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a' \in \mathcal{A}} \frac{\pi_0(a' \mid x)}{\max\{\pi_0(a' \mid x), \tau\}} r(x, a')\right]. \tag{7.4}$$

Exponential smoothing (ES). Instead of hard clipping, ES (Aouali et al. (2023a), Chapter 8) smooths importance weights by raising propensities to a fractional power $\alpha \in (0, 1)$:

$$\hat{V}_{\text{ES}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\pi_0(A_i \mid X_i)^{\alpha}} R_i. \tag{7.5}$$

Its oracle policy balances reward maximization with preference for frequent actions:

$$\pi_*^{\texttt{ES}}(a \mid x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a' \in \mathcal{A}} r(x, a') \pi_0(a' \mid x)^{1-\alpha}\right]. \tag{7.6}$$

Another variant of ES regularizes the entire importance weight as $(\frac{\pi}{\pi_0})^{\beta}$ instead of only the denominator. In contrast to the deterministic policies derived from IPS, cIPS, and the ES formulation above, this approach yields a stochastic oracle policy: $\pi_*^{\texttt{ES}}(a \mid x) \propto r(x, a)^{1/(1-\beta)} \pi_0(a \mid x)$. Other regularizations include logarithmic smoothing (Sakhi et al., 2024), implicit exploration (Gabbianelli et al., 2024), harmonic correction (Metelli et al., 2021), shrinkage (Su et al., 2020). But we do not include as ES and cIPS are already representative of them.

Doubly robust (DR). The DR estimator incorporates a reward model $\hat{r}(x, a)$ to reduce variance and enable generalization to actions outside π_0's support. A common clipped variant is:

$$\hat{V}_{\text{DR}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i \mid X_i)}{\max\{\pi_0(A_i \mid X_i), \tau\}} \left(R_i - \hat{r}(X_i, A_i)\right) + \mathbb{E}_{A \sim \pi(\cdot \mid X_i)}\left[\hat{r}(X_i, A)\right]. \tag{7.7}$$

Its oracle policy interpolates between the reward model prediction and an importance weighting correction for the reward model error:

$$\pi_*^{\texttt{DR}}(a \mid x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a' \in \mathcal{A}} \hat{r}(x, a') + \frac{\pi_0(a' \mid x)}{\max\{\pi_0(a' \mid x), \tau\}} \left(r(x, a') - \hat{r}(x, a')\right)\right]. \tag{7.8}$$

7.1.2 Large-Scale IPS-Based Objectives

In large action spaces, importance weights $\frac{\pi(a|x)}{\pi_0(a|x)}$ can become huge, leading to estimators with high variance. To mitigate this, modern methods compute marginalized importance weights over a lower-dimensional action representation, trading bias for reduced variance.

Marginalized IPS (`MIPS`). `MIPS` (Saito and Joachims, 2022) tackles large action spaces by clustering actions. It maps each action a to a cluster c via a function $h : \mathcal{A} \to \mathcal{C}$, where $|\mathcal{C}| \ll |\mathcal{A}|$. Estimation is then performed at the cluster level:

$$\hat{V}_{\mathrm{MIPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(C_i \mid X_i)}{\pi_0(C_i \mid X_i)} R_i, \quad \text{where } C_i = h(A_i) \text{ and } \pi(c \mid x) = \sum_{a \in c} \pi(a \mid x). \tag{7.9}$$

This cluster-level marginalization introduces bias: the oracle policy only selects the best *cluster* based on its average reward under π_0, and cannot differentiate between actions within that cluster:

$$\pi_*^{\texttt{MIPS}}(c \mid x) = \mathbb{I}\left[c = \operatorname*{argmax}_{c' \in \mathcal{C}} \left\{ \frac{\sum_{a \in c'} \pi_0(a \mid x) r(x,a)}{\sum_{a \in c'} \pi_0(a \mid x)} \right\}\right]. \tag{7.10}$$

Hence, `MIPS` offers no specific guidance for selecting an action within the optimal cluster; any action is considered equally valid. Consequently, one possible induced action-level oracle under uniform tie-breaking is:

$$\pi_*^{\texttt{MIPS}}(a \mid x) = \frac{\mathbb{I}\left[h(a) = \operatorname{argmax}_{c' \in \mathcal{C}} \left\{ \frac{\sum_{a \in c'} \pi_0(a|x) r(x,a)}{\sum_{a \in c'} \pi_0(a|x)} \right\}\right]}{|h(a)|}.$$

where $|h(a)|$ denotes the size of the cluster containing action a.

Conjunct effect modeling (`OffCEM`). Building on `MIPS`, `OffCEM` (Saito et al., 2023) uses a reward model $\hat{r}$ to correct for the cluster-level aggregation bias, in a doubly robust fashion:

$$\hat{V}_{\mathrm{OFFCEM}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\pi(C_i \mid X_i)}{\pi_0(C_i \mid X_i)} \left(R_i - \hat{r}(X_i, A_i)\right) + \mathbb{E}_{A \sim \pi(\cdot|X_i)}[\hat{r}(X_i, A)] \right). \tag{7.11}$$

The resulting oracle policy selects the action that maximizes the model-predicted reward $\hat{r}$, plus a cluster-level correction term that accounts for model error:

$$\pi_*^{\texttt{OffCEM}}(a \mid x) = \mathbb{I}\left[a = \operatorname*{argmax}_{a' \in \mathcal{A}} \left\{ \hat{r}(x,a') + \frac{\sum_{\bar{a} \in h(a')} \pi_0(\bar{a} \mid x)(r(x,\bar{a}) - \hat{r}(x,\bar{a}))}{\sum_{\bar{a} \in h(a')} \pi_0(\bar{a} \mid x)} \right\}\right]. \tag{7.12}$$

Two-stage decomposition (`POTEC`). In this chapter, we see `POTEC` (Saito et al., 2025) as an *optimization strategy of* `OffCEM` (rather than seeing it as a new estimator). It restricts the policy to a cluster-informed form,

$$\pi(a \mid x) = \sum_{c \in \mathcal{C}} \pi^{\mathrm{RM}}(a \mid x, c) \pi^{\mathrm{CL}}(c \mid x),$$

where $\pi^{\text{RM}}(a \mid x, c) = \mathbb{1}[a = \text{argmax}_{a' \in c} \hat{r}(x, a')]$ is fixed, model-based policy that deterministically selects the best action within each cluster. Learning is then simplified to finding the optimal cluster-level policy π^{CL} that maximizes the `OffCEM` objective in Equation (7.11):

$$\hat{V}_{\text{POTEC}}(\pi^{\text{CL}}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\pi^{\text{CL}}(C_i \mid X_i)}{\pi_0(C_i \mid X_i)} (R_i - \hat{r}(X_i, A_i)) + \sum_{c \in \mathcal{C}} \pi^{\text{CL}}(c \mid X_i) \hat{r}^*_c(X_i) \right), \quad (7.13)$$

where $\hat{r}^*_c(x) = \max_{a \in c} \hat{r}(x, a)$ is the estimated reward of the best action in cluster c. This practical decomposition has the same optimal oracle policy as `OffCEM`:

$$\pi^{\texttt{POTEC}}_* = \pi^{\texttt{OffCEM}}_*.$$

Policy convolution (`PC`). Moving beyond hard clustering, `PC` (Sachdeva et al., 2024) leverages the assumption that actions close in an embedding space yield similar rewards. For each action a, it aggregates over its neighborhood of nearest neighbors $N_\epsilon(a) = \{a' : d(a, a') < \epsilon\}$, where d is a pre-defined distance metric (e.g., ℓ_2 distance between action embeddings):

$$\hat{V}_{\text{PC}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(N_\epsilon(A_i) \mid X_i)}{\pi_0(N_\epsilon(A_i) \mid X_i)} R_i, \quad \text{with } \pi(N_\epsilon(a) \mid x) = \sum_{a' \in N_\epsilon(a)} \pi(a' \mid x). \quad (7.14)$$

The induced oracle policy is deterministic: it selects the action a' that maximizes an aggregated neighborhood score. Each logged neighbor $\bar{a} \in N_\epsilon(a')$ contributes its reward $r(x, \bar{a})$, weighted by the conditional probability of observing $\bar{a}$ under the logging policy restricted to its neighborhood.

$$\pi^{\texttt{PC}}_*(a \mid x) = \mathbb{I}\left[a = \underset{a' \in \mathcal{A}}{\text{argmax}} \left\{ \sum_{\bar{a} \in N_\epsilon(a')} \frac{\pi_0(\bar{a} \mid x) r(x, \bar{a})}{\pi_0(N_\epsilon(\bar{a}) \mid x)} \right\} \right]. \quad (7.15)$$

Other recent IPS variants for large action spaces (Peng et al., 2023; Cief et al., 2024; Taufiq et al., 2024) are often extensions of MIPS that relax its core assumptions. We focused on four methods (`MIPS`, `OffCEM`, `POTEC`, and `PC`), which we consider representative of this family. Since these variants largely share the same MIPS foundation and optimization procedure (with the notable exception of `POTEC`), we expect our findings to be generally applicable.

7.1.3 Optimization Challenges

The effectiveness of IPS-based estimators in off-policy learning is often limited by their challenging optimization landscape. These objectives become difficult to optimize when paired with standard, expressive policy classes such as the softmax. This section explores why this occurs and introduces *objective-aware parametrization* as a strategy to mitigate, though not entirely solve, the problem.

To analyze the optimization process, we consider policies parametrized by a softmax function over an *effective action space*[2] $\mathcal{A}_{\text{eff}} \subseteq \mathcal{A}$, which is the set of actions that can be assigned non-zero probability. By default, $\mathcal{A}_{\text{eff}} = \mathcal{A}$, but we explain below why restricting it to match the structure of the estimator's oracle policy can be beneficial. Specifically, the policy takes the form:

$$\pi_\theta(a \mid x) = \frac{\exp(s_\theta(x,a))}{\sum_{a' \in \mathcal{A}_{\text{eff}}} \exp(s_\theta(x,a'))} \mathbb{1}_{a \in \mathcal{A}_{\text{eff}}}, \quad \forall a \in \mathcal{A}, \tag{7.16}$$

where $s_\theta(x,a)$ is a learnable score function. Common choices are linear softmax scores:

$$\text{lightweight: } s_\theta(x,a) = \phi(x,a)^\top \theta, \qquad \text{heavyweight: } s_\theta(x,a) = \phi(x)^\top \theta_a, \tag{7.17}$$

which we call *lightweight parametrization* (a single shared parameter vector θ, corresponding to a joint reward model) and *heavyweight parametrization* (separate parameters θ_a for each action, corresponding to a disjoint reward model).

The size of the effective action space, $K_{\text{eff}} = |\mathcal{A}_{\text{eff}}|$, is the critical factor governing optimization difficulty. The following propositions (proofs in Section D.2, adapted from Chen et al. (2019); Mei et al. (2020a)) reveal the severity of the problem.

First, gradient-based methods can become trapped in suboptimal regions for extended periods.

Proposition 3 (Optimization plateaus)**.** *For any IPS-based estimator $\hat{V}$ that is linear in π, even with a linear softmax policy, there exist problem instances where gradient ascent remains trapped in a suboptimal region for $\mathcal{O}(K_{\text{eff}})$ iterations.*

Second, the optimization landscape has numerous poor local maxima.

Proposition 4 (Local maxima)**.** *Under similar conditions, the optimization landscape for IPS-based objectives can contain a number of local maxima that is exponential in K_{eff}.*

These results highlight that K_{eff} plays a central role in optimization difficulty. The standard choice of $\mathcal{A}_{\text{eff}} = \mathcal{A}$, which sets $K_{\text{eff}} = K$, leads to optimization failure in large action spaces where K can reach millions: learning must navigate a landscape with potentially $\mathcal{O}(K)$-length plateaus and exponentially many local maxima.

Surprisingly, even sophisticated methods designed specifically for large action spaces often fall into this trap. At first glance, methods such as `MIPS`, `OffCEM`, and `PC` appear to operate in a smaller space because their objectives involve marginalized probabilities: $\pi(C_i \mid X_i)$ in `MIPS` and `OffCEM`, or $\pi(N_\epsilon(A_i) \mid X_i)$ in `PC`. However, these marginalized terms are defined as sums over an underlying action-level policy:

$$\pi(C_i \mid X_i) = \sum_{a \in C_i} \pi(a \mid X_i), \qquad \text{and} \qquad \pi(N_\epsilon(a) \mid x) = \sum_{a' \in N_\epsilon(a)} \pi(a' \mid x).$$

[2]The effective action space can also depend on context x, i.e., $\mathcal{A}_{\text{eff}}(x) \subseteq \mathcal{A}$. We omit this dependence for notational simplicity.

Then, if $\pi(a \mid x)$ is a softmax over $\mathcal{A}$, then $K_{\text{eff}} = K$ and Propositions 3 and 4 apply with $K_{\text{eff}} = K$ which is large. The only exception is `POTEC`, which fixes the intra-cluster policy π^{RM} and only optimizes a cluster-level policy π^{CL}. This reduces the effective action space to $\mathcal{A}_{\text{eff}} = \mathcal{C}$ with $K_{\text{eff}} = |\mathcal{C}| \ll K$, directly mitigating the optimization pathologies.

Design implications: objective-aware parametrization

The choice of K_{eff} introduces a fundamental trade-off. A smaller effective action space simplifies the optimization landscape, but risks excluding the optimal action and reduces policy expressiveness. If $\mathcal{A}_{\text{eff}}$ is chosen arbitrarily, it may degrade performance. The challenge is to find the *sweet spot*: a parametrization constrained enough to be optimizable, yet expressive enough to contain the objective's maximizer.

This is precisely where our asymptotic analysis helps. The oracle policy π_*^{METHOD} reveals the minimal sufficient set of actions required to maximize each objective. By aligning the policy parametrization with this structure, we can reduce K_{eff} without sacrificing performance: the core principle of our proposed *objective-aware parametrization*.

For instance, the oracle policies for `IPS`, `cIPS`, and `ES` are confined to the support of the logging policy, $S_0(x)$. This implies that $\mathcal{A}_{\text{eff}} = S_0$ is sufficient, reducing K_{eff} from K to $|S_0| \ll K$. Similarly, for `OffCEM` and `MIPS`, the cluster-level structure of their oracle policies suggests a two-stage decomposition similar to that of `POTEC`, reducing K_{eff} to $|\mathcal{C}|$. We summarize these observations as claims, validated empirically in Section 7.3:

Claim 1. *For `IPS`, `cIPS`, and `ES`, restricting the policy support to S_0 reduces K_{eff} and yields superior learned policies.*

Claim 2. *For `OffCEM` and `MIPS`, a two-stage `POTEC`-style decomposition that optimizes at the cluster level outperforms action-level parametrization.*

While objective-aware parametrization mitigates the optimization pathologies of Propositions 3 and 4 by reducing K_{eff}, it only treats the symptoms without curing the underlying non-concavity. In the next section, we propose a more fundamental shift: abandoning value estimation in favor of inherently tractable objectives.

7.2 Analysis of PWLL objectives

To overcome the optimization challenges of IPS-based objectives, we consider policy-weighted log-likelihood (PWLL) objectives. These methods trade accurate value estimation for a well-behaved, concave optimization landscape, leading to more robust and effective policy learning.

General form. Given a positive weighting function $g(r, p_0)$, the PWLL objective is:

$$\hat{U}_g(\pi) = \frac{1}{n} \sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i)) \log \pi(A_i \mid X_i). \tag{7.18}$$

The key motivation behind PWLL is to replace the linear dependence on the policy in IPS-based estimators, responsible for plateaus and local maxima in Section 7.1, with a concave transformation. Softmax policies are parametrized through scores $s_\theta(x, a)$, and the map $s \mapsto \log \operatorname{softmax}(s)$ is concave. Consequently, for common linear parametrizations in Equation (7.17), the composition $\log \pi_\theta(a \mid x)$ is concave in θ. This removes the optimization pathologies inherent to IPS-based objectives. Proposition 5 (proof in Section D.2.) formalizes this advantage.

Proposition 5. *For linear softmax policies π_θ, the PWLL objective $\hat{U}_g(\pi_\theta)$ is concave in θ. With ℓ_2 regularization, it is strongly concave.*

Proposition 5 makes PWLL appealing for stochastic optimization. In Section D.3, we show that under standard assumptions of bounded feature norms $\|\phi(x, a)\|$ and weights $g(R_i, \pi_0(A_i, X_i))$, these objectives satisfy the regularity conditions necessary to invoke established convergence theorems (Garrigos and Gower, 2023). This allows us to derive problem-dependent convergence guarantees: stochastic gradient ascent attains a global $\mathcal{O}(1/\sqrt{T})$ rate in the general (concave) case (Proposition 12), accelerating to a geometric rate under ℓ_2-regularization (Proposition 13).

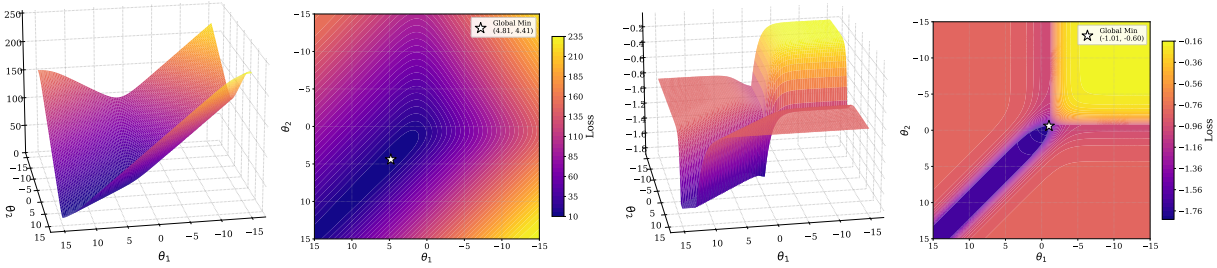
Beyond optimization properties, PWLL also admits a simple statistical interpretation. $\hat{U}_g(\pi)$ in Equation (7.18) is a weighted log-likelihood: the term $\log \pi(A_i \mid X_i)$ performs standard behavior cloning, while the weight $g(R_i, \pi_0(A_i \mid X_i))$ determines how *desirable*[3] each logged sample is. This turns off-policy learning into a form of logging-aware and reward-weighted maximum-likelihood estimation. Different choices of g encode different notions of desirability. For example, the weighting

$$g(r, \pi_0(a \mid x)) = \frac{r}{\max\{\pi_0(a \mid x), \tau\}}$$

emphasizes samples with high reward while reducing the influence of actions that the logging policy selected very frequently. At the same time, the clipping at τ prevents extremely rare actions from receiving disproportionately large weights, ensuring that their contribution is attenuated once $\pi_0(a \mid x)$ falls below the threshold. In this view, desirable samples are those that provide strong reward evidence without allowing very small propensities to dominate the updates. Many other PWLL variants arise from different choices of g (see below), each specifying a distinct prioritization scheme for the logged data, while all benefit from the concavity induced by the logarithmic term.

To illustrate the qualitative difference between PWLL and IPS-based objectives, we construct a simple offline bandit problem with $K = 3$ actions and visualize the resulting optimization landscapes in a two-parameter policy space. Concretely, we consider a non-contextual setting with deterministic mean rewards $r = (0.9, 0.7, 0.2)$ and a logging policy π_0 whose support places almost all mass on action 3 ($\pi_0(1) = 0.002$, $\pi_0(2) = 0.003$, $\pi_0(3) = 0.995$). We generate a fixed dataset of $n = 60$ logged samples (A_i, R_i) by drawing actions $A_i \sim \pi_0$ and binary rewards from the corresponding Bernoulli distributions, $R_i \sim \text{Bern}(r(A_i))$. To obtain a two-dimensional visualization, we parameterize the target

[3] By how desirable an action is, we mean how strongly this action should influence the learned policy.



(a) PWLL (3D view) (b) PWLL (2D projection) (c) IPS-based (3D view) (d) IPS-based (2D projection)

Figure 7.1: Optimization landscapes on a toy example. PWLL (`cLPI`) vs IPS-based (`cIPS`).

policy using a softmax over three logits: $\pi_\theta(a) = e^{\theta_a} / \sum_{b\in[3]} e^{\theta_b}$, fixing the logit associated with action 3 as $\theta_3 = 1$, and letting the remaining two logits be free parameters (θ_1, θ_2).

In Figure 7.1, the PWLL landscape is concave with well-scaled gradients, and optimization trajectories converge reliably from roughly any initialization. In contrast, the IPS-based landscape consists of flat regions, separated by a narrow band of extremely steep curvature. This creates both vanishing and exploding gradients, severe ill-conditioning, and high sensitivity to initialization and learning rate. This aligns with the optimization pathologies in Propositions 3 and 4.

Remark 7 (Beyond linear-softmax policies)**.** *The concavity guarantee of Proposition 5 assumes linear-softmax policies. In many large-scale recommendation systems, a deep encoder is pre-trained and kept fixed, and only a final linear head is optimized for the downstream task; in this case, the policy is still linear-softmax in the trainable parameters, and PWLL objectives retain their concavity. When the full network is trained end-to-end, concavity no longer holds. Yet, PWLL's gradients $g(R_i, \pi_0(A_i|x_i))\nabla_\theta \log \pi_\theta(A_i \mid X_i)$ match the structure of cross-entropy gradients, which are known to produce stable and well-scaled updates in deep architectures. Thus, even without formal guarantees, PWLL maintains substantially more benign optimization dynamics than IPS-based objectives.*

Local policy improvement (`LPI`). Liang and Vlassis (2022) set $g(r, p_0) = r$, which optimizes the log-likelihood of actions weighted by their observed rewards:

$$\hat{U}_{\text{LPI}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} R_i \log \pi(A_i \mid X_i)\,. \tag{7.19}$$

The oracle policy balances reward-seeking with imitation of the logging policy:

$$\pi_*^{\texttt{LPI}}(a \mid x) \propto r(x, a)\pi_0(a \mid x)\,. \tag{7.20}$$

Clipped LPI (`cLPI`). uses importance-weight clipping, setting $g(r, p_0) = \frac{r}{\max(p_0, \tau)}$:

$$\hat{U}_{\text{CLPI}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\max\{\pi_0(A_i \mid X_i), \tau\}} \log \pi(A_i \mid X_i)\,. \tag{7.21}$$

In a similar spirit to `cIPS`, its oracle policy corrects for action frequency under π_0, down-weighting the influence of rare actions due to the clipping:

$$\pi_*^{\texttt{cLPI}}(a \mid x) \propto r(x,a)\frac{\pi_0(a \mid x)}{\max\{\pi_0(a \mid x), \tau\}}. \tag{7.22}$$

KL regularization (`RegKL`). To further amplify the reward signal relative to the logging policy prior, `RegKL` uses an exponential weighting function $g(r, p_0) = \exp(r/\beta)$:

$$\hat{U}_{\text{REGKL}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \exp(R_i/\beta) \log \pi(A_i \mid X_i)\,. \tag{7.23}$$

The oracle policy is proportional to the logging policy, weighted by the exponentiated reward:

$$\pi_*^{\texttt{RegKL}}(a \mid x) \propto \mathbb{E}_{r\sim p(\cdot|x,a)}\big[\exp(r/\beta)\big]\pi_0(a \mid x). \tag{7.24}$$

The temperature parameter β smoothly interpolates between behavior cloning ($\beta \to \infty$) and greedy reward maximization ($\beta \to 0$).

Note that `BPR` (Rendle et al., 2012) can be seen as an approximate PWLL objective, and we included it in our experiments. In fact, this general form of PWLL lends itself to numerous variations by modifying the weighting function g. For instance, one could introduce variants inspired by regularized IPS like exponential smooting (Chapter 8). While many such variants can be proposed for specific use cases, the central message of our work is that the well-behaved optimization landscape of the PWLL family is of greater practical importance than the estimation accuracy of IPS-based objectives. Thus, an exploration of these PWLL variants is beyond our scope. We contend that the foundational methods analyzed above, `LPI`, `cLPI`, and `RegKL`, along with the widely used `BPR` are sufficient to demonstrate the inherent advantages of PWLL objectives.

Finally, PWLL resembles reward- or advantage-weighted behavioral cloning objectives in RL (Nair et al., 2020; Wang et al., 2020; Peng et al., 2019; Peters, 2006). While those methods address multi-step MDPs and often focus on mitigating distributional shift and bootstrapping errors, we focus on offline contextual bandits with large action spaces: identifying objectives and parametrizations that remain optimizable as K grows, rather than accurately estimating $V(\pi)$. PWLL is critic-free and uses logged rewards and propensities through a weighting function $g(R_i, \pi_0(A_i \mid X_i))$ that induces concave optimization landscapes for common policy classes. This yields substantial gains in large-K bandits without the overhead of value-function estimation. PWLL's optimization-centric perspective complements the usual KL-regularized or trust-region interpretations of these RL methods.

7.3 Empirical Analysis

We conduct our empirical evaluation on three large-scale recommendation datasets: `MovieLens` (K = 60k) (Lam and Herlocker, 2016), `Twitch` (K = 200k) (Rappaz et al., 2021), and

GoodReads ($K = 1$M) (Wan et al., 2019). These benchmarks feature action spaces with up to one million items, representing some of the largest settings studied in the offline policy learning literature. For all experiments, we employ the common softmax inner-product policies. We compare methods from both objective families. For IPS-based objectives, we include IPS, ES, DR, MIPS, OffCEM, POTEC, and PC in Section 7.1. For PWLL objectives, we evaluate LPI, cLPI, RegKL, and BPR in Section 7.2. All implementation details are provided in Section D.4.

7.3.1 Optimization is the Main Bottleneck

To test our central hypothesis that *optimization challenges are a more significant barrier than estimation accuracy*, we evaluate how objectives perform under various optimization configurations. If an algorithm's success is highly dependent on specific hyperparameters like batch size or learning rate, it suggests a difficult, non-robust optimization landscape. This experiment directly probes the practical trainability of each method, a key aspect our paper argues is often overlooked.

The results strongly support our claim. As shown in Figure 7.2, *IPS-based objectives are highly sensitive* to batch size and learning rate schedule: minor changes can cause performance collapse, making them difficult to tune and train reliably. In contrast, *PWLL objectives remain robust*, achieving consistently high reward across all configurations. This stability translates directly into better learned policies: *PWLL objectives outperform IPS-based objectives on all datasets*. Even POTEC, a state-of-the-art method designed for large action spaces, is surpassed by the much simpler and easier-to-optimize cLPI.

One might assume that an objective designed for estimation fidelity, such as a low-MSE IPS-based estimator, would naturally yield a better policy. Our findings show this is not the case. The superiority of PWLL objectives, which are poor value estimators by design, provides compelling evidence against this estimator-centric view. This reinforces our main takeaway: in large action space settings, a tractable optimization landscape is a more critical feature for a learning objective than its statistical accuracy. For completeness, an experiment tracking the MSE of methods is given in Section D.4.

The figure also supports Claim 2. Indeed, there is a consistent performance gap between POTEC and OffCEM. Both methods are designed to maximize the same asymptotic objective as we show in Section 7.1; their statistical goals are identical. The divergence in performance, therefore, can be attributed entirely to their differing optimization strategies. POTEC's use of a two-stage, cluster-level optimization proves far more effective than OffCEM's naive, action-level parametrization.

7.3.2 Objective-Aware Parametrization

To empirically validate Claim 1, we compare a naive, whole-action-space parametrization against our proposed objective-aware approach, which restricts the policy's effective action space to the logging policy support, S_0. As shown for the IPS objective in Figure 7.3, the naive approach is highly unstable, with performance collapsing under simple learning configurations. In contrast, the objective-aware version is very robust, achieving high reward consistently across all batch sizes and schedules. This benefit extends even to

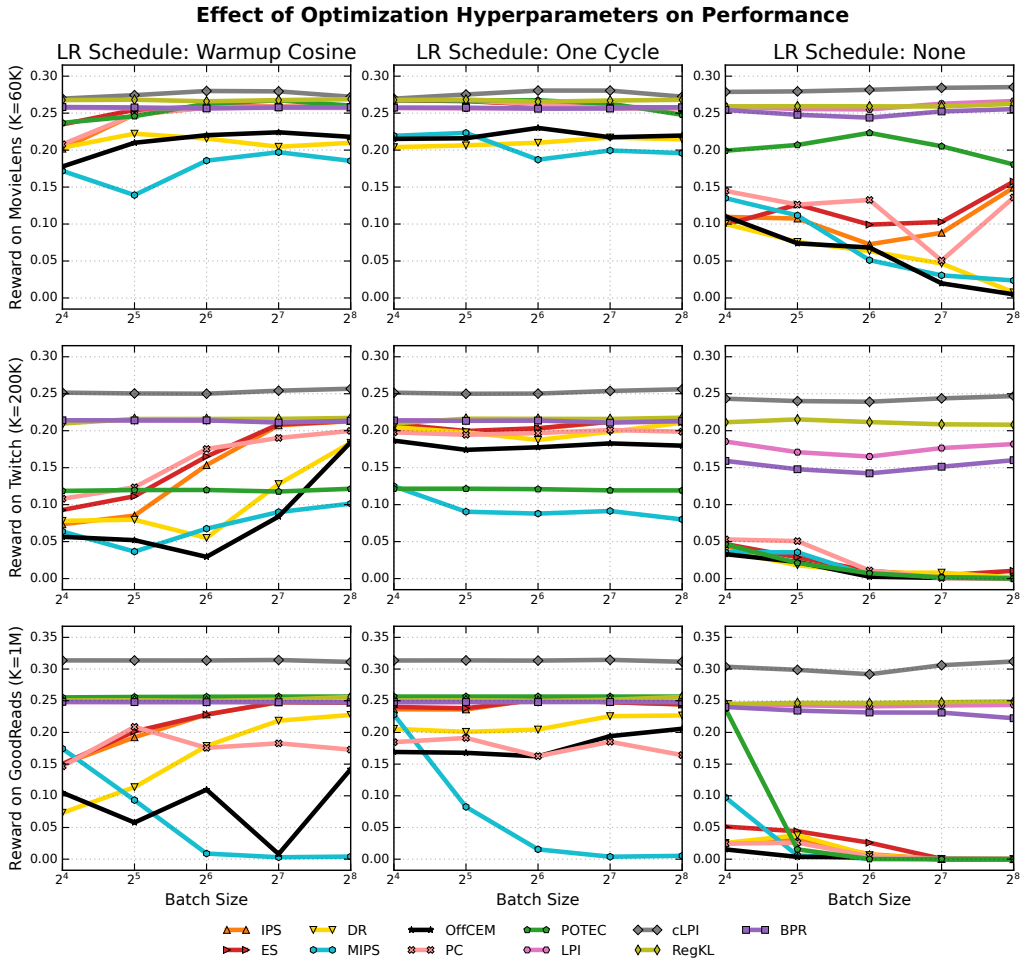


Figure 7.2: Effect of batch size and learning rate schedule on final validation reward using three large-scale datasets. IPS-based objectives are highly sensitive, while PWLL objectives are robust.

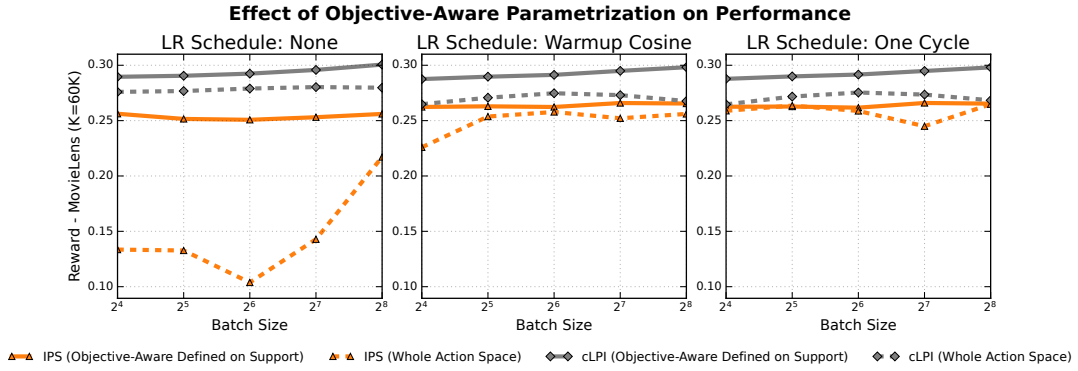


Figure 7.3: The effect of objective-aware parametrization for `IPS` and `cLPI` on `MovieLens`.

inherently stable PWLL objectives like `cLPI`, which achieve even better performance with the restricted support. This provides strong evidence for Claim 1: aligning the policy structure with the objective's inductive bias simplifies the optimization landscape, leading to greater stability and superior learned policies. This finding holds across all datasets, with full results available in Section D.4.

7.4 Conclusion

The dominant approach to off-policy learning focuses on developing sophisticated IPS-based estimators while neglecting a crucial factor: the optimization landscape. We demonstrated, both theoretically and empirically, that this landscape becomes prohibitively difficult to optimize in large action spaces, undermining the practical effectiveness of even state-of-the-art estimators.

Our analysis motivates two strategies. First, objective-aware policy parametrizations align the policy class with the estimator's inductive bias, reducing the effective search space. Second, PWLL objectives abandon value estimation entirely in favor of inherently concave optimization landscapes. Experiments confirm that this focus on optimization tractability yields more robust learning, reduced sensitivity to hyperparameters, and superior policies.

Our work has several limitations. First, PWLL objectives are not value estimators: they cannot be used for off-policy evaluation or policy selection (choosing the best policy from a finite candidate set) when accurate value estimates and their comparison are required. Second, the concavity guarantee of Proposition 5 holds only for linear-softmax policies; when training deep networks end-to-end, PWLL retains favorable gradient structure but loses formal concavity guarantees, although IPS-based objectives face even more severe optimization challenges in this setting. Third, PWLL's oracle policies inherently depend on the logging policy (e.g., $\pi_*^{\texttt{LPI}} \propto r(x,a)\pi_0(a \mid x)$), which may be suboptimal when π_0 has poor coverage of high-reward actions; however, this limitation is shared by IPS-based methods, whose oracle policies similarly depend on π_0's support.

Chapter 8

Principled Pessimism for Exponential Smoothing and Beyond

Contents

Having explored structured direct methods in Chapter 6 and optimization-focused objectives in Chapter 7, we now turn to inverse propensity scoring (IPS). Despite its current practical limitations in large action spaces, many practitioners remain committed to IPS-based methods for their unbiasedness and theoretical guarantees, which enable principled safe off-policy learning. In this chapter, we improve IPS through *exponential smoothing*, a differentiable importance-weight regularization technique that enables a controlled bias-variance trade-off. Then, we adopt the pessimistic framework introduced in Section 5.1, deriving principled uncertainty penalties for our regularized estimators for safe policy learning. Prior work on pessimistic off-policy learning has derived objectives from generalization bounds (Swaminathan and Joachims, 2015a; London and Sandler, 2019), but these approaches suffer from critical limitations: (i) they provide only one-sided bounds that fail to control estimation error in absolute value, limiting their ability to certify estimator quality, (ii) the resulting bounds are intractable and incompatible with stochastic optimization, and (iii) the pessimistic objectives require careful hyperparameter tuning.

We address these limitations by deriving *tractable two-sided* PAC-Bayes generalization bounds that can be optimized directly via stochastic gradient ascent. Unlike prior work (Sakhi et al., 2022), our analysis applies to standard IPS without assuming bounded importance weights, requiring only bounded second moments. Our bounds reveal that the optimal importance-weight smoothing parameter α depends on the quality of the logging policy. Furthermore, our framework generalizes to a broad class of importance-weight regularization techniques, yielding unified pessimistic objectives that enable fair comparison across different importance-weight regularization techniques. We present this extension in the final two sections of this chapter.

The chapter is organized as follows. Section 8.1 presents background on regularized IPS and pessimism. Section 8.2 identifies the shortcomings of hard clipping and introduces our exponential smoothing estimators. Section 8.3 leverages PAC-Bayes theory to derive two-sided generalization bounds within the pessimistic framework. Section 8.4 discusses implications of our results. Section 8.5 demonstrates favorable performance across diverse benchmarks. Finally, Sections 8.6 and 8.7 extend the framework to other importance-weight regularizations and compare them under a unified pessimistic objective.

8.1 Background

We consider the off-policy setting in Section 5.1, where we have access to logged data $\mathcal{D}_n = \{(X_i, A_i, R_i)\}_{i=1}^n$ collected by a known logging policy π_0. As additional notation, we let μ_π be the joint distribution of (X, A, R);

$$\mu_\pi(x, a, r) = \nu(x)\pi(a|x)p(r|x, a)\,, \quad \text{so that} \quad (X_i, A_i, R_i) \sim \mu_{\pi_0}\,.$$

Our goal remains to find a policy $\hat{\pi} \in \Pi$ that maximizes the value

$$V(\pi) = \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}[r(X, A)]\,.$$

8.1.1 Regularized IPS

This chapter focuses on the IPS estimator (Horvitz and Thompson, 1952; Dudík et al., 2012), which estimates the value $V(\pi)$ by re-weighting the samples as

$$\hat{V}_{\text{IPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i\, w(A_i|X_i), \tag{8.1}$$

where $w(a|x) = \pi(a|x)/\pi_0(a|x)$ are the *importance weights*. While IPS provides an unbiased estimate of $V(\pi)$ when the common support condition holds (i.e., $\pi_0(a|x) = 0$ implies $\pi(a|x) = 0$), its variance grows with these importance weights (Swaminathan et al., 2017), which can be arbitrarily large when the target policy π and logging policy π_0 differ significantly. To mitigate this variance issue, it is common to transform the importance weights using regularization functions that introduce controlled bias to reduce variance. A regularized IPS estimator takes the form:

$$\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i\, \hat{w}(A_i|X_i), \tag{8.2}$$

where $\hat{w}(a|x) \leq w(a|x)$ are the regularized importance weights. A common importance-weight regularization approach is clipping where $\hat{w}(a|x) = \min(\frac{\pi(a|x)}{\pi_0(a|x)}, M)$, $M > 0$.

8.1.2 Pessimistic Objectives

Within the pessimistic framework introduced in Section 5.1, we seek to maximize: $\hat{\pi} = \text{argmax}_{\pi \in \Pi}[\hat{V}(\pi) - \text{pen}(\pi)]$ where $\text{pen}(\cdot)$ is a penalty term. The construction of this penalty has been approached in various ways, but generally relies on lower confidence bounds on the policy value:

Evaluation bounds (Metelli et al., 2021) provide confidence intervals for a *fixed* target policy π, showing that with probability at least $1 - \delta$:

$$|V(\pi) - \hat{V}(\pi)| \leq f(\delta, \pi, \pi_0, n). \tag{8.3}$$

Essentially, Equation (8.3) indicates that for a fixed policy $\pi \in \Pi$, the event $|V(\pi) - \hat{V}(\pi)| \leq f(\delta, \pi, \pi_0, n)$ holds with high probability. However, this event depends on the target policy π. Thus Equation (8.3) is useful for evaluating a *single target policy* when having access to *multiple logged data sets* $\mathcal{D}_n$. This poses a problem for off-policy learning, where we optimize over a potentially *infinite space of policies* using a *single logged data set* $\mathcal{D}_n$. This is the fundamental theoretical limitation of using evaluation bounds similar to Equation (8.3) in off-policy learning. While one can transform Equation (8.3) into a generalization bound that holds uniformly over all $\pi \in \Pi$ via a union bound, this typically introduces intractable complexity terms, making the resulting pessimistic objectives, which maximize the lower confidence bound, equally intractable.

One-sided generalization bounds (Swaminathan and Joachims, 2015a; London and Sandler, 2019; Sakhi et al., 2022) address this limitation by providing bounds that hold simultaneously for all policies $\in\in \Pi$. For $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$V(\pi) \geq \hat{V}(\pi) - g(\delta, \Pi, \pi, \pi_0, n), \quad \forall \pi \in \Pi, \tag{8.4}$$

where the function g now depends on the policy space Π. This leads to the pessimistic objective:

$$\hat{\pi} = \underset{\pi \in \Pi}{\text{argmax}}\, \hat{V}(\pi) - g(\delta, \Pi, \pi, \pi_0, n). \tag{8.5}$$

However, one-sided bounds fail to attest to the quality of the estimator. To illustrate, consider a degenerate estimator $\hat{V}^{\text{POOR}}(\pi) = 0$ for all $\pi \in \Pi$. Since $V(\pi) \in [0, 1]$, we trivially have a one-sided bound $V(\pi) \geq \hat{V}^{\text{POOR}}(\pi)$ with probability 1, yet this estimator is entirely uninformative about the true rewards.

Two-sided generalization bounds resolve this issue by controlling both the upper and lower deviations leading to

$$|V(\pi) - \hat{V}(\pi)| \leq g(\delta, \Pi, \pi, \pi_0, n), \quad \forall \pi \in \Pi. \tag{8.6}$$

These bounds ensure the estimator quality and enable oracle inequalities of the form $V(\hat{\pi}) \geq V(\pi_*) - 2g(\delta, \Pi, \pi_*, \pi_0, n)$, where $\hat{\pi}$ is learned using Equation (8.5) and $\pi_* =$

$\operatorname{argmax}_{\pi \in \Pi} V(\pi)$ is the optimal policy. This shows the appeal of pessimism: the sub-optimality gap depends on the bound evaluated at the optimal policy π_*, meaning the estimator only needs to be precise for near-optimal policies rather than uniformly across the policy class Π.

Alternative approaches include heuristics that simplify theoretical bounds for tractability (Swaminathan and Joachims, 2015a; London and Sandler, 2019; Wang et al., 2023), often penalizing by empirical variance or policy divergence while discarding complexity terms. Recent work on implicit pessimism (Gabbianelli et al., 2024; Sakhi et al., 2024) (published after the work in this chapter) shows that careful analysis of specific importance-weight regularizations can yield bounds where the penalty is policy-independent. In this case, maximizing the lower bound reduces to maximizing the estimator directly: the pessimism becomes implicit in the regularization itself.

In this work, we derive a tractable two-sided PAC-Bayesian generalization bound for our exponential smoothing estimator and then generalize it to other importance-weight regularizations.

8.2 Exponential Smoothing

Importance-weight clipping (Swaminathan and Joachims, 2015a) yields the following commonly used estimators

$$\begin{aligned} \texttt{IPS-min} \quad \tilde{V}^{\mathrm{M}}(\pi) &= \frac{1}{n} \sum_{i=1}^{n} R_i \min\left(w(A_i|X_i), M\right), \\ \texttt{IPS-max} \quad \hat{V}^{\tau}(\pi) &= \frac{1}{n} \sum_{i=1}^{n} R_i \frac{\pi(A_i|X_i)}{\max(\pi_0(A_i|X_i), \tau)}. \end{aligned} \tag{8.7}$$

Here `IPS-min` clips the weights while `IPS-max` only clips π_0 in the denominator since π is always smaller than 1. For instance, $M \in \mathbb{R}^+$ in $\tilde{V}^{\mathrm{M}}(\pi)$ trades the bias and variance of the estimator. When M is large, the bias of $\tilde{V}^{\mathrm{M}}(\pi)$ is small but its variance may be large. On the other hand, the variance goes to 0 when $M \approx 0$ since in that case $\tilde{V}^{\mathrm{M}}(\pi) \approx 0$ for any $\pi \in \Pi$. Similarly, $\tau \in [0, 1]$ trades the bias and variance of $\hat{V}^{\tau}(\pi)$ and can be seen as $\tau \approx \frac{1}{M}$.

This *hard* clipping has some limitations. First, $\min(\cdot, M)$ leads to non-differentiable objectives that may require additional care in optimization (Papini et al., 2019). Also, $\min(\cdot, M)$ is constant on $[M, \infty)$ leading to objectives with zero gradients for any policy π that satisfies $w(A_i|X_i) > M$ for any $i \in [n]$. More importantly, hard clipping is sensitive to the choice of the clipping threshold M. In practice, tuning M is challenging and may cause the learned policy to match the logging policy, leading to minimal improvements. To see this, consider the following illustrative example.

For simplicity, suppose that the problem is non-contextual, in which case the reward function r only depends on the actions $a \in \mathcal{A}$. It follows that policies do not depend on $x \in \mathcal{X}$; they are now probability distributions $\pi(\cdot)$ over $\mathcal{A}$. Also, assume that $\mathcal{A} = [100]$ and that the reward received after taking action $a \in [100]$ is binary. That is, $R \sim$

Bern$(r(a))$ where $r(a) = 0.1 - 10^{-3}(a-1)$ is the expected reward of action a, and for any $p \in [0,1]$, Bern(p) is the Bernoulli distribution with parameter p. This means that the best action is 1 and the worst is 100. Finally, the logging policy $\pi_0(\cdot)$ is ϵ-greedy centered at action 50. That is $\pi_0(50) = 1 - \epsilon$, and for any $a \neq 50$, $\pi_0(a) = \frac{\epsilon}{99}$, with $\epsilon = 0.05$.

Now consider 100 deterministic policies $\pi_a(\cdot)$ for $a \in [100]$ such that $\pi_a(\cdot)$ is the Dirac distribution centered at a. In Figure 8.1, we plot the estimated reward of the policies π_a using either `IPS` in Equation (8.1) or `IPS-min` in Equation (8.7). We generate $n = 50$k samples and set $M = 100 = \mathcal{O}(\sqrt{n})$ as suggested by Ionides (2008). *With this choice of* M, `IPS-min` underestimates the reward of all policies π_a for $a \neq 100$ since their weights π_a/π_0 are either 0 or $99/\epsilon > M$. The estimated reward of `IPS-min` is maximized in $\pi_{50} \approx \pi_0$ only. Thus, if we optimize $\tilde{V}^{\text{M}}(\cdot)$ over Dirac policies, we will converge to the logging policy despite its bad performance.

Although the other variant of hard clipping, `IPS-max` in Equation (8.7), is differentiable, it is still sensitive to τ and may induce high bias similar to Figure 8.1. This is due to some loss of information related to the preferences of the logging policy. Indeed, for two actions a and a' such that $\pi_0(a|X_i) \ll \pi_0(a'|X_i) < \tau$ for an observed context X_i, the propensity scores $\pi_0(a|X_i)$ and $\pi_0(a'|X_i)$ will be clipped to the same value τ. Thus the information that, for context X_i, action a' is preferred by the logging policy than action a will be lost.

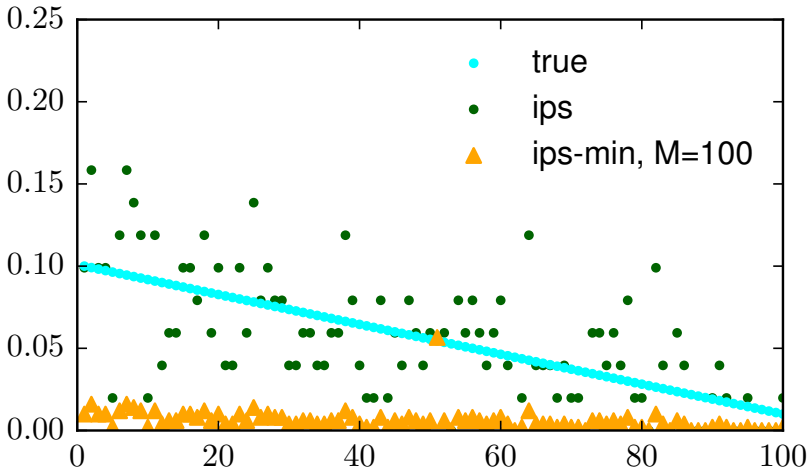


Figure 8.1: Effect of hard clipping on the estimation quality. The x-axis corresponds to actions $a \in [100]$. The y-axis is the estimated reward of each of the 100 policies π_a using either `IPS` or `IPS-min`. The cyan line is the true reward for each policy π_a.

To mitigate this, we propose the following *exponential smoothing* correction for IPS. Our estimators are defined as

$$\texttt{IPS-}\alpha: \ \hat{V}^{\alpha}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i \hat{w}^{\alpha}(A_i|X_i)\,, \ \alpha \in [0,1]\,,$$
$$\texttt{IPS-}\beta: \ \tilde{V}^{\beta}(\pi) = \frac{1}{n}\sum_{i=1}^{n} R_i \tilde{w}^{\beta}_{\pi}(A_i|X_i)\,, \ \beta \in [0,1]\,, \tag{8.8}$$

where $\hat{w}^\alpha(a|x) = \frac{\pi(a|x)}{\pi_0(a|x)^\alpha}$ and $\tilde{w}^\beta_\pi(a|x) = \frac{\pi(a|x)^\beta}{\pi_0(a|x)^\beta}$. Here standard IPS is recovered for $\alpha = 1$ and $\beta = 1$. These estimators yield smooth, everywhere-differentiable objectives and avoid the flat regions induced by hard clipping; this improves optimization in practice. Also, in contrast with `IPS-max` in Equation (8.7), $\hat{V}^\alpha(\pi)$ preserves the preferences of the logging policy. Precisely, for two actions a and a' such that $\pi_0(a|X_i) < \pi_0(a'|X_i)$ for an observed context X_i, we still have $\pi_0(a|X_i)^\alpha < \pi_0(a'|X_i)^\alpha$ and the information that action a' is preferred by the logging policy than action a is preserved.

While a similar correction to `IPS-`β was proposed in Korba and Portier (2022), its use in off-policy learning is novel. Also, Su et al. (2020); Metelli et al. (2021) regularized the importance weights w as $\frac{\lambda_1 w}{\lambda_1 + w^2}, \lambda_1 > 0$ and $\frac{w}{1-\lambda_2+\lambda_2 w}, \lambda_2 \in [0,1]$, respectively. Thus, the expression of both corrections is very different from ours. More importantly, these corrections entail different properties than ours. Roughly speaking, our correction allows us to *simultaneously* **(1)** control a tuning parameter $\alpha \in [0,1]$ that is in a bounded domain $[0,1]$, **(2)** without constraining the resulting importance weights to be bounded, **(3)** and to obtain tractable PAC-Bayes generalization bounds as the correction $\frac{\pi}{\pi_0^\alpha}$ is linear in π; a technical requirement of PAC-Bayes analysis. In contrast, Metelli et al. (2021); Su et al. (2020) do not provide generalization guarantees; they focus on estimation accuracy (e.g., through mean squared error) and only propose heuristics for off-policy learning. Those heuristics are not based on theory, in contrast with ours which is directly derived from our generalization bound. Also, our approach has favorable empirical performance (Section E.3.6).

Although Korba and Portier (2022, Lemma 1) show that smoothing the importance weights similarly to `IPS-`β in Equation (8.8) reduces the variance, it might still be unclear how α and β trade the bias and variance of our estimators in off-policy learning. To see this, let $\alpha \in [0,1]$, then we have

$$
\begin{aligned}
|\mathbb{B}(\hat{V}^\alpha(\pi))| &\leq \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[1 - \pi_0(A|X)^{1-\alpha}\right], \\
\mathbb{V}\left[\hat{V}^\alpha(\pi)\right] &\leq \frac{1}{n}\mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[\frac{\pi(A|X)}{\pi_0(A|X)^{2\alpha-1}}\right],
\end{aligned}
\tag{8.9}
$$

with $\mathbb{B}(\hat{V}^\alpha(\pi)) = \mathbb{E}[\hat{V}^\alpha(\pi)] - V(\pi)$ and $\mathbb{V}[\hat{V}^\alpha(\pi)] = \mathbb{E}[(\hat{V}^\alpha(\pi) - \mathbb{E}[\hat{V}^\alpha(\pi)])^2]$ are respectively the bias and the variance of $\hat{V}^\alpha(\pi)$. The bound of the bias in Equation (8.9) is minimized in $\alpha = 1$ (standard IPS); in which case it is equal to 0 (standard IPS is unbiased). In contrast, the bound of the variance is minimized in $\alpha = 0$. Thus if the variance is small or n is large enough such that $\mathbb{E}[\pi(A|X)/\pi_0(A|X)^{2\alpha-1}]/n \to 0$, then we set $\alpha \to 1$. Otherwise, we set $\alpha \to 0$. This shows that α trades the bias and variance of $\hat{V}^\alpha$. More details and a similar discussion for $\tilde{V}^\beta(\pi)$ are deferred to Section E.1.

8.3 PAC-Bayes Analysis for Off-Policy Learning

We now derive generalization bounds for our estimator. We opt for the PAC-Bayes framework for the following reasons. First, it is known to provide some of the tightest generalization bounds in challenging scenarios (Farid and Majumdar, 2021), for aggregated and randomized predictors (Alquier, 2021). Second, the bounds have a Kullback–Leibler (KL) divergence (Van Erven and Harremos, 2014) term $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})$ that depends on a

fixed prior $\mathbb{P}$ and a *learning posterior* $\mathbb{Q}$ (see Section 8.3.1 for a brief introduction). This quantity can be seen as a complexity measure, similarly to the covering number (Maurer and Pontil, 2009). The difference is that complexity measures are uniform on the space of policies while the KL term in PAC-Bayes depends on the prior $\mathbb{P}$ and the posterior $\mathbb{Q}$. This allows getting sharper bounds when the former is well chosen. Third, the PAC-Bayes perspective fits very well with off-policy learning. In fact, a policy π can be written as an aggregation of predictors under some distribution $\mathbb{Q}$. Thus the prior $\mathbb{P}$ can be associated with the logging policy π_0 that we want to improve upon while the posterior $\mathbb{Q}$ is related to the learning policy π. Fourth, London and Sandler (2019) showed that PAC-Bayes can lead to tractable and scalable objectives, an important consideration for this thesis.

8.3.1 Elements of PAC-Bayes

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be an instance space: e.g., $\mathcal{X}$ and $\mathcal{Y}$ are the input and output space in supervised learning. Let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$ denote a hypothesis space of mappings from $\mathcal{X}$ to $\mathcal{Y}$ (predictors). Also, let $L : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$ be a loss function and assume access to data $\mathcal{D}_n = (Z_i)_{i \in [n]}$ drawn from an unknown distribution $\mathbb{D}$. Let

$$\mathrm{Risk}(h) = \mathbb{E}_{Z \sim \mathbb{D}} \left[L(h, Z) \right]$$

be the risk of $h \in \mathcal{H}$ while

$$\widehat{\mathrm{Risk}}_n(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, Z_i)$$

is its empirical counterpart. Then the main focus in PAC-Bayes is to study the generalization capabilities of random hypotheses $\mathbb{Q}$ on $\mathcal{H}$ by controlling the gap between the expected risk under $\mathbb{Q}$, $\mathbb{E}_{h \sim \mathbb{Q}} \left[\mathrm{Risk}(h) \right]$, and the expected empirical risk under $\mathbb{Q}$, $\mathbb{E}_{h \sim \mathbb{Q}} \left[\widehat{\mathrm{Risk}}_n(h) \right]$.

For example, assume that $L(h, Z) \in [0, 1]$ for any $(h, Z) \in \mathcal{H} \times \mathcal{Z}$, let $\mathbb{P}$ be a *fixed prior* distribution on $\mathcal{H}$ and let $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ over $\mathcal{D}_n \sim \mathbb{D}^n$, the following inequality holds *simultaneously for any posterior* distribution $\mathbb{Q}$ on $\mathcal{H}$:

$$\mathbb{E}_{h \sim \mathbb{Q}} \left[\mathrm{Risk}(h) \right] \leq \mathbb{E}_{h \sim \mathbb{Q}} \left[\widehat{\mathrm{Risk}}_n(h) \right] + \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \log \frac{2\sqrt{n}}{\delta}}{2n}} \,.$$

This was originally proposed by McAllester (1998), and the reader may refer to Alquier (2021); Guedj (2019) for more elaborate introductions of PAC-Bayes theory.

Connection to value functions. The loss L is often chosen as the negative reward, $L(h, Z) = -r(h, Z)$. In this case, minimizing the $\mathrm{Risk}(h)$ is equivalent to maximizing the value function. Thus, PAC-Bayes bounds on the risk directly translate into guarantees on the discrepancy between empirical and true value, providing a principled way to reason about generalization in off-policy learning.

8.3.2 PAC-Bayes for Off-Policy Learning

Let $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{A}\}$ be a hypothesis space of mappings from $\mathcal{X}$ (contexts) to $\mathcal{A}$ (actions). Given a policy π and a context $x \in \mathcal{X}$, the action distribution $\pi(\cdot|x)$ is induced

by a distribution $\mathbb{Q}$ over $\mathcal{H}$ (London and Sandler, 2019) such as

$$\pi(a|x) = \pi_{\mathbb{Q}}(a|x) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{1}_{\{h(x)=a\}}\right] . \tag{8.10}$$

This is not an assumption since any policy π has this form when $\mathcal{H}$ is rich enough (Sakhi et al., 2022, Theorem 2). From Equation (8.10), we observe that policies can be seen as an aggregation $\mathbb{E}_{h\sim\mathbb{Q}}[\cdot]$ (under some distribution $\mathbb{Q}$ on the pre-defined hypothesis space $\mathcal{H}$) of deterministic decision rules $\mathbb{1}_{\{h(x)=a\}}$. This allows formulating off-policy learning as a PAC-Bayes problem. Before showing how this is achieved, we start by providing two practical policies of such form.

Example 1 (softmax and mixed-logit policies). We define the hypothesis space $\mathcal{H} = \left\{h_{\theta,\gamma}\,;\theta\in\mathbb{R}^{dK}, \gamma\in\mathbb{R}^K\right\}$ of mappings $h_{\theta,\gamma}(x) = \operatorname{argmax}_{a\in\mathcal{A}} \phi(x)^\top\theta_a + \gamma_a$. Here $\phi(x)$ outputs a d-dimensional representation of x, and γ_a is a standard Gumbel perturbation, $\gamma_a \sim \mathrm{G}(0,1)$ for any $a\in\mathcal{A}$. Then

$$\begin{aligned}\pi_\theta^{\mathrm{SOF}}(a|x) &= \frac{\exp(\phi(x)^\top\theta_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\theta_{a'})}\,,\\ &\overset{(i)}{=} \mathbb{E}_{\gamma\sim\mathrm{G}(0,1)^K}\left[\mathbb{1}_{\{h_{\theta,\gamma}(x)=a\}}\right] ,\end{aligned} \tag{8.11}$$

where (i) follows from the Gumbel-Max trick (GMT) (Luce, 2012; Maddison et al., 2014). Thus a `softmax` policy $\pi_\theta^{\mathrm{SOF}}$ can be written as in Equation (8.10). Now we also consider random parameters $\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})$ with $\mu\in\mathbb{R}^{dK}$ and $\sigma>0$. Then, let $\mathbb{Q} = \mathcal{N}(\mu,\sigma^2 I_{dK})\times\mathrm{G}(0,1)^K$, it follows that $\pi_{\mathbb{Q}} = \pi_{\mu,\sigma}^{\mathrm{MIXL}}$ is a mixed-logit policy and it reads

$$\begin{aligned}\pi_{\mu,\sigma}^{\mathrm{MIXL}}(a|x) &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_d)}\left[\frac{\exp(\phi(x)^\top\theta_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\theta_{a'})}\right] ,\\ &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_d)\,,\gamma\sim\mathrm{G}(0,1)^K}\left[\mathbb{1}_{\{h_{\theta,\gamma}(x)=a\}}\right] .\end{aligned} \tag{8.12}$$

Example 2 (Gaussian policies): Sakhi et al. (2022) removed the Gumbel noise γ in Equation (8.12) and consequently defined the hypothesis space as $\mathcal{H} = \left\{h_\theta\,;\theta\in\mathbb{R}^{dK}\right\}$ of mappings $h_\theta(x) = \operatorname{argmax}_{a\in\mathcal{A}}\phi(x)^\top\theta_a$ for any $x\in\mathcal{X}$. Then, let $\mathbb{Q} = \mathcal{N}(\mu,\sigma^2 I_{dK})$, it follows that $\pi_{\mathbb{Q}} = \pi_{\mu,\sigma}^{\mathrm{GAUS}}$ reads

$$\pi_{\mu,\sigma}^{\mathrm{GAUS}}(a|x) = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_d)}\left[\mathbb{1}_{\{h_\theta(x)=a\}}\right] . \tag{8.13}$$

To see why removing the Gumbel noise can be beneficial, the reader may refer to Section E.3.2.

After motivating the definition of policies in Equation (8.10), we are in a position to relate our estimators to the general PAC-Bayes framework in Section 8.3.1. One technical requirement of our proof is that the estimator should be linear in π. Thus we focus on $\hat{V}^\alpha(\cdot)$ since $\tilde{V}^\beta(\pi)$ is non-linear in π. Let $h\in\mathcal{H}$, $x\in\mathcal{X}$, $a\in\mathcal{A}$ and $r\in[0,1]$, we define the objective U_α as

$$U_\alpha(h,x,a,r) = \frac{\mathbb{1}_{\{h(x)=a\}}}{\pi_0(a|x)^\alpha} r\,. \tag{8.14}$$

Using the definition in Equation (8.10) and the linearity of the expectation, we have that $\hat{V}^{\alpha}(\cdot)$ in Equation (8.8) can be written as

$$\hat{V}^{\alpha}(\pi_{\mathbb{Q}}) = \mathbb{E}_{h\sim\mathbb{Q}}\left[\frac{1}{n}\sum_{i=1}^{n} U_{\alpha}(h, X_i, A_i, R_i)\right] .$$

Moreover, the expectation of $\hat{V}(\pi_{\mathbb{Q}})$ reads

$$V^{\alpha}(\pi_{\mathbb{Q}}) = \mathbb{E}_{h\sim\mathbb{Q}}\mathbb{E}_{(X,A,R)\sim\mu_{\pi_0}}\left[U_{\alpha}(h, X, A, R)\right] .$$

Finally, the main quantity of interest, the value $V(\pi_{\mathbb{Q}})$, can be expressed in terms of the objective with $\alpha = 1$, U_1, as

$$V(\pi_{\mathbb{Q}}) = \mathbb{E}_{h\sim\mathbb{Q}}\mathbb{E}_{(X,A,R)\sim\mu_{\pi_0}}\left[U_1(h, X, A, R)\right] .$$

Since $\hat{V}^{\alpha}(\pi_{\mathbb{Q}})$ is an unbiased estimator of $V^{\alpha}(\pi_{\mathbb{Q}})$, PAC-Bayes can be used to bound $V^{\alpha}(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})$. This will allow bounding our quantity of interest $V(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})$.

8.3.3 Main Result

To ease the exposition, we assume that the rewards are deterministic. Then, in logged data $\mathcal{D}_n$, $R_i = r(X_i, A_i)$ for any $i \in [n]$. Note that the same result holds for stochastic rewards. We discuss our result and sketch its proof in Section 8.4. The complete proof can be found in Section E.2.1.

Theorem 4. *Let $\lambda > 0$, $n \geq 1$, $\delta \in (0, 1)$, $\alpha \in [0, 1]$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1 - \delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for any posterior $\mathbb{Q}$ on $\mathcal{H}$*

$$|V(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{KL_1(\pi_{\mathbb{Q}})}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{KL_2(\pi_{\mathbb{Q}})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}}) .$$

where $KL_1(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4\sqrt{n}}{\delta}$, *and*

$$KL_2(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{4}{\delta}, \qquad B_n^{\alpha}(\pi_{\mathbb{Q}}) = 1 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[\pi_0^{1-\alpha}(A|X_i)\right] ,$$

$$\operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(A_i|X_i)R_i^2}{\pi_0(A_i|X_i)^{2\alpha}}.$$

We start by clarifying that the prior $\mathbb{P}$ can be any fixed distribution on $\mathcal{H}$. If we have access to $\mathbb{P}_0$ on $\mathcal{H}$ such that $\pi_0 = \pi_{\mathbb{P}_0}$, then it is natural to set $\mathbb{P} = \mathbb{P}_0$. But this is just a choice and one may use priors that do not depend on π_0. Now we explain the main terms in our bound. First, the terms $\mathrm{KL}_1(\pi_{\mathbb{Q}})$ and $\mathrm{KL}_2(\pi_{\mathbb{Q}})$ contain the divergence $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})$ which penalizes posteriors $\mathbb{Q}$ that differ a lot from the prior $\mathbb{P}$. Moreover, $B_n^{\alpha}(\pi_{\mathbb{Q}})$ is the bias conditioned on the contexts $(X_i)_{i\in[n]}$; $B_n^{\alpha}(\pi_{\mathbb{Q}}) = 0$ when $\alpha = 1$ and $B_n^{\alpha}(\pi_{\mathbb{Q}}) > 0$ otherwise. Also, the first term in $\operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}})$ *resembles* the theoretical second moment of

the regularized importance weights $\frac{\pi}{\pi_0^\alpha}$ (without the reward) when they are seen as random variables. Similarly, the second term in $\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})$ *resembles* the empirical second moment of $\frac{\pi}{\pi_0^\alpha}R$ (with the reward). Finally, if $\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})$ is bounded, then we can set $\lambda = 1/\sqrt{n}$, in which case our bound scales as $\mathcal{O}(1/\sqrt{n} + B_n^\alpha(\pi_\mathbb{Q}))$. In practice, we set $\alpha \approx 1$ leading to $B_n^\alpha(\pi_\mathbb{Q}) \approx 0$ and the bound would scale as $\mathcal{O}(1/\sqrt{n})$.

One of the main strengths of our result is that it holds for standard IPS with $\alpha = 1$ under the assumption that $\mathrm{Var}_n^1(\pi_\mathbb{Q})$ is bounded. This assumption is less restrictive than assuming that the importance weight as a random variable, $\pi_\mathbb{Q}(A|X)/\pi_0(A|X)$, is bounded, a required assumption for traditional concentration bounds. In contrast, $\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})$ only involves the *expectations* of the random variables $\pi_\mathbb{Q}(A|X_i)/\pi_0(A|X_i)^{2\alpha}$, and ratios of π_0 evaluated at observed contexts and actions and $(X_i, A_i)_{i\in[n]}$, that have non-zero probabilities under π_0 by definition.

Our result holds for fixed $\lambda > 0$ and $\alpha \in [0, 1]$. In Section E.2.2, we extend this to any potentially data-dependent $\lambda \in (0, 1)$ and $\alpha \in (0, 1]$. The assumption that $R \in [0, 1]$ can be relaxed to $R \in [0, B]$ up to additional factors B^2 and B in $\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})$ and $\mathrm{KL}_1(\pi_\mathbb{Q})$, respectively. Finally, our bound is suitable for stochastic gradient ascent (Robbins and Monro, 1951) since data-dependent quantities are not inside a square root. This is important for scalability.

Limitations. Our bound in Theorem 4 has two main limitations. **(i)** Using it to directly derive a data-independent suboptimality gap bound is not straightforward. This difficulty arises because our bound involves empirical quantities such as $B_n^\alpha(\pi_\mathbb{Q})$ and $\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})$, whose dependence on the logged data prevents expressing the gap purely as a function of n. However, obtaining data-independent suboptimality guarantees was not the goal of this chapter. Instead, our focus was on deriving tractable and theoretically grounded bounds for exponential smoothing, that also perform well in practice when used for pessimistic objectives. **(ii)** Our result provides symmetric deviation bounds that simultaneously control the upper and lower deviations of $\hat{V}^\alpha(\pi_\mathbb{Q})$ from $V(\pi_\mathbb{Q})$. Yet, recent work (Gabbianelli et al., 2024), published after the paper corresponding to this chapter, indicates that the tails of regularized IPS estimators are inherently asymmetric. Consequently, tighter bounds may arise from developing asymmetric two-sided bounds that treat each deviation separately. We explored this direction in our follow-up work (Sakhi et al., 2024), where we derived some of the tightest bounds in the literature, with strong empirical performance.

8.3.4 Adaptive and Data-Driven Tuning of α

Theorem 4 assumes that α is fixed (although we extend it for data-dependent α in Section E.2.2). However, providing a procedure to tune α in an adaptive and data-dependent fashion is important in practice. Thus we propose to set

$$\alpha_* = \operatorname*{argmin}_{\alpha\in[0,1]} B_n^\alpha(\pi_\mathbb{Q}) + \sqrt{\frac{2\mathrm{KL}_2(\pi_\mathbb{Q})\,\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})}{n}}\,, \tag{8.15}$$

where all the terms are defined in Theorem 4. Roughly speaking, α_* establishes a bias-variance trade-off; it minimizes the sum of the bias term $B_n^\alpha(\pi_\mathbb{Q})$ and the square root of the second moment term $\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})$, weighted by $\sqrt{\frac{2\mathrm{KL}_2(\pi_\mathbb{Q})}{n}}$. Here Equation (8.15) is obtained

by minimizing the bound in Theorem 4 with respect to both α and λ as follows. First, we minimize the bound in Theorem 4 with respect to λ; the minimizer is $\lambda_* = \sqrt{\frac{2\mathrm{KL}_2(\pi_{\mathbb{Q}})}{n\,\mathrm{Var}_n^\alpha(\pi_{\mathbb{Q}})}}$. Then, the bound in Theorem 4 evaluated at $\lambda = \lambda_*$ becomes

$$\sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}}) + \sqrt{\frac{2\mathrm{KL}_2(\pi_{\mathbb{Q}})\,\mathrm{Var}_n^\alpha(\pi_{\mathbb{Q}})}{n}}\,. \tag{8.16}$$

Finally, α_* is defined as the minimizer of Equation (8.16) with respect to $\alpha \in [0,1]$, and $\sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}})}{2n}}$ does not appear in Equation (8.15) as it does not depend on α. Note that α_* depends on both logged data $\mathcal{D}_n$ and the learning policy $\pi_{\mathbb{Q}}$. Thus it is adaptive; its value changes in each iteration during optimization.

8.4 Discussion

We start by interpreting and comparing our results to related work. Then, we present the technical challenges in Section 8.4.2. After that, we sketch our proof in Section 8.4.3.

8.4.1 Interpretation and Comparison to Related Work

Theorem 4 gives insight into the number of samples needed so that the performance of $\hat{\pi}$ is close to that of the optimal policy π_*. To simplify the problem, we consider the Gaussian policies in Equation (8.13) and assume that there exists $\mathbb{Q}_* = \mathcal{N}(\mu_*, I_{dK})$ with $\mu_* \in \mathbb{R}^{dK}$ such that the optimal policy is $\pi_* = \pi_{\mathbb{Q}_*}$. Also, we let the prior $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$ and assume that π_0 is uniform. This is possible since as we said before, the prior $\mathbb{P}$ does not have to depend on the logging policy π_0. Then we have that $D_{\mathrm{KL}}(\mathbb{Q}_*\|\mathbb{P}) = \|\mu_* - \mu_0\|^2/2$, $B_n^\alpha(\pi_{\mathbb{Q}_*}) = 1 - 1/K^{1-\alpha}$ and $\mathrm{Var}_n^\alpha(\pi_{\mathbb{Q}_*}) \leq 2K^{2\alpha}$. The last inequality is not tight but it allows getting an easy-to-interpret term that does not depend on n. Now let $\epsilon > 2(1 - K^{\alpha-1})$ for $\alpha \in [1 - \log 2/\log K, 1]$. This condition on α ensures that $\epsilon \in [0,1]$ and it is mild as α is often close to 1. Then, it holds with high probability that

$$n \mathrel{\widetilde{>}} \Big(\frac{\|\mu_* - \mu_0\|^2 + K^{2\alpha}}{\epsilon - 2(1 - K^{\alpha-1})}\Big)^2 \implies V(\hat{\pi}) \geq V(\pi_{\mathbb{Q}_*}) - \epsilon\,,$$

where we omit constant and logarithmic terms in $\widetilde{>}$. This gives an intuition on the sample complexity for our procedure. In particular, fewer samples are needed in four cases. The first is when ϵ is large, which means that we afford to learn a policy whose performance is far from the optimal one. The second is when the prior $\mathbb{P}$ is close to $\mathbb{Q}_*$, that is when $\|\mu_* - \mu_0\|$ is small. This highlights that the choice of the prior $\mathbb{P}$ is important. The third is when the second-moment term $K^{2\alpha}$ is small. The fourth is when the bias $B_n^\alpha(\pi_{\mathbb{Q}_*})$ is small. In particular, when $\alpha = 1$, the bias is 0. In contrast, the second-moment term is minimized in $\alpha = 0$. This is where the choice of α matters. The proofs of these claims and more detail can be found in Section E.2.4.

Our chapter derives a *tractable generalization bound* for an estimator other than clipped IPS in Equation (8.7), which also holds for the standard IPS in Equation (8.1). The bounds in Swaminathan and Joachims (2015a); London and Sandler (2019); Sakhi et al.

(2022) have a multiplicative dependency on the clipping threshold (M or $1/\tau$ in Equation (8.7)). Standard IPS is recovered when $M \to \infty$ (or $\tau = 0$) in which case their bounds are infinite. We successfully avoid any similar dependency on α. Moreover, Swaminathan and Joachims (2015a); London and Sandler (2019) only used their generalization bounds to inspire pessimistic objectives. Although we directly optimize our theoretical bound (Theorem 4) in our experiments, our analysis also inspires a pessimistic objective where we simultaneously penalize the L_2 distance, the variance and the bias. That is, we find $\mu \in \mathbb{R}^{dK}$ that maximizes

$$\hat{V}^{\alpha}(\pi_{\mu}) - \lambda_1 \|\mu - \mu_0\|^2 - \lambda_2 \operatorname{Var}_n^{\alpha}(\pi_{\mu}) - \lambda_3 B_n^{\alpha}(\pi_{\mu}) \,. \tag{8.17}$$

Here λ_1, λ_2 and λ_3 are tunable hyper-parameters, π_{μ} can be the Gaussian policy in Equation (8.13), $\pi_{\mu} = \pi_{\mu,1}^{\text{GAUS}}$, with a fixed $\sigma = 1$, and μ_0 is the mean of the prior $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$. Existing works either penalize the L_2 distance or the variance. For completeness, we also show that this pessimistic objective should be preferred over existing ones in Section E.3.5.

8.4.2 Technical Challenges

London and Sandler (2019); Sakhi et al. (2022) derived PAC-Bayes generalization bounds for the estimator `IPS-max` in Equation (8.7). Extending their analyses to our case is not straightforward. First, their estimator `IPS-max` is upper bounded by $1/\tau$, and thus they relied on traditional techniques for $[0, 1]$-objectives (Alquier, 2021). In contrast, our objective in Equation (8.14) is not upper-bounded, and controlling it without assuming that the importance weights are bounded is challenging.

Moreover, their bounds have a multiplicative dependency on $1/\tau$, hence they explode as $\tau \to 0$. This makes them vacuous for small values of τ and inapplicable to the standard IPS estimator in Equation (8.1) recovered for $\tau = 0$. In contrast, our bound does not have a similar dependency on α and it is also valid for standard IPS recovered for $\alpha = 1$. Moreover, we derive two-sided inequalities rather than one-sided ones for the important reasons that we priorly discussed. This requires carefully controlling in *closed-form* the absolute value of the bias. Prior works only used that the bias is negative which was enough to obtain one-sided inequalities.

Explaining other challenges requires stating a result that inspired our analysis: Kuzborskij and Szepesvári (2019) derived PAC-Bayes generalization bounds for unbounded losses by only controlling their second moments. Recently, Haddouche and Guedj (2022) proposed a similar result using Ville's inequality (Bercu and Touati, 2008). Adapting their theorem to our problem is given Proposition 6. We slightly adapt their proof to get a *two-sided* inequality for a *negative* loss. The proof is deferred to Section E.2.3.

Proposition 6. *Let $\lambda > 0$, $n \geq 1$, $\delta \in (0, 1)$, $\alpha \in [0, 1]$ and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$*

$$|V^{\alpha}(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})| \leq \frac{D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{2}{\delta}}{\lambda n} + \frac{\lambda}{2n} \sum_{i=1}^{n} \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0^{2\alpha}(A_i|X_i)} R_i^2 + \frac{\lambda}{2} \mathbb{E}_{(X,A,R)\sim\mu_{\pi_0}} \left[\frac{\pi_{\mathbb{Q}}(A|X)}{\pi_0^{2\alpha}(A|X)} R^2 \right] , \tag{8.18}$$

There are two main issues with Proposition 6. First, the term $\mathbb{E}_{(X,A,R)\sim\mu_{\pi_0}}\left[\frac{\pi_{\mathbb{Q}}(A|X)}{\pi_0^{2\alpha}(A|X)}R^2\right]$ in Equation (8.18) is intractable. One could bound R^2 by 1, but the resulting term will still be intractable due to the expectation over the unknown distribution of contexts ν. Second, we need an upper bound of $|V(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})|$ while Proposition 6 only provides one for $|V^{\alpha}(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})|$. Thus it remains to quantify the approximation error $|V(\pi_{\mathbb{Q}}) - V^{\alpha}(\pi_{\mathbb{Q}})|$. This will also require computing an expectation over $X \sim \nu$, which is intractable.

8.4.3 Sketch of Proof for Theorem 4

We conclude by showing how the technical challenges above were solved. First, We decompose $V(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})$ as

$$V(\pi_{\mathbb{Q}}) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}}) = I_1 + I_2 + I_3\,, \qquad \text{where}$$

$$I_1 = V(\pi_{\mathbb{Q}}) - \frac{1}{n}\sum_{i=1}^{n} V(\pi_{\mathbb{Q}}|X_i)\,,$$

$$I_2 = \frac{1}{n}\sum_{i=1}^{n} V(\pi_{\mathbb{Q}}|X_i) - \frac{1}{n}\sum_{i=1}^{n} V^{\alpha}(\pi_{\mathbb{Q}}|X_i)\,,$$

$$I_3 = \frac{1}{n}\sum_{i=1}^{n} V^{\alpha}(\pi_{\mathbb{Q}}|X_i) - \hat{V}^{\alpha}(\pi_{\mathbb{Q}})\,,$$

where

$$V(\pi_{\mathbb{Q}}|X_i) = \mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[r(X_i,A)\right]\,, \quad V^{\alpha}(\pi_{\mathbb{Q}}|X_i) = \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{\alpha}}r(X_i,A)\right].$$

I_1 is the estimation error of the empirical mean of the value using n i.i.d. contexts $(X_i)_{i\in[n]}$. This term is introduced to avoid the intractable expectation over $X \sim \nu$. Moreover, I_2 is the bias term conditioned on the contexts $(X_i)_{i\in[n]}$ and we bound it in closed-form. Finally, I_3 is the estimation error of the value conditioned on the contexts $(X_i)_{i\in[n]}$. Again, this conditioning allows us to avoid the intractable expectation over $X \sim \nu$ and to consequently bound $|I_3|$ by tractable terms. First, Alquier (2021, Theorem 3.3) yields that with probability at least $1 - \frac{\delta}{2}$, it holds for any $\mathbb{Q}$ on $\mathcal{H}$ that

$$|I_1| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}}{2n}}\,.$$

Also, $|I_2|$ is bounded similarly to Equation (8.9) as

$$|I_2| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[1 - \pi_0^{1-\alpha}(A|X_i)\right]\,.$$

Bounding $|I_3|$ is achieved by expressing it using martingale difference sequences $(f_i(A_i,h))_{i\in[n]}$ that we construct as follows. Let $(\mathcal{F}_i)_{i\in\{0\}\cup[n]}$ be a filtration adapted to $(S_i)_{i\in[n]}$ where $S_i = (A_\ell)_{\ell\in[i]}$ for any $i \in [n]$, we define

$$f_i\left(A_i,h\right) = \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{1}_{\{h(X_i)=A\}}r(X_i,A)}{\pi_0(A|X_i)^{\alpha}}\right] - \frac{\mathbb{1}_{\{h(X_i)=A_i\}}R_i}{\pi_0(A_i|X_i)^{\alpha}}\,.$$

Then we show that for any $h \in \mathcal{H}$, $(f_i(A_i, h))_{i\in[n]}$ is a martingale difference sequence. After that, we apply Haddouche and Guedj (2022, Theorem 5) and obtain that with probability at least $1-\delta/2$, it holds for any $\mathbb{Q}$ on $\mathcal{H}$ that

$$|\mathbb{E}_{h\sim\mathbb{Q}}[M_n(h)]| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta}}{\lambda} + \frac{\lambda}{2}\mathbb{E}_{h\sim\mathbb{Q}}[\mathrm{Var}_n(h)]\,,$$

where $M_n(h) = \sum_{i=1}^n f_i(A_i,h)$ and $\mathrm{Var}_n(h) = \sum_{i=1}^n f_i(A_i,h)^2 + \mathbb{E}\left[f_i(A_i,h)^2\,|\mathcal{F}_{i-1}\right]$. Then notice that $\mathbb{E}_{h\sim\mathbb{Q}}[M_n(h)]$ can be expressed in terms of I_3 as

$$\mathbb{E}_{h\sim\mathbb{Q}}[M_n(h)] = \sum_{i=1}^n V^\alpha(\pi_\mathbb{Q}|X_i) - n\hat{V}^\alpha(\pi_\mathbb{Q}) = nI_3\,,$$

Moreover, $\mathbb{E}_{h\sim\mathbb{Q}}[\mathrm{Var}_n(h)]$ is bounded by

$$\sum_{i=1}^n \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_\mathbb{Q}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right] + \frac{\pi_\mathbb{Q}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}R_i^2\,.$$

Thus with probability at least $1-\frac{\delta}{2}$, it holds for any $\mathbb{Q}$ that

$$|I_3| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta}}{n\lambda} + \frac{\lambda}{2n}\sum_{i=1}^n \frac{\pi_\mathbb{Q}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}R_i^2 + \frac{\lambda}{2n}\sum_{i=1}^n \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_\mathbb{Q}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right]\,.$$

Our result is obtained by bounding $|I_1| + |I_2| + |I_3|$. One shortcoming of our analysis is that $\mathrm{Var}_n^\alpha(\pi_\mathbb{Q})$ is not exactly and only resembles the sum of the theoretical and empirical second moments of our estimator. Precisely, the terms $\pi_\mathbb{Q}/\pi_0^{2\alpha}$ should be $\pi_\mathbb{Q}^2/\pi_0^{2\alpha}$. This problem arises due to our definition of the martingale difference sequences $(f_i(A_i,h))_{i\in[n]}$ in Equation (8.14). Precisely, in our proof, we compute the square $f_i(A_i,h)^2$. However, the square of an indicator function is the indicator function itself. Thus applying the expectation afterwards, $\mathbb{E}_{h\sim\mathbb{Q}}[f_i(A_i,h)^2]$, leads to $\pi_\mathbb{Q}$ appearing instead of $\pi_\mathbb{Q}^2$. This issue is inherent in the PAC-Bayes formulation and seminal works (London and Sandler, 2019; Sakhi et al., 2022) would suffer the same issue. Solving this would be beneficial and we leave it to future work.

8.5 Experiments for Exponential Smoothing

We briefly present our experiments. More details and discussions can be found in Section E.3. We consider the standard supervised-to-bandit conversion (Agarwal et al., 2014) where we transform a supervised training set $\mathcal{S}_n^{\mathrm{TR}}$ to a logged bandit data $\mathcal{D}_n$ as described in Algorithm 3 in Section E.3.1. Here the action space $\mathcal{A}$ is the label set and the context space $\mathcal{X}$ is the input space. Then, $\mathcal{D}_n$ is used to train our policies. After that, we evaluate the value of the learned policies on the supervised test set $\mathcal{S}_{n_{\mathrm{TS}}}^{\mathrm{TS}}$ as described in Algorithm 4 in Section E.3.1. Roughly speaking, the resulting value quantifies the ability of the learned policy to predict the true labels of the inputs in the test set. This is our performance metric; the higher the better. We use 4 image classification datasets `MNIST` (LeCun et al., 1998), `FashionMNIST` (Xiao et al., 2017), `EMNIST` (Cohen et al., 2017) and `CIFAR100` (Krizhevsky et al., 2009).

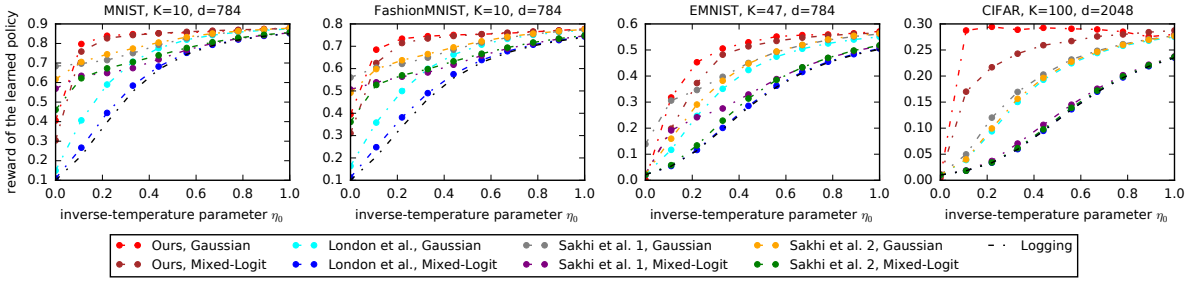


Figure 8.2: The reward of the learned policy using one of the baselines with varying quality of the logging policy $\eta_0 \in [0, 1]$.

The logging policy is defined as $\pi_0 = \pi^{\text{SOF}}_{\eta_0 \cdot \mu_0}$ in Equation (8.11), where $\mu_0 = (\mu_{0,a})_{a \in \mathcal{A}} \in \mathbb{R}^{dK}$ and $\eta_0 \in [0, 1]$ is the inverse-temperature parameter. The higher η_0, the better the performance of π_0. When $\eta_0 = 0$, π_0 is uniform. The parameters μ_0 are learned using 5% of the training set $\mathcal{S}_n^{\text{TR}}$. In our experiments, we consider both, Gaussian and mixed-logit policies, in Equation (8.12) and Equation (8.13), for which we set the prior as $\mathbb{P} = \mathcal{N}(\eta_0 \mu_0, I_{dK})$ and $\mathbb{P} = \mathcal{N}(\eta_0 \mu_0, I_{dK}) \times \mathrm{G}(0, 1)^K$, respectively. Given that μ_0 are learnt on 5% of $\mathcal{S}_n^{\text{TR}}$, we train our policies on the remaining 95% portion of $\mathcal{S}_n^{\text{TR}}$ to match our theory that requires the prior to not depend on training data. The policies are trained using Adam (Kingma and Ba, 2014) with a learning rate of 0.1 for 20 epochs.

Main results. We compare our bound to those in London and Sandler (2019); Sakhi et al. (2022); discarding the intractable bound in Swaminathan and Joachims (2015a) as it requires computing a covering number. Here we do not include the pessimistic objectives in Swaminathan and Joachims (2015a); London and Sandler (2019) since we directly optimize our bounds. But we make such a comparison in Section E.3.5 for completeness, showing the favorable performance of our bound and the newly proposed pessimistic objective in Equation (8.17). Also, we do not compare to Su et al. (2020); Metelli et al. (2021) since they do not provide generalization guarantees; they focus on estimation accuracy and only propose a heuristic for off-policy learning. However, we still show the favorable performance of our approach in off-policy learning compared to Su et al. (2020); Metelli et al. (2021) in Section E.3.6 for completeness.

Prior methods are not named. Thus we refer to them as **(Author, Policy)** where **Author** $\in$ {**Ours, London et al., Sakhi et al. 1, Sakhi et al. 2**} and **Policy** $\in$ {**Gaussian, Mixed-Logit**}. Here **Ours**, **London et al.**, **Sakhi et al. 1** and **Sakhi et al. 2** correspond to Theorem 4, London and Sandler (2019, Theorem 1), Sakhi et al. (2022, Proposition 1), and Sakhi et al. (2022, Proposition 3), respectively. Since we have two classes of policies, each bound leads to two baselines. For example, London and Sandler (2019, Theorem 1) leads to **(London et al., Gaussian)** and **(London et al., Mixed-Logit)**. More details are provided in Section E.3.3.

In Figure 8.2, we report the value of the learned policies. Here we fix $\tau = 1/\sqrt[4]{n} \approx 0.06$ and $\alpha = 1 - 1/\sqrt[4]{n} \approx 0.94$ so that when n is large enough, both $\hat{V}^{\tau}(\pi)$ and $\hat{V}^{\alpha}(\pi)$ approach $\hat{V}^{\text{IPS}}(\pi)$ (Ionides, 2008). This is because standard IPS should be preferred when $n \to \infty$. To have a fair comparison, we fixed α instead of tuning it in an adaptive fashion as described in Section 8.3.4. However, we also provide the results with an adaptive α

in Figure 8.3. Let us start with interpreting Figure 8.2 (with fixed α and τ). Overall, our method outperforms all the baselines. We also observe that Gaussian policies behave better than mixed-logit policies. However, this is less significant for our method where the performances of both Gaussian and mixed-logit policies are comparable. Moreover, our method reaches the maximum value even when the logging policy has an average performance. In contrast, the baselines only reach their best value when the logging policy is well-performing ($\eta_0 \approx 1$), in which case minor to no improvements are made. Finally, the baselines induce a better value when the logging policy is uniform ($\eta_0 = 0$). But our method has a better value when $\eta_0 > 0$, which is more common in practice.

Larger action spaces. The experiments above did not consider very large values of K. However, Chapter 7 evaluated IPS-based methods, including exponential smoothing and clippped IPS, on datasets with up to one million actions. In those experiments, exponential smoothing outperformed clipped IPS, though the improvements were modest compared to the gains observed here.

Choice of hyperparameters. Our choice of τ and α does not affect the above conclusions. In Figure 8.3 (left-hand side), we compare our method with the best baseline, **(Sakhi et al. 2)** with Gaussian policies, for 20 evenly spaced values of $\tau \in (0, 1)$ and $\alpha \in (0, 1)$. We also include the results using the adaptive tuning procedure of α described in Section 8.3.4 (green curve). This procedure is reliable since the performance with an adaptive α (green curve) is comparable with the best possible choice of α. Also, our method consistently outperforms the best baseline **(Sakhi et al. 2)** with the best value of τ when the logging policy is not uniform ($\eta_0 > 0$). Also, there is no very bad choice of α, in contrast with $\tau = 10^{-5}$ (dark blue plot) which led to minimal improvement upon all logging policies. This might be due to the $1/\tau$ dependency in existing bounds.

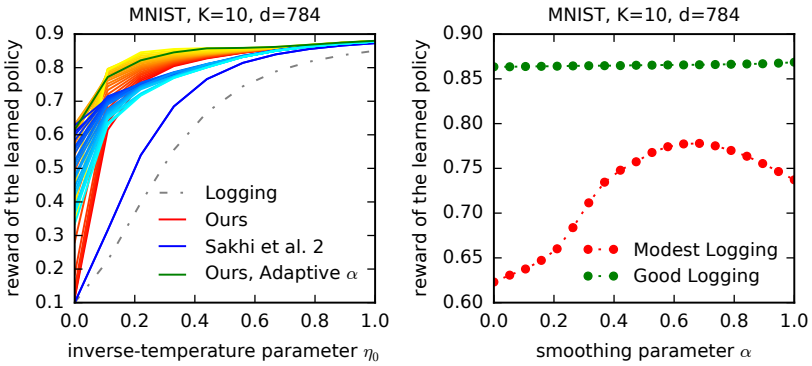


Figure 8.3: On the left-hand side is the reward of the learned policy with varying $\tau \in (0, 1)$, $\alpha \in (0, 1)$ and $\eta_0 \in [0, 1]$, and for an adaptive α using the procedure in Section 8.3.4 (green curve). The blue-to-cyan and red-to-yellow colors correspond to varying values of τ and α, respectively. The lighter the color, the higher the value of τ or α. The green curve corresponds to the reward of the learned policy with an adaptive and data-dependent α (Section 8.3.4). On the right-hand side is the *average* reward of the learned policies using our method across the modest and good logging groups, $\eta_0 \in [0, 0.5]$ (red) and $\eta_0 \in [0.5, 1]$ (green), respectively.

To see the effect of α, we consider the following experiment. We split the logging policies into two groups. The first is called *modest logging* which corresponds to logging policies π_0 whose η_0 is between 0 and 0.5. This group includes the uniform policy and other average-performing policies. The second is called *good logging* and it includes the logging policies whose η_0 is between 0.5 and 1. Then, for each α, we compute the average value of the learned policy, with that value of α, across these two groups. This leads to the two red and green curves in Figure 8.3 (right-hand side). Overall, we observe that $\alpha \approx 0.7$ leads to the best performance across the modest logging group. Thus when the performance of the logging policy is bad or average, which is common in practice, importance-weight regularization can be critical. In contrast, when the performance of the logging policy is already good and n is large enough, importance-weight regularization might not be needed and $\alpha \approx 1$ would also lead to good performance. This is one of the main strengths of our bound; it holds for the standard IPS recovered with $\alpha = 1$. This result goes against the belief that clipped IPS should always be preferred to standard IPS. Here, our bound applied to standard IPS outperformed clipping by a large margin when the logging policy is relatively well-performing. Similar results for the other datasets are deferred to Section E.3.4.

8.6 Extension to Other Regularizations

The experiments above demonstrated that exponential smoothing substantially outperforms clipping. However, we compared exponential smoothing with our pessimistic objective against clipping with pessimistic objectives specifically designed for it. This makes it difficult to isolate whether the gains stem from exponential smoothing as a regularization technique or from our pessimistic objective. Moreover, exponential smoothing and clipping are only two instances within a broader class of importance-weight regularizations.

While numerous methods have been proposed to stabilize IPS through importance-weight transformations (Bottou et al., 2013; Swaminathan and Joachims, 2015a; Su et al., 2020; Metelli et al., 2021), most focus on estimation accuracy rather than learning performance. As highlighted in Chapter 7, improved estimation does not necessarily yield improved policies, motivating a reassessment of importance-weight regularization specifically within the learning paradigm. Moreover, existing approaches followed a case-by-case basis: each regularization technique comes with its own theoretical analysis and corresponding pessimistic objective. This inconsistency makes it impossible to determine whether empirical improvements arise from the regularizer itself or from its specific objective formulation.

This reveals a critical gap: the absence of a unified framework providing principled pessimistic objectives across diverse importance-weight regularizations. We address this by developing a generic PAC-Bayesian generalization bound that applies uniformly to a broad family of regularizations, enabling fair comparison within a single theoretical framework.

Recall that the regularized IPS estimator has the form:

$$\hat{V}(\pi) = \frac{1}{n} \sum_{i=1}^{n} R_i \, \hat{w}(X_i, A_i), \tag{8.19}$$

where $\hat{w}(X, A)$ are the regularized importance weights. We further assume that $\hat{w}(X, A) = g(\pi(A|X), \pi_0(A|X))$ for some function $g : [0, 1] \times [0, 1] \to \mathbb{R}^+$. Examples of $\hat{w}$ include clipping (`Clip`) (London and Sandler, 2019), exponential smoothing (`ES`) (Aouali et al., 2023a), implicit exploration (`IX`) (Gabbianelli et al., 2024), and harmonic (`Har`) (Metelli et al., 2021), defined as

$$\begin{aligned}
\texttt{Clip}: \quad & \hat{w}(x, a) = \frac{\pi(a \mid x)}{\max(\pi_0(a \mid x), \tau)}, \ \tau \in [0, 1], \\
\texttt{ES}: \quad & \hat{w}(x, a) = \frac{\pi(a \mid x)}{\pi_0(a \mid x)^\alpha}, \ \alpha \in [0, 1], \\
\texttt{IX}: \quad & \hat{w}(x, a) = \frac{\pi(a \mid x)}{\pi_0(a \mid x) + \gamma}, \ \gamma \in [0, 1], \\
\texttt{Har}: \quad & \hat{w}(x, a) = \frac{w(x, a)}{(1 - \lambda)w(x, a) + \lambda}, \ \lambda \in [0, 1].
\end{aligned} \tag{8.20}$$

8.6.1 Generalization Bounds

PAC-Bayes theory (Section 8.3) allows bounding $\left|\mathbb{E}_{\theta\sim\mathbb{Q}}[V(\pi_\theta) - \hat{V}(\pi_\theta)]\right|$, with

$$V(\pi_\theta) = \mathbb{E}_{X\sim\nu, A\sim\pi_\theta(\cdot|X)}[r(X, A)], \qquad \hat{V}(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_\theta(X_i, A_i)R_i,$$

where we make the dependence of $\hat{w}_\theta$ on θ explicit to avoid confusion when taking the expectation $\mathbb{E}_{\theta\sim\mathbb{Q}}$. Below is our first general result that extends Theorem 4 to any regularization function g, instead of just exponential smoothing. Its proof follows exactly the same techniques we employed for Theorem 4.

Theorem 5. *Let $\lambda > 0$, $n \geq 1$, $\delta \in (0, 1)$, and let $\mathbb{P}$ be a fixed prior on Θ. The following inequality holds with probability at least $1 - \delta$ for any distribution $\mathbb{Q}$ on Θ:*

$$\left|\mathbb{E}_{\theta\sim\mathbb{Q}}[V(\pi_\theta) - \hat{V}(\pi_\theta)]\right| \leq \sqrt{\frac{KL_1(\pi_\mathbb{Q})}{2n}} + \frac{KL_2(\pi_\mathbb{Q})}{n\lambda} + B_n(\mathbb{Q}) + \frac{\lambda}{2}\mathrm{Var}_n(\mathbb{Q}), \tag{8.21}$$

where $KL_1(\pi_\mathbb{Q}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}$, $KL_2(\pi_\mathbb{Q}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta}$, and

$$\begin{aligned}
\mathrm{Var}_n(\mathbb{Q}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}[\hat{w}_\theta(X_i, A)^2] + \hat{w}_\theta(X_i, A_i)^2 R_i^2\right], \\
B_n(\mathbb{Q}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{A\in\mathcal{A}} \mathbb{E}_{\theta\sim\mathbb{Q}}\left[|\pi_\theta(A|X_i) - \pi_0(A|X_i)\hat{w}_\theta(X_i, A)|\right].
\end{aligned}$$

The terms in the above bound have similar interpretations to those in Theorem 4.

Linear vs. non-linear regularization. If $\hat{w}(X, A)$ is linear in $\pi_\theta(X, A)$ (i.e., g linear in its first variable), then $\hat{V}$ is also linear in π_θ, yielding

$$\left|\mathbb{E}_{\theta\sim\mathbb{Q}}[V(\pi_\theta) - \hat{V}(\pi_\theta)]\right| = \left|V(\pi_\mathbb{Q}) - \hat{V}(\pi_\mathbb{Q})\right|,$$

where we define (similar to Section 8.3.2)

$$\pi_{\mathbb{Q}} = \mathbb{E}_{\theta\sim\mathbb{Q}}[\pi_\theta]\,. \tag{8.22}$$

As seen in Section 8.3.2, this technique allows translating the bound in Theorem 5, which controls $\left|\mathbb{E}_{\theta\sim\mathbb{Q}}[V(\pi_\theta) - \hat{V}(\pi_\theta)]\right|$, into a bound that controls $|V(\pi_{\mathbb{Q}}) - \hat{V}(\pi_{\mathbb{Q}})|$, the quantity of interest in off-policy learning. The main requirement is to find linear importance-weight regularizations and policies that satisfy Equation (8.22). Fortunately, many importance-weight regularizations, such as `Clip`, `IX`, and `ES` in Equation (8.20), are linear in π, and several practical policies adhere to the formulation in Equation (8.22); see Section 8.3.2 for an in-depth explanation of such policies, including softmax, and Gaussian policies.

In Corollary 1, we specialize Theorem 5 to linear importance-weight regularizations of the form $\hat{w}_\theta(x,a) = \frac{\pi_\theta(a|x)}{h(\pi_0(a|x))}$, where $h(\pi_0(a|x)) \geq \pi_0(a|x)$ for all $(x,a) \in \mathcal{X}\times\mathcal{A}$. We additionally assume that the base policies π_θ are deterministic, i.e., $\pi_\theta(a \mid x) \in \{0,1\}$, which implies $\pi_\theta(a|x)^2 = \pi_\theta(a|x)$. This assumption is only needed here and it is mild: the PAC-Bayes policies $\pi_{\mathbb{Q}}$ defined in Equation (8.22) are mixtures of deterministic policies under $\mathbb{Q}$, and common policy classes such as softmax, mixed-logit, and Gaussian policies admit such representations (Section 8.3.2). Under these assumptions, Theorem 5 yields the following result.

Corollary 1. *Assume the regularized importance weights can be written as* $\hat{w}_\theta(x,a) = \frac{\pi_\theta(a|x)}{h(\pi_0(a|x))}$ *with* $h : [0,1] \to \mathbb{R}^+$ *verifies* $h(p) \geq p$ *for any* $p \in [0,1]$. *Moreover, for any distribution* $\mathbb{Q}$ *in the parameter space* Θ, *we define* $\pi_{\mathbb{Q}} = \mathbb{E}_{\theta\sim\mathbb{Q}}[\pi_\theta]$ *where* π_θ *is binary. Then, let* $\lambda > 0$, $n \geq 1$, $\delta \in (0,1)$, *and let* $\mathbb{P}$ *be a fixed prior on* Θ, *The following inequality holds with probability at least* $1-\delta$ *for any distribution* $\mathbb{Q}$ *on* Θ

$$\left|V(\pi_{\mathbb{Q}}) - \hat{V}(\pi_{\mathbb{Q}})\right| \leq \sqrt{\frac{KL_1(\mathbb{Q})}{2n}} + B_n(\pi_{\mathbb{Q}}) + \frac{KL_2(\mathbb{Q})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n(\pi_{\mathbb{Q}})\,, \tag{8.23}$$

where $KL_1(\mathbb{Q})$ *and* $KL_2(\mathbb{Q})$ *are defined in Theorem 5, and*

$$\operatorname{Var}_n(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{h(\pi_0(A|X_i))^2}\right] + \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{h(\pi_0(A_i|X_i))^2}R_i^2\,,$$

$$B_n(\pi_{\mathbb{Q}}) = 1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{A\in\mathcal{A}} \pi_0(A|X_i)\frac{\pi_{\mathbb{Q}}(A|X_i)}{h(\pi_0(A|X_i))}\,.$$

The main benefit of Corollary 1 compared to Theorem 5 is that it eliminates the need for the expectation $\mathbb{E}_{\theta\sim\mathbb{Q}}[\cdot]$, which is now embedded in the definition of policies in Equation (8.22). For example, Corollary 1 allows us to recover the main result of `ES` above Aouali et al. (2023a) when $h(p) = p^\alpha$, $\alpha \in [0,1]$. Similarly, we can apply it to `IX` (Gabbianelli et al., 2024) by setting $h(p) = p + \gamma$, $\gamma \geq 0$, and to `Clip` (London and Sandler, 2019) by setting $h(p) = \max(p,\tau)$, $\tau \in [0,1]$. However, if $\hat{w}_\theta(x,a)$ is not linear in $\pi_\theta(a|x)$, then this technique cannot be used, and the original expectation $\mathbb{E}_{\theta\sim\mathbb{Q}}[\cdot]$ in Theorem 5 must be retained.

8.6.2 Pessimistic Objectives

Theorem 5 yields two pessimistic objectives.

Bound optimization. The first approach directly maximizes the lower bound from Theorem 5:

$$\operatorname*{argmax}_{\mathbb{Q}} \mathbb{E}_{\theta\sim\mathbb{Q}}\left[\hat{V}(\pi_\theta)\right] - \sqrt{\frac{\mathrm{KL}_1(\mathbb{Q})}{2n}} - B_n(\mathbb{Q}) - \frac{\mathrm{KL}_2(\mathbb{Q})}{n\lambda} - \frac{\lambda}{2}\operatorname{Var}_n(\mathbb{Q})\,, \tag{8.24}$$

The main challenge is that the objective involves expectations under $\mathbb{Q}$. We address this using the *local reparameterization trick* (Kingma et al., 2015), which expresses gradients of expectations as expectations of gradients, estimated via Monte Carlo sampling. Specifically, we consider softmax policies $\pi_\theta^{\mathrm{SOF}}(a|x)$ from Equation (8.11) and set $\mathbb{Q} = \mathcal{N}(\mu, \sigma^2 I_{dK})$ with learnable parameters $\mu \in \mathbb{R}^{dK}$ and $\sigma > 0$. All terms in Equation (8.24) take the form $\mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}[f(\pi_\theta^{\mathrm{SOF}}(a|x))]$, which can be rewritten as:

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}\left[f(\pi_\theta^{\mathrm{SOF}}(a|x))\right] \\
&= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\|\phi(x)\|_2^2 I_K)}\left[f\left(\frac{\exp(\phi(x)^\top\mu_a + \sigma\epsilon_a)}{\sum_{a'\in\mathcal{A}}\exp(\phi(x)^\top\mu_{a'} + \sigma\epsilon_{a'})}\right)\right].
\end{aligned}$$

This expectation is approximated by sampling $\epsilon_i \sim \mathcal{N}(0, \|\phi(x)\|_2^2 I_K)$ and computing the empirical mean; gradients are estimated similarly. However, this approach can exhibit high variance when K is large. For linear importance-weight regularizations, this can be mitigated by optimizing the bound in Corollary 1. For the general case, we propose a practical alternative.

Heuristic optimization. The second approach avoids the challenges of direct bound optimization at the cost of additional hyperparameters. Inspired by Theorem 5, we maximize the estimated value penalized by bias, variance, and proximity to the logging policy:

$$\hat{V}(\pi_\theta) - \lambda_1\|\theta - \theta_0\|^2 - \lambda_2\tilde{\operatorname{Var}}_n(\pi_\theta) - \lambda_3\tilde{B}_n(\pi_\theta)\,, \tag{8.25}$$

where $\tilde{\operatorname{Var}}_n(\pi_\theta)$ and $\tilde{B}_n(\pi_\theta)$ are the terms inside the expectations in $\operatorname{Var}_n(\mathbb{Q})$ and $B_n(\mathbb{Q})$, respectively, θ_0 parameterizes the logging policy π_0, and $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters.

Both objectives in Equations (8.24) and (8.25) are amenable to stochastic gradient optimization and are generic across importance-weight regularizations, enabling fair comparison. We empirically compare these objectives and evaluate different regularization techniques in Section 8.7.

8.7 Experiments for Other Regularizations

We adopt the experimental setting of Section 8.5 and conduct two main experiments. In Section 8.7.1, we fix the importance-weight regularization to `Clip` (Equation (8.20)) and compare our pessimistic objective against PAC-Bayesian objectives from the literature specifically designed for clipping. The goal is to demonstrate that our objective not

only applies more broadly but also outperforms existing alternatives. In Section 8.7.2, having validated our pessimistic objective, we fix it and compare across importance-weight regularizations. The goal is to determine whether any particular regularization technique yields superior off-policy learning performance.

8.7.1 Varying Pessimistic Objectives, Fixed Regularization

We examine the impact of different pessimistic objectives on learned policy performance, fixing the importance-weight regularization to `Clip`: $\hat{w}(x,a) = \frac{\pi(a|x)}{\max(\pi_0(a|x),\tau)}$ in Equation (8.20), with $\tau = 1/\sqrt[4]{n}$ following Ionides (2008). For fair comparison, we consider PAC-Bayesian objectives from prior work where the theoretical bound is optimized directly. Specifically, we include **London et al.** (London and Sandler, 2019, Theorem 1), and two bounds from Sakhi et al. (2022): **Sakhi et al. 1** (Sakhi et al., 2022, Proposition 1), based on Catoni (2007), and **Sakhi et al. 2** (Sakhi et al., 2022, Proposition 3), a Bernstein-type bound. Since these baselines use linear importance-weight regularization (Section 8.6.1), we compare against our bound in Corollary 1. Following Sakhi et al. (2022); Aouali et al. (2023a), we optimize over Gaussian policies (Equation (8.13)), which perform better in this setting. We also include the logging policy as a baseline.

Figure 8.4 plots the reward of learned policies as a function of logging policy quality $\eta_0 \in [0,1]$. Our objective outperforms all baselines across a wide range of logging policies. Thus, in addition to being generic across importance-weight regularizations, our approach proves more effective than objectives tailored specifically for `Clip`. This advantage holds when η_0 is not too close to zero: a realistic scenario where logging policies typically outperform uniform random selection. Note that all methods (including ours) improve upon the logging policy (dashed black lines).

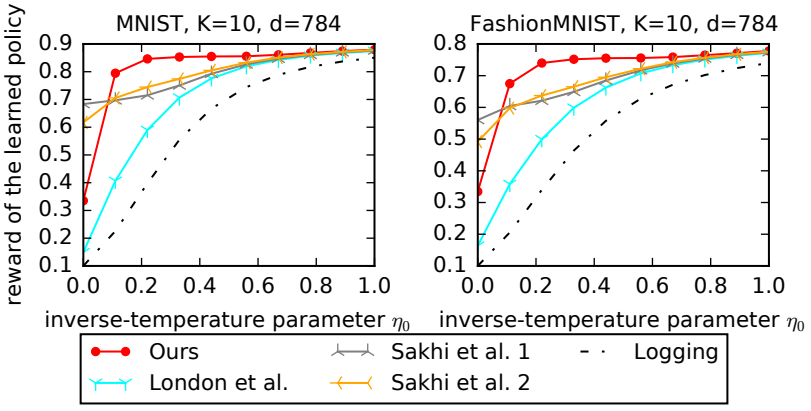


Figure 8.4: Performance of the learned policy with different PAC-Bayes pessimistic objectives (our Corollary 1 and those in London and Sandler (2019); Sakhi et al. (2022)) using the `Clip` IPS estimator in Equation (8.20) .

8.7.2 Varying Regularization, Fixed Pessimistic Objective

Having demonstrated the favorable performance of our pessimistic objective, we now compare different importance-weight regularization techniques: `Clip`, `Har`, `IX`, and `ES`

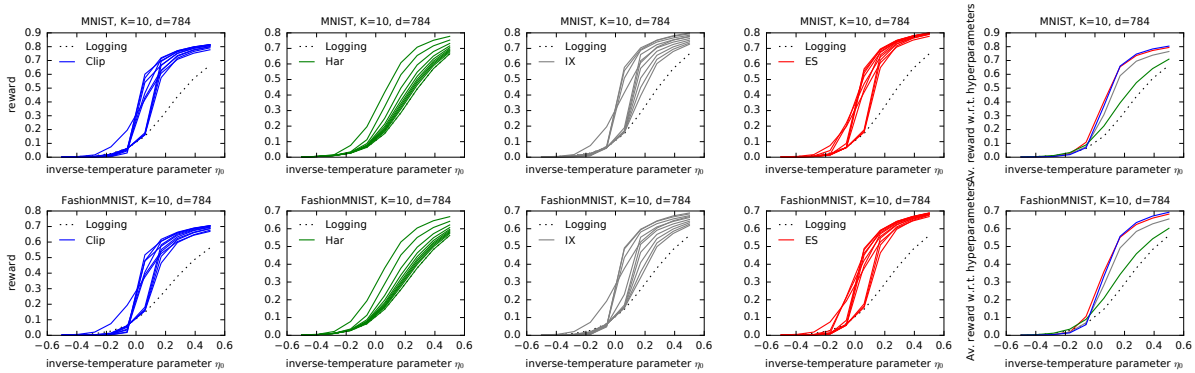


Figure 8.5: Performance of the policy learned by **Bound optimization** (i.e, Equation (8.24)) for different importance-weight regularizations. The x-axis reflects the quality of the logging policy $\eta_0 \in [-0.5, 0.5]$. In the first four columns, we plot the reward of the learned policy using a fixed importance-weight regularization technique (`Clip`, `Har`, `IX`, or `ES` as defined in Equation (8.20)) for various values of its hyperparameter within $[0, 1]$. In the last column, we report the mean reward across these hyperparameter values.

(Equation (8.20)). We evaluate both pessimistic objectives from Section 8.6.2, optimizing over softmax policies. For bound optimization, we use Theorem 5 rather than Corollary 1, since `Har` is non-linear in π. We set λ to its optimal value λ_* minimizing the bound. While our theory requires λ to be fixed a priori (since λ_* is data-dependent), we found this yields good empirical performance. For heuristic optimization (Equation (8.25)), we set $\lambda_1 = \lambda_2 = \lambda_3 = 10^{-5}$.

Figures 8.5 and 8.6 present learned policy rewards as a function of logging policy quality η_0, for bound optimization and heuristic optimization respectively. We vary $\eta_0 \in [-0.5, 0.5]$, including logging policies worse than uniform ($\eta_0 < 0$) to highlight settings requiring stronger regularization, though such scenarios are rarely encountered in practice. Rows correspond to `MNIST` and `FashionMNIST`. The first four columns show results for each regularization technique across hyperparameter values in $[0, 1]$; the last column reports mean reward across hyperparameters for each regularization technique to assess sensitivity to hyperparameters.

Bound optimization (Figure 8.5). All regularizations improve over the logging policy (all curves above the dashed baseline), with `Har` showing less improvement. `Clip`, `IX`, and `ES` achieve comparable performance despite regularizing importance weights differently. These results align with the generality of our bound and suggest that the choice of regularization has limited impact when optimizing the theoretical bound directly.

Heuristic optimization (Figure 8.6). Heuristic optimization achieves better performance than bound optimization, likely due to practical limitations of Monte Carlo estimation in high dimensions (Section 8.6.2). The far-right column reveals comparable average performance across regularizations, with two exceptions: `ES` outperforms the others while `Har` underperforms. This clarifies our results from Section 8.5: the superior performance of exponential smoothing is more related to our pessimistic objective than the smooth regularization itself. Here, the smooth regularization adds some improvements compared

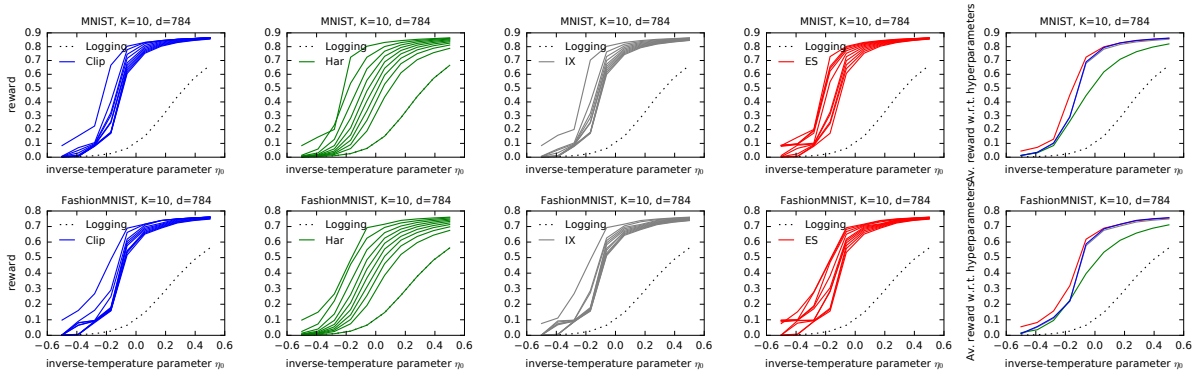


Figure 8.6: Performance of the policy learned by **Heuristic optimization** in Equation (8.25) for different importance-weight regularizations. The x-axis reflects the quality of the logging policy $\eta_0 \in [-0.5, 0.5]$. In the first four columns, we plot the reward of the learned policy using a fixed importance-weight regularization technique (`Clip`, `Har`, `IX`, or `ES` as defined in Equation (8.20)) for various values of its hyperparameter within $[0, 1]$. In the last column, we report the mean reward across these hyperparameter values.

to others, but the improvments are not significant compared to the improvments we get by simply changing the pessimistic learnin principle even with the standard clipping regularization(Figure 8.4)

Larger action spaces. The experiments in this section did not consider very large values of K. However, Chapter 7 evaluated numerous IPS-based methods on datasets with up to one million actions. In those experiments, exponential smoothing outperformed other IPS-based methods, though the improvements were modest. Combined with the results above, this reinforces our conclusion: the choice of the objective has a larger impact on learning performance than the choice of importance-weight regularization, which primarily affects estimation accuracy.

8.8 Conclusion

In this chapter, we investigated importance-weight regularization techniques within the pessimistic paradigm, with particular focus on exponential smoothing as a principled alternative to hard clipping. Our key contributions include: (i) tractable two-sided PAC-Bayes generalization bounds that, unlike prior work, apply to both regularized and standard IPS estimators and are amenable to stochastic gradient optimization; and (ii) the first unified framework for comparing diverse importance-weight regularizations under a common pessimistic objective. This work addresses fundamental theoretical limitations in existing approaches, including the reliance on one-sided inequalities and the misapplication of evaluation bounds in off-policy learning. Rather than using theoretical bounds merely as inspiration for heuristics, we directly optimize them, representing a first step toward making IPS-based pessimism more practical.

Our work has two primary limitations. First, the inclusion of empirical bias and variance terms in our bounds makes deriving data-independent suboptimality gaps challenging.

Second, two-sided bounds for regularized IPS can be loose as they treat both tails symmetrically, whereas recent work indicates significant asymmetry between lower and upper tails. We address both limitations in Sakhi et al. (2024), which investigates tail-specific bounds to achieve significantly tighter guarantees and sharp suboptimality results.

This chapter serves practitioners committed to IPS-based methods, whose appeal is well-founded: unbiasedness, theoretical guarantees, and the ability to derive principled pessimistic objectives for safe policy learning in high-stakes scenarios. However, from a purely empirical perspective, particularly in very large action spaces, we would favor the PWLL objectives introduced in Chapter 7, which consistently outperform IPS-based methods. The reader can find a direct comparison of these approaches on datasets with up to one million actions in that chapter.

Chapter 9

Conclusions and Future Work

This thesis addressed, from both a practical and theoretical perspective, the core obstacle to deploying contextual bandits in modern applications: *scalability to large action spaces while maintaining computational tractability.*

On-Policy Learning (Part I). We introduced structured Bayesian models that enable principled information sharing across actions and derived scalable exploration algorithms. `meTS` in Chapter 3 couples action parameters through shared latent effects, yielding regret and complexity that scale with an *effective* number of actions rather than K. `dTS` in Chapter 4 further develops this idea by using a pre-trained diffusion model to encode richer structure. These algorithms perform well in practice in their theoretical form, without additional tweaks or hyperparameter tuning.

Off-Policy Learning (Part II). We tackled both pillars: DM and IPS approaches. `sDM` in Chapter 6 extends the structured modeling of Part I to the offline regime. We then showed in Chapter 7 that, in large action spaces, *optimization* matters more than estimation: estimator-based objectives induce highly non-concave landscapes, whereas policy-weighted log-likelihoods produce concave objectives for common policy classes and win decisively at scale. Finally, we developed a principled pessimistic framework for regularized IPS in Chapter 8: smooth importance-weight regularization (exponential smoothing) paired with two-sided PAC-Bayes bounds, and a unified analysis that clarifies when regularization matters and how to compare it across methods. Our additional work on *logarithmic smoothing* (Sakhi et al., 2024) sharpens the concentration analysis further and yields tighter learning guarantees.

Thesis message and practical implications. Scaling to large action spaces causes methods that perform well in small settings (e.g., standard IPS) to fail at scale. This thesis advances three design principles to address this challenge: (i) encode *structure* to shrink the effective action space; (ii) prioritize *objectives with favorable optimization properties* over faithful but intractable estimators; and (iii) when relying on IPS with pessimism, couple *differentiable* importance-weight corrections with theoretically grounded, *data-driven* bounds amenable to stochastic gradient descent. Together, these principles yield algorithms that are statistically efficient, computationally tractable, and numerically stable.

Future work. This thesis opens several promising directions for future research. A key theoretical challenge is to establish *robust guarantees under model misspecification*, extending the Bayesian analysis of `sDM`, `meTS`, and `dTS` beyond the well-specified setting. For on-policy learning, developing a comprehensive *nonlinear diffusion theory for Thompson sampling* remains an open problem. In the off-policy setting, future work could investigate the extensions and applications of our methods to LLM and diffusion model fine-tuning, where the objective closely mirrors offline contextual bandit objectives. Moreover, integrating these approaches into *large-scale recommender pipelines* requires efficient action retrieval, slate constraints, and systems-level optimization. Some of these aspects, such as coupling decision-making with approximate maximum inner product search, were explored in our applied studies (see Additional Contributions in Section 1.3) but omitted from this manuscript. These practical directions have a tangible impact on the online advertising industry and beyond, and are worth pursuing.

Chapter A

Supplementary Materials for Chapter 3

A.1 Preliminaries

In this section, we recall some basic properties of matrix operations.

(a) **The mixed-product property.** We have that $(\mathrm{A} \otimes \mathrm{B})(\mathrm{C} \otimes \mathrm{D}) = \mathrm{AC} \otimes \mathrm{BD}$ for any matrices $\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}$ such that the products AC and BD exist.

(b) **Transpose.** We have that $(\mathrm{A} \otimes \mathrm{B})^\top = \mathrm{A}^\top \otimes \mathrm{B}^\top$ for any matrices A, B.

(c) **Vectorization.** Let $\mathrm{A} \in \mathbb{R}^{n \times m}, \mathrm{B} \in \mathbb{R}^{m \times p}$, then $\mathrm{Vec}(\mathrm{AB}) = (I_p \otimes \mathrm{A})\,\mathrm{Vec}(\mathrm{B}) = (\mathrm{B}^\top \otimes I_n)\,\mathrm{Vec}(\mathrm{A})$.

(d) For any matrix A, we have that $I_1 \otimes \mathrm{A} = \mathrm{A}$.

(e) For any positive semi-definite matrices A and B, we have that $\lambda_1(\mathrm{A} \otimes \mathrm{B}) = \lambda_1(\mathrm{A})\lambda_1(\mathrm{B})$.

(f) For any matrix A and any positive semi-definite matrix B such that the product $\mathrm{A}^\top \mathrm{BA}$ exists, the following inequality holds $\lambda_1(\mathrm{A}^\top \mathrm{BA}) \leq \lambda_1(\mathrm{B})\lambda_1(\mathrm{A}^\top \mathrm{A})$.

A.2 Posterior Derivations

Here we provide the derivations of the effect posterior and action posteriors for the setting presented in Section 3.1.1. Precisely, we present the proof for Proposition 1 in Section A.2.1 and the proof of Proposition 2 in Section A.2.2.

A.2.1 Effect Posterior Derivation

Proof of Proposition 1 (derivation of q_t). First, from basic properties of matrix operations, we observe that the mean of the action parameter can be rewritten using Kronecker products. Specifically, $\sum_{\ell \in [L]} b_{a,\ell} \psi_\ell = \Gamma_a \Psi$, where $\Psi = (\psi_\ell)_{\ell \in [L]} \in \mathbb{R}^{Ld}$ is the concatenated effect vector and $\Gamma_a = b_a^\top \otimes I_d \in \mathbb{R}^{d \times Ld}$. Thus, the model in Equation (3.2) (up to round

$t \in [T]$) can be written as

$$\begin{aligned} \Psi &\sim \mathcal{N}(\mu_\Psi, \Sigma_\Psi)\,, \\ \theta_a \mid \Psi &\sim \mathcal{N}(\Gamma_a \Psi, \Sigma_{0,a})\,, && \forall a \in \mathcal{A}\,, \\ R_i \mid X_i, A_i, \theta, \Psi &\sim \mathcal{N}(X_i^\top \theta_{A_i}, \sigma^2)\,, && \forall i \in [t-1]\,. \end{aligned} \tag{A.1}$$

Under this model, conditional on $(\theta_a)_{a\in\mathcal{A}}$ and $(X_i, A_i)_{i<t}$, the rewards $(R_i)_{i<t}$ are independent and each R_i depends on Ψ only through θ_{A_i}. Hence

$$p((R_i)_{i<t} \mid (X_i, A_i)_{i<t}, \Psi) = \int_{\theta\in\mathbb{R}^{dK}} p((R_i)_{i<t} \mid (X_i, A_i)_{i<t}, \theta)\, p(\theta \mid \Psi)\, d\theta.$$

Moreover, since $p(\theta \mid \Psi) = \prod_{a\in\mathcal{A}} p_{0,a}(\theta_a \mid \Psi)$ and

$$p((R_i)_{i<t} \mid (X_i, A_i)_{i<t}, \theta) = \prod_{a\in\mathcal{A}} \prod_{i\in S_{t,a}} \mathcal{N}(R_i; X_i^\top \theta_a, \sigma^2) = \prod_{a\in\mathcal{A}} \mathcal{L}_{t,a}(\theta_a),$$

the integral factorizes across arms:

$$p((R_i)_{i<t} \mid (X_i, A_i)_{i<t}, \Psi) = \prod_{a\in\mathcal{A}} \int \mathcal{L}_{t,a}(\theta_a)\, p_{0,a}(\theta_a \mid \Psi)\, d\theta_a.$$

It follows that the joint effect posterior in round t reads

$$\begin{aligned} q_t(\Psi) \;&\propto\; p((R_i)_{i<t} \mid (X_i, A_i)_{i<t}, \Psi)\, q_0(\Psi)\,, && \text{(A.2)} \\ &= \prod_{a\in\mathcal{A}} \int_{\theta_a} \mathcal{L}_{t,a}(\theta_a) p_{0,a}(\theta_a \mid \Psi)\, \mathrm{d}\theta_a\, q_0(\Psi) \\ &= \prod_{a\in\mathcal{A}} \underbrace{\int_{\theta_a} \mathcal{L}_{t,a}(\theta_a) \mathcal{N}\left(\theta_a; \Gamma_a \Psi, \Sigma_{0,a}\right) \mathrm{d}\theta_a}_{\mathcal{I}_a(\Psi)}\, \mathcal{N}(\Psi; \mu_\Psi, \Sigma_\Psi)\,, && \text{(A.3)} \end{aligned}$$

where $\mathcal{L}_{t,a}(\theta_a) = \prod_{i\in S_{t,a}} \mathcal{N}(R_i; X_i^\top \theta_a, \sigma^2)$.

We compute the integral term $\mathcal{I}_a(\Psi)$ using Lemma 1. Specifically, we obtain that $\mathcal{I}_a(\Psi)$ is proportional to a Gaussian density on Ψ, denoted $\mathcal{N}(\Psi; \bar{\mu}_{t,a}, \bar{\Sigma}_{t,a})$, where

$$\begin{aligned} \bar{\Sigma}_{t,a}^{-1} &= \Gamma_a^\top \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right) \Gamma_a\,, \\ \bar{\mu}_{t,a} &= \bar{\Sigma}_{t,a} \left(\Gamma_a^\top \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1} B_{t,a}\right)\,, \end{aligned}$$

and $G_{t,a}$ and $B_{t,a}$ are defined in Section 3.2.2. Consequently, the effect posterior $q_t(\Psi)$ is proportional to the product of $K+1$ multivariate Gaussian distributions: the prior $\mathcal{N}(\mu_\Psi, \Sigma_\Psi)$ and the likelihood contributions $\mathcal{N}(\bar{\mu}_{t,a}, \bar{\Sigma}_{t,a})$ for each $a \in \mathcal{A}$. Since the product of Gaussians is Gaussian, $q_t = \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t)$, where the precision matrix is the sum of the individual precisions:

$$\bar{\Sigma}_t^{-1} = \Sigma_\Psi^{-1} + \sum_{a\in\mathcal{A}} \bar{\Sigma}_{t,a}^{-1} = \Sigma_\Psi^{-1} + \sum_{a\in\mathcal{A}} \Gamma_a^\top \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right) \Gamma_a\,.$$

Using that $\Gamma_a = b_a^\top \otimes I_d$, we rewrite the term inside the sum as:

$$\begin{aligned}\Gamma_a^\top \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right)\Gamma_a &= (b_a \otimes I_d)\left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right)(b_a^\top \otimes I_d) \\ &= (b_a b_a^\top) \otimes \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right) .\end{aligned}$$

Similarly, for the mean $\bar{\mu}_t$, we have:

$$\begin{aligned}\bar{\mu}_t &= \bar{\Sigma}_t \left(\Sigma_\Psi^{-1}\mu_\Psi + \sum_{a\in\mathcal{A}} \bar{\Sigma}_{t,a}^{-1}\bar{\mu}_{t,a}\right) \\ &= \bar{\Sigma}_t \left(\Sigma_\Psi^{-1}\mu_\Psi + \sum_{a\in\mathcal{A}} \Gamma_a^\top\Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}B_{t,a}\right) .\end{aligned}$$

Using the mixed (Kronecker) product property,

$$\Gamma_a^\top\Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}B_{t,a} = b_a \otimes \left(\Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}B_{t,a}\right) .$$

This recovers the expressions in Proposition 1. □

To reduce clutter in the following lemma, we fix an action $a \in \mathcal{A}$ and a round t. We drop the sub-indices a and t, so that we have the following correspondences:

$$\Gamma \leftarrow \Gamma_a\,, \quad \Sigma_0 \leftarrow \Sigma_{0,a}\,, \quad N \leftarrow N_{t,a}\,, \quad \theta \leftarrow \theta_a\,, \quad (X_i, R_i)_{i\in[N]} \leftarrow (X_i, R_i)_{i\in S_{t,a}}\,,$$

Lemma 1 (Gaussian posterior update)**.** *Let $\Gamma \in \mathbb{R}^{d\times Ld}$, $\Sigma_0 \in \mathbb{R}^{d\times d}$, and $\sigma > 0$. Consider a dataset of N observations $(X_i, R_i)_{i=1}^N$. Then,*

$$\int_\theta \left(\prod_{i=1}^N \mathcal{N}(R_i; X_i^\top\theta, \sigma^2)\right)\mathcal{N}(\theta; \Gamma\Psi, \Sigma_0)\,\mathrm{d}\theta \propto \mathcal{N}\left(\Psi; \mu_N, \Sigma_N\right) ,$$

where

$$\begin{aligned}\Sigma_N^{-1} &= \Gamma^\top\left(\Sigma_0^{-1} - \Sigma_0^{-1}\left(G_N + \Sigma_0^{-1}\right)^{-1}\Sigma_0^{-1}\right)\Gamma\,, \\ \Sigma_N^{-1}\mu_N &= \left(\Gamma^\top\Sigma_0^{-1}\left(G_N + \Sigma_0^{-1}\right)^{-1}B_N\right) .\end{aligned}$$

with $G_N = \sigma^{-2}\sum_{i=1}^N X_iX_i^\top$ and $B_N = \sigma^{-2}\sum_{i=1}^N R_iX_i$.

Proof. Let $v = \sigma^{-2}$ and $\Lambda_0 = \Sigma_0^{-1}$. We denote the integral in the lemma by $f(\Psi)$. Completing the square for θ, we have:

$$\begin{aligned}f(\Psi) &\propto \int_\theta \exp\left[-\frac{v}{2}\sum_{i=1}^N (R_i - X_i^\top\theta)^2 - \frac{1}{2}(\theta - \Gamma\Psi)^\top\Lambda_0(\theta - \Gamma\Psi)\right]\mathrm{d}\theta \\ &\propto \int_\theta \exp\Big[-\frac{1}{2}\Big(\theta^\top \underbrace{\left(v\sum_{i=1}^N X_iX_i^\top + \Lambda_0\right)}_{V_N^{-1}}\theta - 2\theta^\top\left(v\sum_{i=1}^N R_iX_i + \Lambda_0\Gamma\Psi\right) \\ &\qquad\qquad + (\Gamma\Psi)^\top\Lambda_0(\Gamma\Psi)\Big)\Big]\,\mathrm{d}\theta\,.\end{aligned}$$

To reduce clutter, let

$$G_N = v\sum_{i=1}^{N} X_i X_i^\top\,, \quad V_N = (G_N + \Lambda_0)^{-1}\,, \quad U_N = V_N^{-1}\,,$$
$$B_N = v\sum_{i=1}^{N} R_i X_i \quad \text{and} \quad \beta_N = V_N\left(B_N + \Lambda_0\Gamma\Psi\right)\,.$$

We have that $U_N V_N = V_N U_N = I_d$, and thus

$$\begin{aligned}
f(\Psi) &\propto \int_\theta \exp\left[-\frac{1}{2}\left(\theta^\top U_N\theta - 2\theta^\top U_N V_N\left(B_N + \Lambda_0\Gamma\Psi\right) + (\Gamma\Psi)^\top\Lambda_0(\Gamma\Psi)\right)\right]\mathrm{d}\theta\,,\\
&= \int_\theta \exp\left[-\frac{1}{2}\left(\theta^\top U_N\theta - 2\theta^\top U_N\beta_N + (\Gamma\Psi)^\top\Lambda_0(\Gamma\Psi)\right)\right]\mathrm{d}\theta\,,\\
&= \int_\theta \exp\left[-\frac{1}{2}\left((\theta-\beta_N)^\top U_N(\theta-\beta_N) - \beta_N^\top U_N\beta_N + (\Gamma\Psi)^\top\Lambda_0(\Gamma\Psi)\right)\right]\mathrm{d}\theta\,,\\
&\propto \exp\left[-\frac{1}{2}\left(-\beta_N^\top U_N\beta_N + (\Gamma\Psi)^\top\Lambda_0(\Gamma\Psi)\right)\right]\,,\\
&= \exp\left[-\frac{1}{2}\left(-\left(B_N + \Lambda_0\Gamma\Psi\right)^\top V_N\left(B_N + \Lambda_0\Gamma\Psi\right) + (\Gamma\Psi)^\top\Lambda_0(\Gamma\Psi)\right)\right]\,,\\
&\propto \exp\left[-\frac{1}{2}\left(\Psi^\top\Gamma^\top\left(\Lambda_0 - \Lambda_0 V_N\Lambda_0\right)\Gamma\Psi - 2\Psi^\top\left(\Gamma^\top\Lambda_0 V_N B_N\right)\right)\right]\,,\\
&= \exp\left[-\frac{1}{2}\Psi^\top\Sigma_N^{-1}\Psi + \Psi^\top\Sigma_N^{-1}\mu_N\right]\,,
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_N^{-1} &= \Gamma^\top\left(\Lambda_0 - \Lambda_0 V_N\Lambda_0\right)\Gamma\,,\\
\Sigma_N^{-1}\mu_N &= \left(\Gamma^\top\Lambda_0 V_N B_N\right)\,.
\end{aligned} \tag{A.4}$$

Plugging the expression of V_N concludes the proof. □

A.2.2 Action Posterior Derivation

Proof of Proposition 2 (Derivation of $p_{t,a}$). This proposition is a direct application of Lemma 2; in which case we get that the posterior $p_{t,a}$ is a multivariate Gaussian distribution $\mathcal{N}(\tilde{\mu}_{t,a}, \tilde{\Sigma}_{t,a})$, where

$$\begin{aligned}
\tilde{\Sigma}_{t,a}^{-1} &= G_{t,a} + \Sigma_{0,a}^{-1}\,,\\
\tilde{\mu}_{t,a} &= \tilde{\Sigma}_{t,a}\left(B_{t,a} + \Sigma_{0,a}^{-1}\sum_{\ell=1}^{L} b_{a,\ell}\psi_{t,\ell}\right)\,.
\end{aligned}$$

□

To reduce clutter, we consider a fixed action $a \in [K]$ and round $t \in [T]$, and drop subindexing by t and a in Lemma 2. In summary, fix $a \in [K]$ and $t \in [T]$ such that we have the following correspondences:

$$b_\ell \leftarrow b_{a,\ell}\,, \quad \Sigma_0 \leftarrow \Sigma_{0,a}\,, \quad N \leftarrow N_{t,a}\,, \quad \theta \leftarrow \theta_a\,, \quad (X_i, R_i)_{i\in[N]} \leftarrow (X_i, R_i)_{i\in S_{t,a}}\,.$$

Lemma 2. *Consider the following model*

$$\begin{aligned} \theta \mid \Psi &\sim \mathcal{N}\left(\sum_{\ell=1}^{L} b_\ell \psi_\ell, \Sigma_0\right), \\ R_i \mid X_i, \theta &\sim \mathcal{N}\left(X_i^\top \theta, \sigma^2\right), \qquad \forall i \in [N]. \end{aligned}$$

Let $H = \{X_1, R_1, \ldots, X_N, R_N\}$ *then we have that* $p(\theta \mid \Psi, H) = \mathcal{N}\left(\theta; \tilde{\mu}_N, \tilde{\Sigma}_N\right)$, *where*

$$\begin{aligned} \tilde{\Sigma}_N^{-1} &= \sigma^{-2} \sum_{i=1}^{N} X_i X_i^\top + \Sigma_0^{-1}, \\ \tilde{\mu}_N &= \tilde{\Sigma}_N \left(\sigma^{-2} \sum_{i=1}^{N} X_i R_i + \Sigma_0^{-1} \sum_{\ell=1}^{L} b_\ell \psi_\ell\right). \end{aligned}$$

Proof. Let $v = \sigma^{-2}$, $\Lambda_0 = \Sigma_0^{-1}$. Then the action posterior decomposes as

$$\begin{aligned} p(\theta \mid \Psi, H) &\propto p((R_i)_{i\in[N]} \mid \Psi, \theta, (X_i)_{i\in[N]}) p(\theta \mid \Psi), \\ &= p((R_i)_{i\in[N]} \mid \theta, (X_i)_{i\in[N]}) p(\theta \mid \Psi), \\ &= \prod_{i=1}^{N} \mathcal{N}(R_i; X_i^\top \theta, \sigma^2) \mathcal{N}(\theta; \sum_{\ell=1}^{L} b_\ell \psi_\ell, \Sigma_0), \\ &= \exp\Big[-\frac{1}{2}\Big(v \sum_{i=1}^{N} (R_i^2 - 2R_i X_i^\top \theta + (X_i^\top \theta)^2) + \theta^\top \Lambda_0 \theta - 2\theta^\top \Lambda_0 \sum_{\ell=1}^{L} b_\ell \psi_\ell \\ &\qquad + \left(\sum_{\ell=1}^{L} b_\ell \psi_\ell\right)^\top \Lambda_0 \left(\sum_{\ell=1}^{L} b_\ell \psi_\ell\right)\Big)\Big], \\ &\propto \exp\left[-\frac{1}{2}\left(\theta^\top (v \sum_{i=1}^{N} X_i X_i^\top + \Lambda_0)\theta - 2\theta^\top \left(v \sum_{i=1}^{N} X_i R_i + \Lambda_0 \sum_{\ell=1}^{L} b_\ell \psi_\ell \right)\right)\right], \\ &\propto \mathcal{N}\left(\theta; \tilde{\mu}_N, \left(\tilde{\Lambda}_N\right)^{-1}\right), \end{aligned}$$

where $\tilde{\Lambda}_N = v \sum_{i=1}^{N} X_i X_i^\top + \Lambda_0$, and $\tilde{\Lambda}_N \tilde{\mu}_N = v \sum_{i=1}^{N} X_i R_i + \Lambda_0 \sum_{\ell=1}^{L} b_\ell \psi_\ell$. □

A.3 Regret Proofs

In this section, we establish a more general version of Theorem 1. As explained in Section 3.1.1, we analyze `meTS` in the linear setting under the assumption of a *fully well-specified* model. That is, the true action parameters and rewards are generated according to the same hierarchical structure assumed by `meTS`:

$$\begin{aligned} \Psi_* &\sim \mathcal{N}(\mu_\Psi, \Sigma_\Psi), \qquad (A.5) \\ \theta_{*,a} \mid \Psi_* &\sim \mathcal{N}\Big(\sum_{\ell=1}^{L} b_{a,\ell} \psi_{*,\ell}, \Sigma_{0,a}\Big), \qquad \forall a \in \mathcal{A}, \\ R_t \mid X_t, A_t, \theta_*, \Psi_* &\sim \mathcal{N}(X_t^\top \theta_{*,A_t}, \sigma^2), \qquad \forall t \in [T], \end{aligned}$$

where the subscript $*$ denotes the true action and latent parameters.

To derive the regret bound, we proceed as follows: First, we provide a compact problem formulation in Section A.3.1. Next, we employ total covariance decomposition to derive the posterior covariance of $\theta_{*,a} \mid H_t$ in Section A.3.2. Finally, we present preliminary eigenvalue results in Section A.3.3 before completing the proof in Section A.3.4.

A.3.1 Problem Reformulation for Regret Analysis

Here, we aim at rewriting Equation (A.5) in a compact form to simplify regret analysis. We first introduce K independent multivariate Gaussian variables $Z_a \sim \mathcal{N}(0, \Sigma_{0,a})$ for $a \in [K]$, and the following matrix

$$\Psi_{*,\text{mat}} = [\psi_{*,1}, \dots, \psi_{*,L}] \in \mathbb{R}^{d \times L} .$$

First, we have that $\text{Vec}(\Psi_{*,\text{mat}}) = \Psi_*$ where Ψ_* is defined in Equation (A.5). Moreover notice that $\sum_{\ell=1}^{L} b_{a,\ell}\psi_{*,\ell} = \Psi_{*,\text{mat}} b_a$, where $b_a = (b_{a,\ell})_{\ell \in [L]}$ and thus given matrix $\Psi_{*,\text{mat}}$ we have that

$$\theta_{*,a} = \Psi_{*,\text{mat}} b_a + Z_a , \qquad \forall a \in [K] . \tag{A.6}$$

We vectorize Equation (A.6) to obtain

$$\theta_{*,a} = \text{Vec}(\theta_{*,a}) = \text{Vec}(\Psi_{*,\text{mat}} b_a + Z_a) = \text{Vec}(\Psi_{*,\text{mat}} b_a) + Z_a , \tag{A.7}$$

where we used that if $X \in \mathbb{R}^d$ (a column vector), then $X = \text{Vec}(X)$ and that $\text{Vec}(\cdot)$ is a linear transformation. Also, we know from (c) in Section A.1 that $\text{Vec}(\text{AB}) = (\text{B}^\top \otimes I_n)\,\text{Vec}(\text{A})$ for any $\text{A} \in \mathbb{R}^{n \times m}, \text{B} \in \mathbb{R}^{m \times p}$. Therefore,

$$\theta_{*,a} = \Gamma_a \Psi_* + Z_a , \tag{A.8}$$

where $\Gamma_a = b_a^\top \otimes I_d$ and we used that $\text{Vec}(\Psi_{*,\text{mat}}) = \Psi_*$. It follows that

$$\theta_{*,a} \mid \Psi_* \sim \mathcal{N}(\Gamma_a \Psi_*, \Sigma_{0,a}) , \tag{A.9}$$

This allows us to rewrite our model as a single-parent hierarchical model

$$\begin{aligned} \Psi_* &\sim \mathcal{N}(\mu_\Psi, \Sigma_\Psi) , && \\ \theta_{*,a} \mid \Psi_* &\sim \mathcal{N}(\Gamma_a \Psi_*, \Sigma_{0,a}) , && \forall a \in [K] , \\ R_t \mid X_t, A_t, \theta_*, \Psi_* &\sim \mathcal{N}(X_t^\top \theta_{*,A_t}, \sigma^2) , && \forall t \in [T] . \end{aligned} \tag{A.10}$$

A.3.2 Derivation of $\text{cov}\left[\theta_{*,a} \,|\, H_t\right]$

Let

$$G_{t,a} = \sigma^{-2} \sum_{i \in S_{t,a}} X_i X_i^\top , \quad B_{t,a} = \sigma^{-2} \sum_{i \in S_{t,a}} R_i X_i .$$

Lemma 3 (Expression of $\operatorname{cov}[\theta_{*,a} \mid H_t]$)**.** *Consider the model in Equation (A.10), then we have*

$$\hat{\Sigma}_{t,a} = \operatorname{cov}[\theta_{*,a} \mid H_t] = \tilde{\Sigma}_{t,a} + \tilde{\Sigma}_{t,a}\Sigma_{0,a}^{-1}\Gamma_a\bar{\Sigma}_t\Gamma_a^\top\Sigma_{0,a}^{-1}\tilde{\Sigma}_{t,a}\,, \qquad \forall a \in [K]\,.$$

where

$$\begin{aligned}\bar{\Sigma}_t &= \left(\Sigma_\Psi^{-1} + \sum_{a=1}^{K} b_a b_a^\top \otimes \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right)\right)^{-1} \\ \tilde{\Sigma}_{t,a} &= \left(G_{t,a} + \Sigma_{0,a}^{-1}\right)^{-1}\,.\end{aligned}$$

Proof. Before proceeding with the proof, we emphasize that

$$\operatorname{cov}[\Psi_* \mid H_t] = \operatorname{cov}[\Psi \mid H_t] = \bar{\Sigma}_t\,, \qquad \mathbb{E}[\Psi_* \mid H_t] = \mathbb{E}[\Psi \mid H_t] = \bar{\mu}_t\,,$$

and

$$\operatorname{cov}[\theta_{*,a} \mid \Psi_*, H_t] = \operatorname{cov}[\theta_a \mid \Psi, H_t] = \tilde{\Sigma}_{t,a}\,, \qquad \mathbb{E}[\theta_{*,a} \mid \Psi_*, H_t] = \mathbb{E}[\theta_a \mid \Psi, H_t] = \tilde{\mu}_{t,a}\,,$$

where the explicit expressions of these covariances and expectations are provided in Proposition 1 and Proposition 2, respectively. These equalities hold because the true action parameters and rewards are assumed to follow the exact generative process defined by the `meTS` model.

Now let $\Lambda_{0,a} = \Sigma_{0,a}^{-1}$. Proposition 2 and the fact that $\sum_{\ell\in[L]} b_{a,\ell}\psi_{*,\ell} = \Gamma_a\Psi_*$ where $\Gamma_a = b_a^\top \otimes I_d$ (Section A.3.1) yield

$$\begin{aligned}\operatorname{cov}[\theta_{*,a} \mid \Psi_*, H_t] &= (G_{t,a} + \Lambda_{0,a})^{-1} \\ \mathbb{E}[\theta_{*,a} \mid \Psi_*, H_t] &= \operatorname{cov}[\theta_{*,a} \mid \Psi_*, H_t]\,(B_{t,a} + \Lambda_{0,a}\Gamma_a\Psi_*)\end{aligned}$$

First, given H_t, $\operatorname{cov}[\theta_{*,a} \mid \Psi_*, H_t] = (G_{t,a} + \Lambda_{0,a})^{-1}$ is constant (does not depend on Ψ_*). Thus

$$\mathbb{E}[\operatorname{cov}[\theta_{*,a} \mid \Psi_*, H_t] \mid H_t] = \operatorname{cov}[\theta_{*,a} \mid \Psi_*, H_t] = (G_{t,a} + \Lambda_{0,a})^{-1}\,.$$

In addition, given H_t, both $(G_{t,a} + \Lambda_{0,a})^{-1}$ and $B_{t,a}$ are constant. Thus

$$\begin{aligned}\operatorname{cov}[\mathbb{E}[\theta_{*,a} \mid \Psi_*, H_t] \mid H_t] &= \operatorname{cov}[\operatorname{cov}[\theta_{*,a} \mid \Psi_*, H_t]\,\Lambda_{0,a}\Gamma_a\Psi_* \mid H_t] \\ &= (G_{t,a} + \Lambda_{0,a})^{-1}\Lambda_{0,a}\Gamma_a\operatorname{cov}[\Psi_* \mid H_t]\,\Gamma_a^\top\Lambda_{0,a}\,(G_{t,a} + \Lambda_{0,a})^{-1} \\ &= (G_{t,a} + \Lambda_{0,a})^{-1}\Lambda_{0,a}\Gamma_a\bar{\Sigma}_t\Gamma_a^\top\Lambda_{0,a}\,(G_{t,a} + \Lambda_{0,a})^{-1}\,.\end{aligned}$$

Finally, total covariance decomposition (Weiss, 2005) concludes the proof. □

A.3.3 Preliminary Eigenvalues Results

Next we present some preliminary upper bounds on the maximum eigenvalues of our covariance matrices.

- **Definitions:** Let $\lambda_{1,0} = \max_{a\in[K]} \lambda_1(\Sigma_{0,a})$, $\lambda_{d,0} = \min_{a\in[K]} \lambda_d(\Sigma_{0,a})$, $\lambda_{1,\Psi} = \lambda_1(\Sigma_\Psi)$, and $\kappa_b = \max_{a\in[K]} \|b_a\|_2^2$.
- **upper bound of $\lambda_1(\Gamma_a\Gamma_a^\top)$:**

$$\lambda_1(\Gamma_a\Gamma_a^\top) \leq \kappa_b\,, \qquad \forall a \in [K]\,. \tag{A.11}$$

  Similarly, we have that

$$\lambda_1(\Gamma_a^\top\Gamma_a) \leq \kappa_b\,, \qquad \forall a \in [K]\,. \tag{A.12}$$

- **upper bound of $\lambda_1(\hat{\Sigma}_{t,a})$:**

$$\lambda_1(\hat{\Sigma}_{t,a}) \leq \lambda_{1,0} + \frac{\lambda_{1,0}^2\lambda_{1,\Psi}\kappa_b}{\lambda_{d,0}^2}\,, \qquad \forall a \in [K]\,. \tag{A.13}$$

- **upper bound of $\lambda_1(\Sigma_\Psi^{\frac{1}{2}}\bar{\Sigma}_{T+1}^{-1}\Sigma_\Psi^{\frac{1}{2}})$:**

$$\lambda_1(\Sigma_\Psi^{\frac{1}{2}}\bar{\Sigma}_{T+1}^{-1}\Sigma_\Psi^{\frac{1}{2}}) \leq 1 + K\lambda_{1,\Psi}\kappa_b\Big(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda_{1,0}^2\big(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\lambda_{d,0}}\big)}\Big)\,. \tag{A.14}$$

Proof. We start with Equation (A.11). First, recall that $\Gamma_a = b_a^\top \otimes I_d$ for any $a \in [K]$. Thus $\Gamma_a\Gamma_a^\top = (b_a^\top \otimes I_d)(b_a \otimes I_d) = \|b_a\|_2^2 I_d$ for any $a \in [K]$. Then $\lambda_1(\Gamma_a\Gamma_a^\top) = \|b_a\|_2^2 \leq \kappa_b$. The second result follows from the fact that $\lambda_1(\Gamma_a\Gamma_a^\top) = \lambda_1(\Gamma_a^\top\Gamma_a)$.

Now we prove the result in Equation (A.13). This follows from the expression of $\hat{\Sigma}_{t,a}$ in Lemma 3. Precisely, we have that

$$\hat{\Sigma}_{t,a} = \tilde{\Sigma}_{t,a} + \tilde{\Sigma}_{t,a}\Sigma_{0,a}^{-1}\Gamma_a\bar{\Sigma}_t\Gamma_a^\top\Sigma_{0,a}^{-1}\tilde{\Sigma}_{t,a}\,, \qquad \forall a \in [K]\,.$$

where $\tilde{\Sigma}_{t,a} = \big(G_{t,a} + \Sigma_{0,a}^{-1}\big)^{-1}$. Thus Weyl's inequality combined with the properties in Section A.1 yields that

$$\lambda_1(\hat{\Sigma}_{t,a}) \leq \lambda_1(\tilde{\Sigma}_{t,a}) + \lambda_1(\tilde{\Sigma}_{t,a})\lambda_1(\Sigma_{0,a}^{-1})\lambda_1(\Gamma_a\bar{\Sigma}_t\Gamma_a^\top)\lambda_1(\Sigma_{0,a}^{-1})\lambda_1(\tilde{\Sigma}_{t,a}) \leq \lambda_{1,0} + \frac{\lambda_{1,0}^2\lambda_{1,\Psi}\kappa_b}{\lambda_{d,0}^2}$$

In the last inequality, we used that $\lambda_1(\Gamma_a\bar{\Sigma}_t\Gamma_a^\top) \leq \lambda_1(\bar{\Sigma}_t)\lambda_1(\Gamma_a\Gamma_a^\top)$, ((f) in Section A.1), $\lambda_1(\Sigma_{0,a}^{-1}) \leq \frac{1}{\lambda_{d,0}}$, and $\lambda_1(\tilde{\Sigma}_{t,a}) \leq \lambda_{1,0}$.

Finally, we prove the result in Equation (A.14). First, we rewrite the precision matrix of the effect posterior $\bar{\Sigma}_t^{-1}$ using the compact notation introduced in Section A.3.1. Precisely, it follows from Equation (A.4) that

$$\begin{aligned}\bar{\Sigma}_t^{-1} &\overset{(i)}{=} \Sigma_\Psi^{-1} + \sum_{a=1}^K \Gamma_a^\top\left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right)\Gamma_a\,,\\ &\overset{(ii)}{=} \Sigma_\Psi^{-1} + \sum_{a=1}^K \Gamma_a^\top\left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}\tilde{\Sigma}_{t,a}\Sigma_{0,a}^{-1}\right)\Gamma_a\,.\end{aligned}$$

Here, (i) and (ii) are the same; (ii) follows from plugging $\tilde{\Sigma}_{t,a} = (G_{t,a} + \Sigma_{0,a}^{-1})^{-1}$ in (i). Then we have that

$$
\begin{aligned}
&\lambda_1(\Sigma_\Psi^{\frac{1}{2}} \bar{\Sigma}_{T+1}^{-1} \Sigma_\Psi^{\frac{1}{2}}) \\
&= \lambda_1\Big(I_{Ld} + \sum_{a=1}^{K} \Sigma_\Psi^{\frac{1}{2}} \Gamma_a^\top \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}\tilde{\Sigma}_{T+1,a}\Sigma_{0,a}^{-1}\right) \Gamma_a \Sigma_\Psi^{\frac{1}{2}}\Big) \\
&\leq 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \lambda_1\Big(\Gamma_a^\top \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}\tilde{\Sigma}_{T+1,a}\Sigma_{0,a}^{-1}\right) \Gamma_a\Big), \\
&\leq 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \lambda_1(\Gamma_a^\top \Gamma_a) \lambda_1\left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}\tilde{\Sigma}_{T+1,a}\Sigma_{0,a}^{-1}\right) \\
&\leq 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \kappa_b \left(\lambda_1\left(\Sigma_{0,a}^{-1}\right) + \lambda_1\left(-\Sigma_{0,a}^{-1}\tilde{\Sigma}_{T+1,a}\Sigma_{0,a}^{-1}\right)\right), \\
&\leq 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \kappa_b \left(\frac{1}{\lambda_{d,0}} - \lambda_d\left(\Sigma_{0,a}^{-1}\tilde{\Sigma}_{T+1,a}\Sigma_{0,a}^{-1}\right)\right) \\
&\leq 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \kappa_b \left(\frac{1}{\lambda_{d,0}} - \lambda_d\left(\Sigma_{0,a}^{-1}\right) \lambda_d\left(\tilde{\Sigma}_{T+1,a}\right) \lambda_d\left(\Sigma_{0,a}^{-1}\right)\right), \\
&\leq 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \kappa_b \left(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda_{1,0}^2}\lambda_d\left(\tilde{\Sigma}_{T+1,a}\right)\right) \\
&= 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \kappa_b \left(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda_{1,0}^2 \lambda_1\left(G_{T+1,a} + \Sigma_{0,a}^{-1}\right)}\right), \\
&\leq 1 + \lambda_{1,\Psi} \sum_{a=1}^{K} \kappa_b \left(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda_{1,0}^2 \left(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\lambda_{d,0}}\right)}\right) \\
&= 1 + K\lambda_{1,\Psi}\kappa_b \left(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda_{1,0}^2 \left(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\lambda_{d,0}}\right)}\right).
\end{aligned}
$$

□

A.3.4 Regret Proof

Here we prove a more general version of Theorem 1 where we do not assume that the covariance matrices $\Sigma_{0,a}$ and Σ_Ψ are diagonal. We still assume that there exists $\kappa_x > 0$ such that $\|X_t\|_2^2 \leq \kappa_x$ for any $t \in [T]$.

Theorem 6 (General version of Theorem 1)**.** *For any $\delta \in (0,1)$, the Bayes regret of* `meTS` *in the mixed-effect model in Section 3.1.1 is bounded as*

$$
\mathcal{BR}(T) \leq \sqrt{2T\left(\mathcal{R}^{\text{A}}(T) + \mathcal{R}^{\text{E}}(T)\right)\log(1/\delta)} + cT\delta\,, \tag{A.15}
$$

with $c = \sqrt{\frac{2}{\pi}\kappa_x\big(\lambda_{1,0} + \frac{\lambda_{1,0}^2\lambda_{1,\Psi}\kappa_b}{\lambda_{d,0}^2}\big)K}$, $\kappa_b = \max_{a\in[K]} \|b_a\|_2^2$, $\lambda_{1,0} = \max_{a\in[K]} \lambda_1(\Sigma_{0,a})$, $\lambda_{d,0} = \min_{a\in[K]} \lambda_d(\Sigma_{0,a})$, $\lambda_{1,\Psi} = \lambda_1(\Sigma_\Psi)$ *and*

$$\mathcal{R}^{\text{A}}(T) = dKc_{\text{A}} \log\big(1 + \frac{T\kappa_x\lambda_{1,0}}{\sigma^2 d}\big)\,,\ c_{\text{A}} = \frac{\kappa_x\lambda_{1,0}}{\log(1+\frac{\kappa_x\lambda_{1,0}}{\sigma^2})}\,,$$

$$\mathcal{R}^{\text{E}}(T) = dLc_{\text{E}} \log\Big(1 + K\kappa_b\lambda_{1,\Psi}\Big(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda_{1,0}^2\big(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\lambda_{d,0}}\big)}\Big)\Big)\,,\ c_{\text{E}} = \frac{\kappa_x\kappa_b\lambda_{1,0}^2\lambda_{1,\Psi}\big(1+\frac{\kappa_x\lambda_{1,0}}{\sigma^2}\big)}{\lambda_{d,0}^2\log\big(1+\frac{\kappa_x\kappa_b\lambda_{1,0}^2\lambda_{1,\Psi}}{\sigma^2\lambda_{d,0}^2}\big)}\,.$$

In particular, the result in Theorem 1 is retrieved when $\lambda_{1,0} = \lambda_{d,0} = \sigma_0^2$, *and* $\lambda_{1,\Psi} = \sigma_\Psi^2$.

Proof. Consider our model rewritten in Equation (A.10). Then, the posterior distribution of the action parameter $\theta_{*,a} \mid H_t$ is a multivariate Gaussian distribution $\mathcal{N}(\hat{\mu}_{t,a}, \hat{\Sigma}_{t,a})$ for some $\hat{\mu}_{t,a} \in \mathbb{R}^d$ and $\hat{\Sigma}_{t,a} \in \mathbb{R}^{d\times d}$ (Lemma 3). Now we let $\theta_{t,*} = (X_t^\top\theta_{*,a})_{a\in[K]} \in \mathbb{R}^K$ be the concatenation of the expected rewards of actions in round t. Notice that the context X_t is known in round t, and thus we include it in the history H_t. This is important, with slight abuse of notation, H_t now denotes $H_t \leftarrow H_t \cup \{X_t\}$. Then, the joint posterior of the expected rewards, $\theta_{t,*} \mid H_t$, is also a multivariate Gaussian $\mathcal{N}(\check{\theta}_t, \check{\Sigma}_t)$ for $\check{\theta}_t = (X_t^\top\hat{\mu}_{t,a})_{a\in[K]} \in \mathbb{R}^K$ and some covariance $\check{\Sigma}_t \in \mathbb{R}^{K\times K}$. This follows from the properties of Gaussian distributions (Koller and Friedman, 2009) and the fact that X_t is now included in H_t. Let $\boldsymbol{A}_t \in \{0,1\}^K$ and $\boldsymbol{A}_{t,*} \in \{0,1\}^K$ be indicator vectors of the taken action A_t and optimal action $A_{t,*}$, respectively (The vector representations are in bold letters while the integer representations are in regular letters). Then the Bayes regret can be rewritten and consequently decomposed following standard analysis (Russo and Van Roy, 2014) as

$$\begin{aligned}
\mathcal{BR}(T) &= \mathbb{E}\left[\sum_{t=1}^T X_t^\top\theta_{*,A_{t,*}} - X_t^\top\theta_{*,A_t}\right]\,, && \text{(A.16)}\\
&= \mathbb{E}\Big[\sum_{t=1}^T \boldsymbol{A}_{t,*}^\top\theta_{t,*} - \boldsymbol{A}_t^\top\theta_{t,*}\Big]\,, && \text{(A.17)}\\
&= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{A}_{t,*}^\top(\theta_{t,*} - \check{\theta}_t) \,\middle|\, H_t\right]\right] + \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{A}_t^\top(\check{\theta}_t - \theta_{t,*}) \,\middle|\, H_t\right]\right]\,.
\end{aligned}$$

This follows from the fact that $\check{\theta}_t = (X_t^\top\hat{\mu}_{t,i})_{i\in[K]}$ is deterministic given H_t (since H_t now includes X_t), and that $\boldsymbol{A}_{t,*}$ and $\boldsymbol{A}_t$ are i.i.d. given H_t. Moreover, given H_t, $\check{\theta}_t - \theta_{t,*}$ is a zero-mean multivariate random variable independent of $\boldsymbol{A}_t$ and thus $\mathbb{E}[\boldsymbol{A}_t^\top(\check{\theta}_t - \theta_{t,*}) \mid H_t] = 0$. Therefore, we only need to bound the first term in (A.16). With slight abuse of notation, let $\boldsymbol{\mathcal{A}}$ be the set of all possible indicator vectors of actions $a \in [K]$. Precisely, an action $a \in [K]$ is also represented by an indicator vector $\boldsymbol{a} \in \boldsymbol{\mathcal{A}} \subset \{0,1\}^K$ (in bold letter). Then we define the following events

$$E_{t,\boldsymbol{a}}(\delta) = \Big\{|\boldsymbol{a}^\top(\theta_{t,*} - \check{\theta}_t)| \le \sqrt{2\log(1/\delta)}\|\boldsymbol{a}\|_{\check{\Sigma}_t}\Big\}\,, \qquad \forall\delta\in(0,1)\,,\ \forall\boldsymbol{a}\in\boldsymbol{\mathcal{A}}\,.$$

Fix history H_t, we split the expectation over the two complementary events $E_{t,\boldsymbol{A}_{t,*}}(\delta)$ and

$\bar{E}_{t,\boldsymbol{A}_{t,*}}(\delta)$, and use the Cauchy-Schwarz inequality to obtain

$$\mathbb{E}\left[\boldsymbol{A}_{t,*}^\top(\theta_{t,*}-\check{\theta}_t)\,\middle|\,H_t\right] \le \sqrt{2\log(1/\delta)}\,\mathbb{E}\left[\|\boldsymbol{A}_{t,*}\|_{\check{\Sigma}_t}\,\middle|\,H_t\right] + \mathbb{E}\left[\boldsymbol{A}_{t,*}^\top(\theta_{t,*}-\check{\theta}_t)\mathbb{1}\left\{\bar{E}_{t,\boldsymbol{A}_{t,*}}(\delta)\right\}\,\middle|\,H_t\right]. \tag{A.18}$$

Now the second term in Equation (A.18) can be bounded as follows. For any $\boldsymbol{a}\in\boldsymbol{\mathcal{A}}$, let $Z_{\boldsymbol{a}} = \boldsymbol{a}^\top(\theta_{t,*}-\check{\theta}_t)$. Then we have that

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{A}_{t,*}^\top(\theta_{t,*}-\check{\theta}_t)\mathbb{1}\left\{\bar{E}_{t,\boldsymbol{A}_{t,*}}(\delta)\right\}\,\middle|\,H_t\right]\\
&\quad\overset{(i)}{=}\mathbb{E}\left[Z_{\boldsymbol{A}_{t,*}}\mathbb{1}\left\{|Z_{\boldsymbol{A}_{t,*}}|>\sqrt{2\log(1/\delta)}\|\boldsymbol{A}_{t,*}\|_{\check{\Sigma}_t}\right\}\,\middle|\,H_t\right],\\
&\quad\overset{(ii)}{\le}\mathbb{E}\left[|Z_{\boldsymbol{A}_{t,*}}|\mathbb{1}\left\{|Z_{\boldsymbol{A}_{t,*}}|>\sqrt{2\log(1/\delta)}\|\boldsymbol{A}_{t,*}\|_{\check{\Sigma}_t}\right\}\,\middle|\,H_t\right],\\
&\quad\overset{(iii)}{\le}\sum_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\mathbb{E}\left[|Z_{\boldsymbol{a}}|\mathbb{1}\left\{|Z_{\boldsymbol{a}}|>\sqrt{2\log(1/\delta)}\|\boldsymbol{a}\|_{\check{\Sigma}_t}\right\}\,\middle|\,H_t\right],\\
&\quad\overset{(iv)}{\le}\sum_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\frac{2}{\|\boldsymbol{a}\|_{\check{\Sigma}_t}\sqrt{2\pi}}\int_{u=\sqrt{2\log(1/\delta)}\|\boldsymbol{a}\|_{\check{\Sigma}_t}}^{\infty}u\exp\left[-\frac{u^2}{2\|\boldsymbol{a}\|_{\check{\Sigma}_t}^2}\right]\mathrm{d}u,\\
&\quad\overset{(v)}{\le}\sum_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\|\boldsymbol{a}\|_{\check{\Sigma}_t}\frac{2}{\sqrt{2\pi}}\int_{u=\sqrt{2\log(1/\delta)}}^{\infty}u\exp\left[-\frac{u^2}{2}\right]\mathrm{d}u\overset{(vi)}{\le}\sqrt{\frac{2}{\pi}}\lambda_{\max,t}K\delta\,.
\end{aligned}\tag{A.19}$$

In (i), we simply rewrite the terms using the random variable $Z_{\boldsymbol{A}_{t,*}}$. In (ii), we use the fact that $Z_{\boldsymbol{A}_{t,*}}\le|Z_{\boldsymbol{A}_{t,*}}|$. In (iii), we upper bound the expectation of the random variable $|Z_{\boldsymbol{A}_{t,*}}|\mathbb{1}\left\{|Z_{\boldsymbol{A}_{t,*}}|>\sqrt{2\log(1/\delta)}\|\boldsymbol{A}_{t,*}\|_{\check{\Sigma}_t}\right\}$ with the sum of the expectations of $|Z_{\boldsymbol{a}}|\mathbb{1}\left\{|Z_{\boldsymbol{a}}|>\sqrt{2\log(1/\delta)}\|\boldsymbol{a}\|_{\check{\Sigma}_t}\right\}$ for $\boldsymbol{a}\in\boldsymbol{\mathcal{A}}$ since all these random variables are non-negative. Moreover, (iv) follows from the facts that given H_t, $Z_{\boldsymbol{a}}\sim\mathcal{N}(0,\|\boldsymbol{a}\|_{\check{\Sigma}_t}^2)$, and that if $Z\sim\mathcal{N}(0,\sigma^2)$, then for any $\epsilon\ge0$, $\mathbb{P}(|Z|>\epsilon)\le2\mathbb{P}(Z>\epsilon)$. In (v), we use the change of variables $u\leftarrow u/\|\boldsymbol{a}\|_{\check{\Sigma}_t}$. Finally, in (vi), we compute the integral and set $\lambda_{\max,t}=\max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}}\|\boldsymbol{a}\|_{\check{\Sigma}_t}$. We combine Equation (A.18) and Equation (A.19) with the fact that $\boldsymbol{A}_t$ and $\boldsymbol{A}_{t,*}$ are i.i.d. given H_t to obtain that

$$\mathbb{E}\left[\boldsymbol{A}_{t,*}^\top(\theta_{t,*}-\check{\theta}_t)\,\middle|\,H_t\right]\le\sqrt{2\log(1/\delta)}\,\mathbb{E}\left[\|\boldsymbol{A}_t\|_{\check{\Sigma}_t}\,\middle|\,H_t\right]+\sqrt{\frac{2}{\pi}}\lambda_{\max,t}K\delta\,. \tag{A.20}$$

The bound in Equation (A.20) holds for any history H_t and thus we take an additional expectation and get that

$$\begin{aligned}
\mathcal{BR}(T)=\mathbb{E}\left[\sum_{t=1}^T\boldsymbol{A}_{t,*}^\top\theta_{t,*}-\boldsymbol{A}_t^\top\theta_{t,*}\right]&\le\sqrt{2\log(1/\delta)}\,\mathbb{E}\left[\sum_{t=1}^T\|\boldsymbol{A}_t\|_{\check{\Sigma}_t}\right]+\sqrt{\frac{2}{\pi}}K\delta\sum_{t=1}^T\lambda_{\max,t}\,,\\
&\overset{(i)}{\le}\sqrt{2T\log(1/\delta)}\,\mathbb{E}\left[\sqrt{\sum_{t=1}^T\|\boldsymbol{A}_t\|_{\check{\Sigma}_t}^2}\right]+\sqrt{\frac{2}{\pi}}K\delta\sum_{t=1}^T\lambda_{\max,t}\,,\\
&\overset{(ii)}{\le}\sqrt{2T\log(1/\delta)}\sqrt{\mathbb{E}\left[\sum_{t=1}^T\|\boldsymbol{A}_t\|_{\check{\Sigma}_t}^2\right]}+\sqrt{\frac{2}{\pi}}K\delta\sum_{t=1}^T\lambda_{\max,t}\,,
\end{aligned}$$

where we use the Cauchy-Schwarz inequality in (i), and (ii) follows from the concavity of the square root. Now note that any $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$ is an indicator vector and that $\check{\Sigma}_t$ is the covariance of the joint posterior of the expected rewards $(X_t^\top \theta_{*,a})_{a\in[K]} \mid H_t$. Therefore, for any $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$, $\|\boldsymbol{a}\|^2_{\check{\Sigma}_t} = \check{\sigma}_a^2$ is the variance of $X_t^\top \theta_{*,a} \mid H_t$. But we know that $\theta_{*,a} \mid H_t$ is a multivariate Gaussian and its covariance is $\hat{\Sigma}_{t,a}$ (Lemma 3). Thus the variance of $X^\top \theta_{*,a} \mid H_t$ is $\check{\sigma}_a^2 = X_t^\top \hat{\Sigma}_{t,a} X_t$. It follows that for any $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$, $\|\boldsymbol{a}\|^2_{\check{\Sigma}_t} = X_t^\top \hat{\Sigma}_{t,a} X_t = \|X_t\|^2_{\hat{\Sigma}_{t,a}}$. In particular, $\|\boldsymbol{A}_t\|^2_{\check{\Sigma}_t} = X_t^\top \hat{\Sigma}_{t,A_t} X_t$. Combining this with Equation (A.13) yields that $\lambda_{\max,t} = \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \|\boldsymbol{a}\|_{\check{\Sigma}_t} = \max_{a\in\mathcal{A}} \|X_t\|_{\hat{\Sigma}_{t,a}} \le \max_{a\in\mathcal{A}} \sqrt{\lambda_1(\hat{\Sigma}_{t,a})\kappa_x} \le \sqrt{\left(\lambda_{1,0} + \frac{\lambda_{1,0}^2 \lambda_{1,\Psi}\kappa_b}{\lambda_{d,0}^2}\right)\kappa_x}$. Then we let $c = \sqrt{\frac{2}{\pi}\left(\lambda_{1,0} + \frac{\lambda_{1,0}^2 \lambda_{1,\Psi}\kappa_b}{\lambda_{d,0}^2}\right)\kappa_x} K$ which allows us to write

$$\mathcal{BR}(T) \le \sqrt{2T\log(1/\delta)} \sqrt{\mathbb{E}\left[\sum_{t=1}^T \|X_t\|^2_{\hat{\Sigma}_{t,A_t}}\right]} + cT\delta\,. \tag{A.21}$$

Now we focus on the the term $\sqrt{\mathbb{E}\left[\sum_{t=1}^T \|X_t\|^2_{\hat{\Sigma}_{t,A_t}}\right]}$ that we decompose and bound as

$$\begin{aligned}
\|X_t\|^2_{\hat{\Sigma}_{t,A_t}} &= \sigma^2 \frac{X_t^\top \hat{\Sigma}_{t,A_t} X_t}{\sigma^2} \overset{(i)}{=} \sigma^2 \left(\sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} X_t + \sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t\right)\,, \\
&\overset{(ii)}{\le} c_{\mathrm{A}} \log(1 + \sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} X_t) + c_1 \log(1 + \sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t)\,,
\end{aligned} \tag{A.22}$$

where (i) follows from $\hat{\Sigma}_{t,A_t} = \tilde{\Sigma}_{t,A_t} + \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t}$, and we use the following inequality in (ii)

$$x = \frac{x}{\log(1+x)} \log(1+x) \le \left(\max_{x\in[0,u]} \frac{x}{\log(1+x)}\right) \log(1+x) = \frac{u}{\log(1+u)} \log(1+x)\,,$$

which holds for any $x \in [0, u]$, where constants c_{A} and c_1 are derived as

$$c_{\mathrm{A}} = \frac{\kappa_x \lambda_{1,0}}{\log(1 + \sigma^{-2}\kappa_x \lambda_{1,0})}\,, \quad c_1 = \frac{c_\Psi}{\log(1+\sigma^{-2} c_\Psi)}\,, \quad c_\Psi = \frac{\kappa_x \kappa_b \lambda_{1,0}^2 \lambda_{1,\Psi}}{\lambda_{d,0}^2}\,,$$

The derivation of c_{A} uses that

$$X_t^\top \tilde{\Sigma}_{t,A_t} X_t \le \lambda_1(\tilde{\Sigma}_{t,A_t}) \|X_t\|^2 \le \lambda_d^{-1}(\Sigma_{0,A_t}^{-1} + G_{t,A_t})\kappa_x \le \lambda_d^{-1}(\Sigma_{0,A_t}^{-1})\kappa_x = \lambda_1(\Sigma_{0,A_t})\kappa_x \le \lambda_{1,0}\kappa_x\,.$$

The derivation of c_1 follows from

$$\begin{aligned}
X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t &\le \lambda_1^2(\tilde{\Sigma}_{t,A_t}) \lambda_1^2(\Sigma_{0,A_t}^{-1}) \lambda_1(\Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top) \kappa_x\,, \\
&\le \frac{\lambda_1^2(\Sigma_{0,A_t}) \lambda_{1,\Psi} \lambda_1(\Gamma_{A_t} \Gamma_{A_t}^\top) \kappa_x}{\lambda_d^2(\Sigma_{0,A_t})}\,, \\
&\le \frac{\lambda_{1,0}^2 \lambda_{1,\Psi} \kappa_b \kappa_x}{\lambda_{d,0}^2}\,.
\end{aligned}$$

The first inequality follows from Weyl's inequality and the fact that $\lambda_1(\bar{\Sigma}_t) \leq \lambda_1(\Sigma_\Psi) = \lambda_{1,\Psi}$ and $\lambda_1(\tilde{\Sigma}_{t,A_t}) \leq \lambda_1(\Sigma_{0,A_t})$. Now we focus on bounding the logarithmic terms in Equation (A.22).

First Term in Equation (A.22) We first rewrite this term as

$$\begin{aligned}\log(1+\sigma^{-2}X_t^\top \tilde{\Sigma}_{t,A_t} X_t) &\overset{(i)}{=} \log\det(I_d + \sigma^{-2}\tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}})\,,\\ &= \log\det(\tilde{\Sigma}_{t,A_t}^{-1} + \sigma^{-2}X_t X_t^\top) - \log\det(\tilde{\Sigma}_{t,A_t}^{-1})\,,\\ &= \log\det(\tilde{\Sigma}_{t+1,A_t}^{-1}) - \log\det(\tilde{\Sigma}_{t,A_t}^{-1})\,,\end{aligned}$$

where (i) follows from the Weinstein–Aronszajn identity. Now note that for any $a \neq A_t$, the arm-a precision does not update at round t, hence $\tilde{\Sigma}_{t+1,a}^{-1} = \tilde{\Sigma}_{t,a}^{-1}$ and the increment is zero; therefore we may sum over all $a \in [K]$ without changing the value. Then, we sum over all rounds $t \in [T]$, and get a telescoping that leads to

$$\begin{aligned}&\sum_{t=1}^{T} \log\det(I_d + \sigma^{-2}\tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}}) = \sum_{t=1}^{T} \log\det(\tilde{\Sigma}_{t+1,A_t}^{-1}) - \log\det(\tilde{\Sigma}_{t,A_t}^{-1})\,,\\ &= \sum_{t=1}^{T}\sum_{a=1}^{K} \log\det(\tilde{\Sigma}_{t+1,a}^{-1}) - \log\det(\tilde{\Sigma}_{t,a}^{-1}) = \sum_{a=1}^{K}\sum_{t=1}^{T} \log\det(\tilde{\Sigma}_{t+1,a}^{-1}) - \log\det(\tilde{\Sigma}_{t,a}^{-1})\,,\\ &= \sum_{a=1}^{K} \log\det(\tilde{\Sigma}_{T+1,a}^{-1}) - \log\det(\tilde{\Sigma}_{1,a}^{-1})\,,\\ &\overset{(i)}{=} \sum_{a=1}^{K} \log\det(\Sigma_{0,a}^{\frac{1}{2}}\tilde{\Sigma}_{T+1,a}^{-1}\Sigma_{0,a}^{\frac{1}{2}}) \overset{(ii)}{\leq} \sum_{a=1}^{K} d\log\left(\frac{1}{d}\operatorname{Tr}(\Sigma_{0,a}^{\frac{1}{2}}\tilde{\Sigma}_{T+1,a}^{-1}\Sigma_{0,a}^{\frac{1}{2}})\right)\\ &\leq \sum_{a=1}^{K} d\log\left(1 + \frac{\kappa_x \lambda_1(\Sigma_{0,a})T}{\sigma^2 d}\right) \leq Kd\log\left(1 + \frac{\kappa_x\lambda_{1,0}T}{\sigma^2 d}\right)\,.\end{aligned}$$

where (i) follows from the fact that $\tilde{\Sigma}_{1,a} = \Sigma_{0,a}$ and we use the inequality of arithmetic and geometric means in (ii).

Second Term in Equation (A.22) First, we rewrite the covariance matrix of the effect posterior $\bar{\Sigma}_t$ using the compact notation introduced in Section A.3.1. Precisely, it follows from Equation (A.4) that

$$\begin{aligned}\bar{\Sigma}_t^{-1} &\overset{(i)}{=} \Sigma_\Psi^{-1} + \sum_{a=1}^{K} \Gamma_a^\top \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}(G_{t,a} + \Sigma_{0,a}^{-1})^{-1}\Sigma_{0,a}^{-1}\right)\Gamma_a\,,\\ &\overset{(ii)}{=} \Sigma_\Psi^{-1} + \sum_{a=1}^{K} \Gamma_a^\top \left(\Sigma_{0,a}^{-1} - \Sigma_{0,a}^{-1}\tilde{\Sigma}_{t,a}\Sigma_{0,a}^{-1}\right)\Gamma_a\,.\end{aligned} \tag{A.23}$$

Recall that (i) and (ii) are the same; (ii) follows from plugging $\tilde{\Sigma}_{t,a} = (G_{t,a} + \Sigma_{0,a}^{-1})^{-1}$ in

(i). Now let $u = \sigma^{-1}\tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t$. Then it follows from (ii) in Equation (A.23) that

$$\begin{aligned}
\bar{\Sigma}_{t+1}^{-1} - \bar{\Sigma}_t^{-1} &= \Gamma_{A_t}^\top \left(\Sigma_{0,A_t}^{-1} - \Sigma_{0,A_t}^{-1} (\tilde{\Sigma}_{t,A_t}^{-1} + \sigma^{-2} X_t X_t^\top)^{-1} \Sigma_{0,A_t}^{-1} - (\Sigma_{0,A_t}^{-1} - \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1}) \right) \Gamma_{A_t} \,, \\
&= \Gamma_{A_t}^\top \left(\Sigma_{0,A_t}^{-1} (\tilde{\Sigma}_{t,A_t} - (\tilde{\Sigma}_{t,A_t}^{-1} + \sigma^{-2} X_t X_t^\top)^{-1}) \Sigma_{0,A_t}^{-1} \right) \Gamma_{A_t} \,, \\
&= \Gamma_{A_t}^\top \left(\Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} (I_d - (I_d + \sigma^{-2} \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}})^{-1}) \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} \Sigma_{0,A_t}^{-1} \right) \Gamma_{A_t} \,, \\
&= \Gamma_{A_t}^\top \left(\Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} (I_d - (I_d + uu^\top)^{-1}) \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} \Sigma_{0,A_t}^{-1} \right) \Gamma_{A_t} \,, \\
&\overset{(i)}{=} \Gamma_{A_t}^\top \left(\Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} \frac{uu^\top}{1 + u^\top u} \tilde{\Sigma}_{t,A_t}^{\frac{1}{2}} \Sigma_{0,A_t}^{-1} \right) \Gamma_{A_t} \,, \\
&= \sigma^{-2} \Gamma_{A_t}^\top \left(\Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} \frac{X_t X_t^\top}{1 + u^\top u} \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \right) \Gamma_{A_t} \,. \qquad \text{(A.24)}
\end{aligned}$$

In (i) we use the Sherman-Morrison formula. Moreover, we have that $\|X_t\|^2 \leq \kappa_x$. Therefore,

$$1 + u^\top u = 1 + \sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} X_t \leq 1 + \sigma^{-2} \kappa_x \lambda_1(\Sigma_{0,A_t}) \leq 1 + \sigma^{-2} \kappa_x \lambda_{1,0} = c_2 \,.$$

This allows us to bound the second logarithmic term in Equation (A.22) as

$$\begin{aligned}
&\log(1 + \sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t) \,, \\
&\overset{(i)}{\leq} c_2 \log(1 + c_2^{-1} \sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t) \,, \\
&\overset{(ii)}{=} c_2 \log\det(I_{Ld} + c_2^{-1} \sigma^{-2} \bar{\Sigma}_t^{\frac{1}{2}} \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t^{\frac{1}{2}}) \,, \\
&\overset{(iii)}{=} c_2 \left[\log\det(\bar{\Sigma}_t^{-1} + c_2^{-1} \sigma^{-2} \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t}) - \log\det(\bar{\Sigma}_t^{-1}) \right] \,, \\
&\overset{(iv)}{\leq} c_2 \left[\log\det(\bar{\Sigma}_t^{-1} + \sigma^{-2} \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} \frac{X_t X_t^\top}{1 + u^\top u} \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t}) - \log\det(\bar{\Sigma}_t^{-1}) \right] \,, \\
&\overset{(v)}{=} c_2 \left[\log\det(\bar{\Sigma}_{t+1}^{-1}) - \log\det(\bar{\Sigma}_t^{-1}) \right] \,.
\end{aligned}$$

Here (i) follows from the fact that $\log(1+x) \leq c_2 \log(1 + c_2^{-1} x)$ for any $x \geq 0$ and $c_2 \geq 1$. In (ii), we use the Weinstein–Aronszajn identity. In (iii), we use the log product formula and the fact that the det is a multiplicative map. In (iv), we use that $c_2^{-1} \leq 1/(1 + u^\top u)$. Finally, (v) follows from Equation (A.24). Now we sum over all rounds and get telescoping

$$\begin{aligned}
&\sum_{t=1}^{T} \log(1 + \sigma^{-2} X_t^\top \tilde{\Sigma}_{t,A_t} \Sigma_{0,A_t}^{-1} \Gamma_{A_t} \bar{\Sigma}_t \Gamma_{A_t}^\top \Sigma_{0,A_t}^{-1} \tilde{\Sigma}_{t,A_t} X_t) \,, \\
&\quad \leq c_2 \left[\log\det(\bar{\Sigma}_{T+1}^{-1}) - \log\det(\bar{\Sigma}_1^{-1}) \right] = c_2 \log\det(\Sigma_\Psi^{\frac{1}{2}} \bar{\Sigma}_{T+1}^{-1} \Sigma_\Psi^{\frac{1}{2}}) \,, \\
&\quad \overset{(i)}{\leq} c_2 Ld \log\left(\frac{1}{Ld} \operatorname{Tr}(\Sigma_\Psi^{\frac{1}{2}} \bar{\Sigma}_{T+1}^{-1} \Sigma_\Psi^{\frac{1}{2}}) \right) \,, \\
&\quad \overset{(ii)}{\leq} c_2 Ld \log(\lambda_1(\Sigma_\Psi^{\frac{1}{2}} \bar{\Sigma}_{T+1}^{-1} \Sigma_\Psi^{\frac{1}{2}})) \overset{(iii)}{\leq} c_2 Ld \log\left(1 + K \kappa_b \lambda_{1,\Psi} \left(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda_{1,0}^2 \left(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\lambda_{d,0}} \right)} \right) \right) \,,
\end{aligned}$$

In (i) we use the inequality of arithmetic and geometric means. In (ii) we bound all eigenvalues in the trace by the maximum eigenvalue. In (iii) we use the result in Equation (A.14). We combine the upper bounds for both logarithmic terms and get

$$\mathbb{E}\left[\sum_{t=1}^{T} \|X_t\|^2_{\hat{\Sigma}_{t,A_t}}\right] \leq Kdc_{\text{A}} \log\left(1 + \frac{\kappa_x \lambda_{1,0} T}{\sigma^2 d}\right) + Ldc_1c_2 \log\left(1 + K\kappa_b \lambda_{1,\Psi}\left(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda^2_{1,0}\left(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\lambda_{d,0}}\right)}\right)\right).$$

Finally, we set $c_{\text{E}} = c_1c_2$, which concludes the proof for the general case. To retrieve the result in Theorem 1, we only need to set $\lambda_{1,0} = \lambda_{d,0} = \sigma_0^2$ and $\lambda_{1,\Psi} = \sigma_\Psi^2$ since we assumed that $\Sigma_\Psi = \sigma_\Psi^2 I_{Ld}$ and that $\Sigma_{0,a} = \sigma_0^2 I_d$ for any $a \in [K]$. In that case, the second term simplifies as

$$\begin{aligned}\log\left(1 + K\kappa_b \lambda_{1,\Psi}\left(\frac{1}{\lambda_{d,0}} - \frac{1}{\lambda^2_{1,0}\left(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\lambda_{d,0}}\right)}\right)\right) &= \log\left(1 + K\kappa_b \sigma_\Psi^2\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_0^4\left(\frac{\kappa_x T}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}\right)\right), \\ &= \log\left(1 + K\kappa_b \sigma_\Psi^2 \frac{T\kappa_x}{T\kappa_x\sigma_0^2 + \sigma^2}\right).\end{aligned}$$

□

A.4 Additional Experiments

We provide additional experiments where we evaluate `meTS` using synthetic and real-world problems, and compare it to baselines that either ignore or partially use effect parameters. In each plot, we report the averages and standard errors of the quantities. Both settings are described in Section 3.4.

A.4.1 Synthetic Experiments

In Figures A.1 and A.2, we report regret from 12 experiments with horizon $T = 5000$, where we vary K and d and use both linear and logistic rewards. For the linear setting, we compare `meTS-Lin` (Section 3.2.2), `LinUCB` (Abbasi-Yadkori et al., 2011), `LinTS` (Agrawal and Goyal, 2013a) and `HierTS` (Hong et al., 2022b). For the logistic setting, we compare `meTS-GLM` (Section 3.2.3), `meTS-Lin` (Section 3.2.2), `UCB-GLM` (Li et al., 2017), `GLM-TS` (Chapelle and Li, 2012) and `HierTS` (Hong et al., 2022b). We also include the factored approximation of `meTS` (`meTS-Lin-Fa` and `meTS-GLM-Fa`). In all experiments, we observe that `meTS-Lin` and `meTS-Fa` outperform other baselines that ignore the effect parameters or incorporate them partially. We also notice that the gain in performance becomes smaller when K/L decreases.

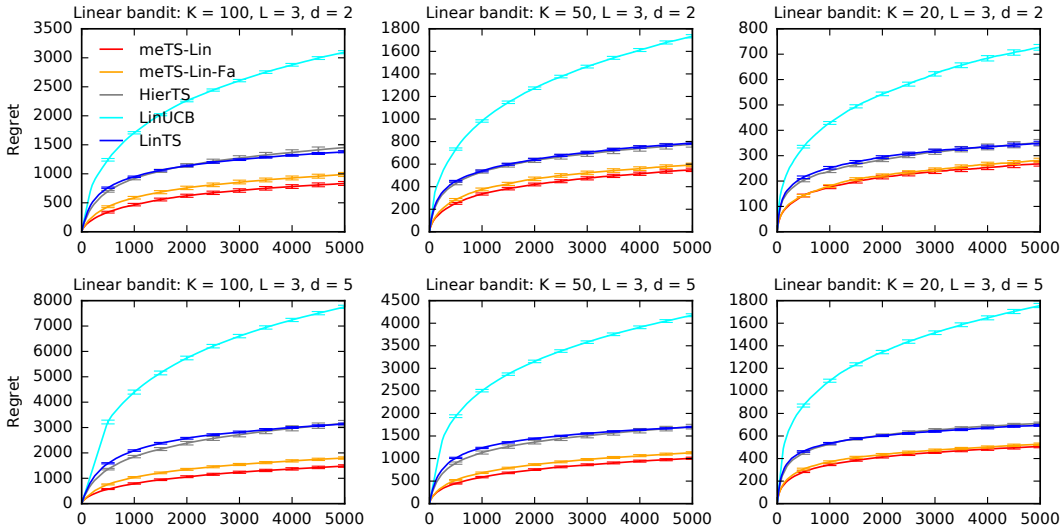


Figure A.1: Regret of `meTS-Lin` on synthetic linear bandit problems with varying feature dimension $d \in \{2, 5\}$ and number of actions $K \in \{20, 50, 100\}$.

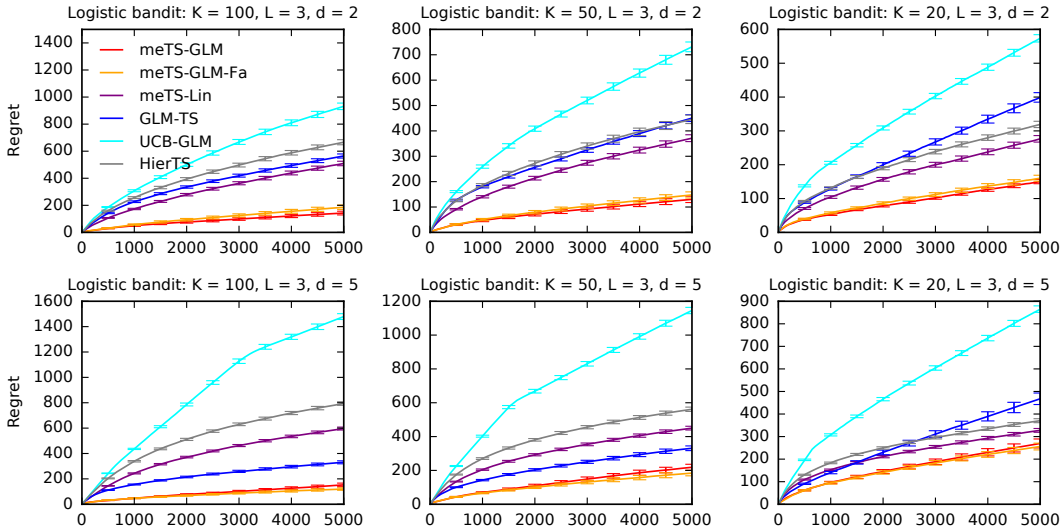


Figure A.2: Regret of `meTS-GLM` on synthetic logistic bandit problems with varying feature dimension $d \in \{2, 5\}$ and number of actions $K \in \{20, 50, 100\}$.

A.4.2 MovieLens Experiments

We plot the regret of `meTS` and the baselines up to $T = 5000$ rounds in Figures A.3 and A.4. We observe that `meTS` outperforms the other baselines. This is despite the fact that we did not fine-tune the mixing weights, which attests to the robustness of our approach to model misspecification. Similarly to the synthetic problems, we observe that the gap in performance between `meTS` and other baselines is less significant when K/L is small.

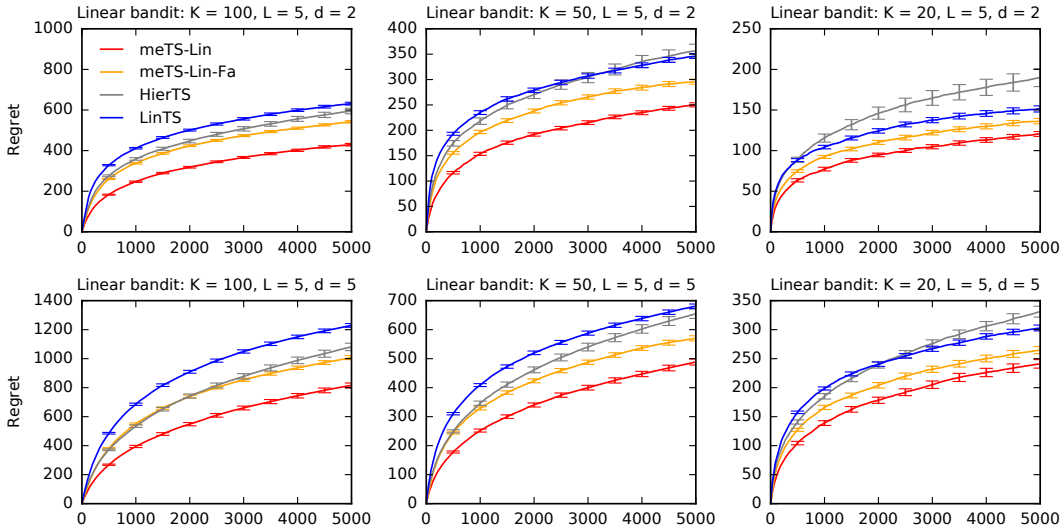


Figure A.3: Regret of `meTS-Lin` on the MovieLens dataset with linear rewards and varying feature dimension $d \in \{2, 5\}$ and number of actions $K \in \{20, 50, 100\}$.

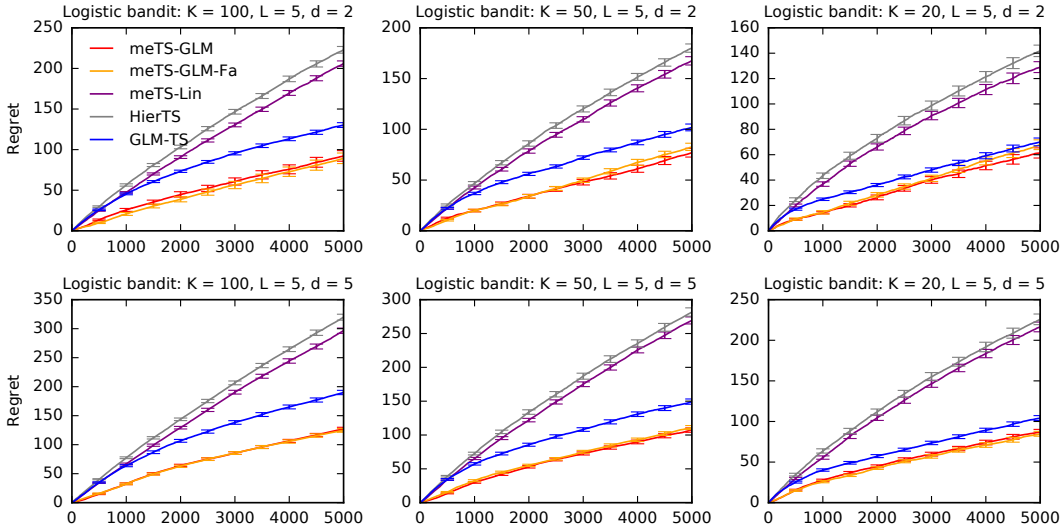


Figure A.4: Regret of `meTS-GLM` on the MovieLens dataset with logistic rewards and varying feature dimension $d \in \{2, 5\}$ and number of actions $K \in \{20, 50, 100\}$.

A.4.3 Robustness to Model Misspecification

We conduct additional synthetic experiments where the hyper-parameters do not match the parameters of the bandit environment to assess the robustness of our approach to misspecification. We provide results for this experiment in Figure A.5. Here we consider the setting described in Section 3.4.1 except that the true hyper-parameters are misspecified as follows. At each run, we sample uniformly 4 misspecification constants c_1, c_2, c_3, and c_4 from $(0, 2)$ and set the hyper-parameters of `meTS-Lin` as $c_1\Sigma_\Psi$, $c_2\mu_\Psi$, $c_3\Sigma_{0,a}$, and $c_4 b_a$ for any $a \in [K]$; where Σ_Ψ, μ_Ψ, $\Sigma_{0,a}$, and b_a for $a \in [K]$ are the true hyper-parameters. Model

misspecification is only applied to `meTS-Lin` and we refer to it as `meTS-Lin-mis`. We compare it to `meTS-Lin` and the other baselines, all with the true hyper-parameters. Although the baselines are not misspecified, `meTS-Lin-mis` still performs better. `meTS-Lin-mis` also performs similarly to `meTS-Lin` (with true hyper-parameters).

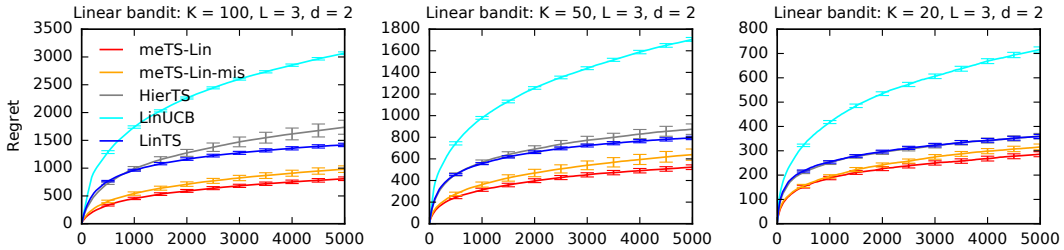


Figure A.5: Regret of *misspecified* `meTS-Lin` on synthetic bandit problems with a varying number of actions K. Here, the *misspecified* `meTS`, `meTS-Lin-mis`, is compared to baselines with true hyper-parameters.

A.4.4 Effect of Action Uncertainty

As we mentioned in Section 3.4.1 and predicted by our Bayes regret bound, learning the effect parameters is most beneficial when they are more uncertain than the action parameters. In this section, we support this claim by conducting an experiment where the initial uncertainty of action parameters is greater than the initial uncertainty of the effect parameters. Precisely, we consider the setting described in Section 3.4.1 except that we set $\Sigma_\Psi = I_{Ld}$ and $\Sigma_{0,a} = 3I_d$ for all $a \in [K]$. We report the results in Figure A.6. By comparing Figure A.6 to Figure 3.2, we observe that `meTS-Lin` still outperforms the baselines but the gap in performance shrinks when the action parameters are more uncertain than the effect parameters.

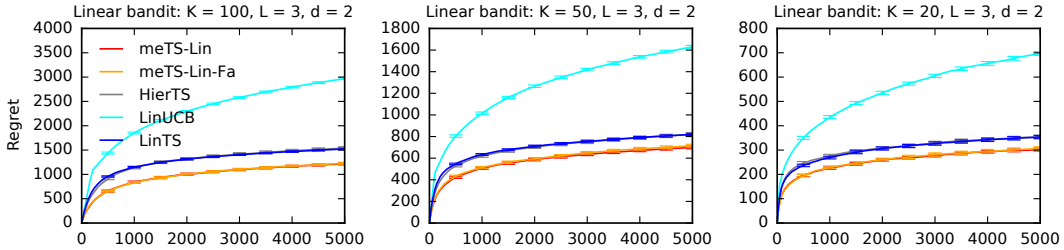


Figure A.6: Regret of `meTS-Lin` on synthetic bandit problems with a varying number of actions K, where the action parameters are more uncertain than the effect parameters.

Chapter B

Supplementary Materials for Chapter 4

Contents

B.1 Posterior for Linear Diffusion Models

Our posterior approximation builds on the simplified setting where the diffusion model is fully linear, i.e., each link function f_ℓ is linear in ψ_ℓ. This linear case, studied in our earlier workshop paper (Aouali, 2023), serves as the analytical foundation for our posterior approximation used in the general non-linear case. In Section B.2, we show how the exact posteriors derived in this linear setting inspire our efficient approximation, which extends naturally to practical diffusion models that are typically highly non-linear.

B.1.1 Linear Diffusion Models

Here, we assume the link functions f_ℓ are linear such as $f_\ell(\psi_\ell) = \mathrm{W}_\ell \psi_\ell$ for $\ell \in [L]$, where $\mathrm{W}_\ell \in \mathbb{R}^{d \times d}$ are *known mixing matrices*. Then, Equation (4.1) becomes a linear Gaussian system (LGS) (Bishop, 2006) and can be summarized as follows

$$\begin{aligned} \psi_L &\sim \mathcal{N}(0, \Sigma_{L+1}) \,, && \\ \psi_{\ell-1} \mid \psi_\ell &\sim \mathcal{N}(\mathrm{W}_\ell \psi_\ell, \Sigma_\ell) \,, && \forall \ell \in [L]/\{1\} \,, \\ \theta_a \mid \psi_1 &\sim \mathcal{N}(\mathrm{W}_1 \psi_1, \Sigma_1) \,, && \forall a \in [K] \,, \\ R_t \mid X_t, A_t, \theta, (\psi_\ell)_{\ell \in [L]} &\sim p(\cdot \mid X_t; \theta_{A_t}) \,, && \forall t \in [T] \,. \end{aligned} \tag{B.1}$$

This model is important because it yields closed-form posteriors when the reward distribution is linear-Gaussian, i.e., $p(\cdot \mid x; \theta_a) = \mathcal{N}(\cdot; x^\top \theta_a, \sigma^2)$. This allows bounding the Bayes regret of `sDM`. For practice, the posterior expressions are used to motivate efficient approximations for the general case in Equation (4.1) as we show in Section 4.2.1.

B.1.2 Posterior Expressions for Linear Diffusion Models

Recall that the reward distribution is modeled as a generalized linear model (GLM) (McCullagh and Nelder, 1989), allowing for non-linear rewards even when the diffusion links are linear. This non-linearity in the reward distribution prevents closed-form posteriors. However, since the non-linearity arises only through the reward likelihood, we approximate it by a Gaussian, leading to efficient posterior updates that are exact whenever the reward model itself is Gaussian; a special case of the GLM framework. Precisely, let $\hat{B}_{t,a}$ and $\hat{G}_{t,a}$ denote the MLE (see the remark below for practical considerations) and the Hessian of the negative log-likelihood, respectively:

$$\hat{B}_{t,a} = \operatorname*{argmax}_{\theta_a \in \mathbb{R}^d} \sum_{i \in S_{t,a}} \log p(R_i \mid X_i; \theta_a), \qquad \hat{G}_{t,a} = \sum_{i \in S_{t,a}} \dot{g}\big(X_i^\top \hat{B}_{t,a}\big) X_i X_i^\top, \tag{B.2}$$

where $S_{t,a} = \{i \in [t-1] : A_i = a\}$ is the set of rounds in which action a was taken up to round t. We approximate the likelihood as

$$\prod_{i \in S_{t,a}} p(R_i \mid X_i; \theta_a) \;\propto\; \exp\Big(-\tfrac{1}{2}(\theta_a - \hat{B}_{t,a})^\top \hat{G}_{t,a} (\theta_a - \hat{B}_{t,a})\Big) \,, \tag{B.3}$$

which makes all subsequent posteriors Gaussian. Once this approximation is done, all other derivations of the action posterior and latent posteriors are exact.

Remark 8. *The MLE may be ill-posed. In practice, we maximize an ℓ_2-regularized estimator.*

Action posterior. The conditional action posterior becomes

$$p(\theta_a \mid \psi_1, H_{t,a}) \;\approx\; \mathcal{N}\big(\theta_a; \hat{\mu}_{t,a}, \hat{\Sigma}_{t,a}\big),$$

with parameters

$$\hat{\Sigma}_{t,a}^{-1} = \Sigma_1^{-1} + \hat{G}_{t,a}, \qquad \hat{\mu}_{t,a} = \hat{\Sigma}_{t,a}\Big(\Sigma_1^{-1} \mathrm{W}_1 \psi_1 + \hat{G}_{t,a} \hat{B}_{t,a}\Big) \,. \tag{B.4}$$

Latent posteriors. For each $\ell \in [L] \setminus \{1\}$, the conditional latent posterior is

$$p(\psi_{\ell-1} \mid \psi_\ell, H_t) \;\approx\; \mathcal{N}\big(\psi_{\ell-1}; \bar{\mu}_{t,\ell-1}, \bar{\Sigma}_{t,\ell-1}\big),$$

where

$$\bar{\Sigma}_{t,\ell-1}^{-1} = \Sigma_\ell^{-1} + \bar{G}_{t,\ell-1}, \qquad \bar{\mu}_{t,\ell-1} = \bar{\Sigma}_{t,\ell-1}\big(\Sigma_\ell^{-1}\mathrm{W}_\ell\psi_\ell + \bar{B}_{t,\ell-1}\big)\,. \tag{B.5}$$

The top-layer posterior is

$$p(\psi_L \mid H_t) \;\approx\; \mathcal{N}\big(\psi_L; \bar{\mu}_{t,L}, \bar{\Sigma}_{t,L}\big),$$

with

$$\bar{\Sigma}_{t,L}^{-1} = \Sigma_{L+1}^{-1} + \bar{G}_{t,L}, \qquad \bar{\mu}_{t,L} = \bar{\Sigma}_{t,L}\bar{B}_{t,L}. \tag{B.6}$$

Recursive updates. The matrices $\bar{G}_{t,\ell}$ and $\bar{B}_{t,\ell}$ for $\ell \in [L]$ are defined recursively. The base recursion is

$$\bar{G}_{t,1} = \mathrm{W}_1^\top \sum_{a=1}^{K} \big(\Sigma_1^{-1} - \Sigma_1^{-1}\hat{\Sigma}_{t,a}\Sigma_1^{-1}\big)\mathrm{W}_1, \qquad \bar{B}_{t,1} = \mathrm{W}_1^\top\Sigma_1^{-1}\sum_{a=1}^{K}\hat{\Sigma}_{t,a}\hat{G}_{t,a}\hat{B}_{t,a}. \tag{B.7}$$

Then, for $\ell \in [L] \setminus \{1\}$, the recursive step is

$$\bar{G}_{t,\ell} = \mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\,\mathrm{W}_\ell, \qquad \bar{B}_{t,\ell} = \mathrm{W}_\ell^\top\Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\bar{B}_{t,\ell-1}. \tag{B.8}$$

Discussion. This completes the derivation of the linear posterior approximation. All posteriors are Gaussian and exact whenever the reward distribution follows a linear-Gaussian model, i.e.

$$p(\cdot \mid x; \theta_a) = \mathcal{N}(\cdot; x^\top\theta_a, \sigma^2).$$

In this case, the above posterior updates coincide with the exact Bayesian updates, while for general GLMs they serve as efficient and accurate approximations.

B.2 Posterior for Non-Linear Diffusion Models

The general diffusion model (Equation (4.1), which is our case of interest) involves two sources of non-linearity: (i) the reward distribution $p(\cdot \mid x; \theta)$, which may follow a non-linear generalized linear model (GLM), and (ii) the diffusion links $f_\ell(\psi_\ell)$, which can be arbitrary non-linear functions. Both sources make the posterior intractable, and therefore two approximations are needed.

First approximation (likelihood). We first approximate the reward likelihood by a Gaussian density (as we did above in Equation (B.3)). After this substitution, the model becomes conditionally Gaussian given the latent variables. This step is exact when the reward model is linear-Gaussian, and approximate otherwise.

Second approximation (diffusion hierarchy). Even after the likelihood is approximated, the diffusion hierarchy remains non-linear because of the non-linear mappings f_ℓ. To handle this, we reuse the exact Gaussian posteriors derived for the linear diffusion case (Section B.1.2) and generalize them as follows:

- Replace each linear mapping $\mathrm{W}_\ell \psi_\ell$ by its non-linear counterpart $f_\ell(\psi_\ell)$, which represents the mean of the diffusion prior at layer ℓ.
- Remove matrix multiplications involving W_ℓ in the recursive updates.

This step can be viewed as extending the linear-Gaussian posterior updates to a general non-linear setting. This allows fast sampling and updating of the posterior, without heavy standard posterior approximation techniques. Of course, this is a purely empirical and heuristic based approximation that does not come with guarantees but performs very well in practice.

Resulting approximation. The two steps above yield a posterior where each conditional factor $p(\theta_a \mid \psi_1, H_{t,a})$ and $p(\psi_{\ell-1} \mid \psi_\ell, H_t)$ remains Gaussian with updated means and covariances, while the overall model retains the hierarchical diffusion structure. The approximation satisfies two desirable properties: it exactly recovers the diffusion prior when no data is available, and as more data is observed, the likelihood terms dominate and the prior influence fades naturally.

B.3 Connection to Two-Level Hierarchies

The linear diffusion Equation (B.1) can be marginalized into a 2-level hierarchy using two different strategies. To simplify, we let $\Sigma_\ell = \sigma_\ell^2 I_d$. The first one yields,

$$\begin{aligned} \psi_L &\sim \mathcal{N}(0, \sigma_{L+1}^2 \mathrm{B}_L \mathrm{B}_L^\top)\,, \\ \theta_a \mid \psi_L &\sim \mathcal{N}(\psi_L,\, \Omega_1)\,, \qquad\qquad \forall a \in [K]\,, \end{aligned} \tag{B.9}$$

with $\Omega_1 = \sigma_1^2 I_d + \sum_{\ell=1}^{L-1} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top$ and $\mathrm{B}_\ell = \prod_{i=1}^{\ell} \mathrm{W}_i$. The second strategy yields,

$$\begin{aligned} \psi_1 &\sim \mathcal{N}(0, \Omega_2)\,, \\ \theta_a \mid \psi_1 &\sim \mathcal{N}(\psi_1,\, \sigma_1^2 I_d)\,, \qquad\qquad \forall a \in [K]\,, \end{aligned} \tag{B.10}$$

where $\Omega_2 = \sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \mathrm{B}_\ell \mathrm{B}_\ell^\top$. Recently, `HierTS` (Hong et al., 2022b) was developed for such two-level graphical models, and we call `HierTS` under Equation (B.9) by `HierTS-1` and `HierTS` under Equation (B.10) by `HierTS-2`. Then, we start by highlighting the differences between these two variants of `HierTS`. First, their regret bounds scale as

$$\texttt{HierTS-1} : \tilde{\mathcal{O}}\Big(\sqrt{Td(K\sum_{\ell=1}^{L}\sigma_\ell^2 + L\sigma_{L+1}^2)}\Big)\,, \quad \texttt{HierTS-2} : \tilde{\mathcal{O}}\Big(\sqrt{Td(K\sigma_1^2 + \sum_{\ell=1}^{L}\sigma_{\ell+1}^2)}\Big)\,.$$

When $K \approx L$, the regret bounds of `HierTS-1` and `HierTS-2` are similar. However, when $K > L$, `HierTS-2` outperforms `HierTS-1`. This is because `HierTS-2` puts more uncertainty on a single d-dimensional latent parameter ψ_1, rather than K individual d-dimensional action parameters θ_a. More importantly, `HierTS-1` implicitly assumes that action parameters θ_a are conditionally independent given ψ_L, which is not true. Consequently, `HierTS-2` outperforms `HierTS-1`. Note that, under the linear diffusion model Equation (B.1), `sDM` and `HierTS-2` have roughly similar regret bounds. Specifically, their regret bounds dependency on K is identical, where both methods involve multiplying K by σ_1^2, and both enjoy improved performance compared to `HierTS-1`. That said, note that

Theorem 7 and Proposition 7 provide an understanding of how sDM's regret scales under linear link functions f_ℓ, and do not say that using sDM is better than using HierTS when the link functions f_ℓ are linear since the latter can be obtained by a proper marginalization of latent parameters (i.e., HierTS-2 instead of HierTS-1). While such a comparison is not the goal of this work, we still provide it for completeness next.

When the mixing matrices W_ℓ are dense (i.e., assumption **(A5)** is not applicable), sDM and HierTS-2 have comparable regret bounds and computational efficiency. However, under the sparsity assumption **(A5)** and with mixing matrices that allow for conditional independence of ψ_1 coordinates given ψ_2, sDM enjoys a computational advantage over HierTS-2. This advantage explains why works focusing on multi-level hierarchies typically benchmark their algorithms against two-level structures akin to HierTS-1, rather than the more competitive HierTS-2. This is also consistent with prior works in Bayesian bandits using multi-level hierarchies, such as Tree-based priors (Hong et al., 2022a), which compared their method to HierTS-1. In line with this, we also compared sDM with HierTS-1 in our experiments. But this is only given for completeness as this is not the aim of Theorem 7 and Proposition 7. More importantly, HierTS is inapplicable in the general case in Equation (4.1) with non-linear link functions since the latent parameters cannot be analytically marginalized.

B.4 Formal Theory

We analyze sDM assuming that: **(A0)** The true environment parameters θ_* and $\psi_{*,\ell}$ are drawn from the same prior distribution used by sDM, as is standard in Bayes regret analysis (Russo and Van Roy, 2014); we thus use θ_* and θ ($\psi_{*,\ell}$ and ψ_ℓ) interchangeably throughout. **(A1)** The rewards are linear $p(\cdot \mid x; \theta_a) = \mathcal{N}(\cdot; x^\top \theta_a, \sigma^2)$. **(A2)** The link functions f_ℓ are linear such as $f_\ell(\psi_\ell) = \mathrm{W}_\ell \psi_\ell$ for $\ell \in [L]$, where $\mathrm{W}_\ell \in \mathbb{R}^{d \times d}$ are *known mixing matrices*. This leads to a structure with L layers of linear Gaussian relationships detailed in Section B.1.1. In particular, this leads to closed-form posteriors given in Section B.1.2 that inspired our approximation and enable theory similar to linear bandits (Agrawal and Goyal, 2013a). However, proofs are not the same, and technical challenges remain (explained in Section B.5).

Although our result holds for milder assumptions, we make additional simplifications for clarity and interpretability. We assume that **(A3)** Contexts satisfy $\|X_t\|_2^2 = 1$ for any $t \in [T]$. Note that **(A3)** can be relaxed to any contexts X_t with bounded norms $\|X_t\|_2$. **(A4)** Mixing matrices and covariances satisfy $\lambda_1(\mathrm{W}_\ell^\top \mathrm{W}_\ell) = 1$ for any $\ell \in [L]$ and $\Sigma_\ell = \sigma_\ell^2 I_d$ for any $\ell \in [L+1]$. In this section, we write $\tilde{\mathcal{O}}$ for the big-O notation up to polylogarithmic factors. We start by stating our bound for sDM.

Theorem 7. *Let* $\sigma_{MAX}^2 = \max_{\ell \in [L+1]} 1 + \frac{\sigma_\ell^2}{\sigma^2}$. *There exists a constant* $c > 0$ *such that for any* $\delta \in (0, 1)$, *the Bayes regret of* sDM *under* **(A1)**, **(A2)**, **(A3)** *and* **(A4)** *is bounded*

as

$$\mathcal{BR}(T) \leq \sqrt{2T\big(\mathcal{R}^{ACT}(T) + \sum_{\ell=1}^{L} \mathcal{R}_\ell^{LAT}\big)\log(1/\delta)\big)} + cT\delta\,,$$
$$\mathcal{R}^{ACT}(T) = c_0 dK \log\big(1 + \frac{T\sigma_1^2}{d\sigma^2}\big),\ c_0 = \frac{\sigma_1^2}{\log\Big(1 + \frac{\sigma_1^2}{\sigma^2}\Big)}\,,$$
$$\mathcal{R}_\ell^{LAT} = c_\ell d \log\big(1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\big), c_\ell = \frac{\sigma_{\ell+1}^2 \sigma_{MAX}^{2\ell}}{\log\Big(1 + \frac{\sigma_{\ell+1}^2}{\sigma^2}\Big)}, \tag{B.11}$$

Equation (B.11) holds for any $\delta \in (0,1)$. In particular, the term $cT\delta$ is constant when $\delta = 1/T$. Then, the bound is $\tilde{\mathcal{O}}\Big(\sqrt{T(dK\sigma_1^2 + d\sum_{\ell=1}^{L} \sigma_{\ell+1}^2 \sigma_{\text{MAX}}^{2\ell})}\Big)$, and this dependence on the horizon T aligns with prior Bayes regret bounds. The bound comprises $L+1$ main terms, $\mathcal{R}^{\text{ACT}}(T)$ and $\mathcal{R}_\ell^{\text{LAT}}$ for $\ell \in [L]$. First, $\mathcal{R}^{\text{ACT}}(T)$ relates to action parameters learning, conforming to a standard form (Lu and Van Roy, 2019). Similarly, $\mathcal{R}_\ell^{\text{LAT}}$ is associated with learning the ℓ-th latent parameter.

To include more structure, we propose the *sparsity* assumption **(A5)** $\mathrm{W}_\ell = (\bar{\mathrm{W}}_\ell, 0_{d,d-d_\ell})$, where $\bar{\mathrm{W}}_\ell \in \mathbb{R}^{d\times d_\ell}$ for any $\ell \in [L]$. Note that **(A5)** is not an assumption when $d_\ell = d$ for any $\ell \in [L]$. Notably, **(A5)** incorporates a plausible structural characteristic that a diffusion model could capture.

Proposition 7 (Sparsity)**.** *Let* $\sigma_{MAX}^2 = \max_{\ell\in[L+1]} 1 + \frac{\sigma_\ell^2}{\sigma^2}$. *There exists a constant* $c > 0$ *such that for any* $\delta \in (0,1)$*, the Bayes regret of* `sDM` *under* **(A1)**, **(A2)**, **(A3)**, **(A4)** *and* **(A5)** *is bounded as*

$$\mathcal{BR}(T) \leq \sqrt{2T\big(\mathcal{R}^{ACT}(T) + \sum_{\ell=1}^{L} \tilde{\mathcal{R}}_\ell^{LAT}\big)\log(1/\delta)\big)} + cT\delta\,,$$
$$\tilde{\mathcal{R}}^{ACT}(T) = c_0 dK \log\big(1 + \frac{T\sigma_1^2}{d\sigma^2}\big),\ c_0 = \frac{\sigma_1^2}{\log\Big(1 + \frac{\sigma_1^2}{\sigma^2}\Big)}\,,$$
$$\mathcal{R}_\ell^{LAT} = c_\ell d_\ell \log\big(1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\big), c_\ell = \frac{\sigma_{\ell+1}^2 \sigma_{MAX}^{2\ell}}{\log\Big(1 + \frac{\sigma_{\ell+1}^2}{\sigma^2}\Big)}, \tag{B.12}$$

From Proposition 7, our bounds scales as

$$\mathcal{BR}(T) = \tilde{\mathcal{O}}\Big(\sqrt{T(dK\sigma_1^2 + \sum_{\ell=1}^{L} d_\ell \sigma_{\ell+1}^2 \sigma_{\text{MAX}}^{2\ell})}\Big)\,. \tag{B.13}$$

B.5 Regret proof

Important notation clarification. Throughout this proof, we operate under the standard Bayesian bandit framework where the true environment parameters θ_* are drawn

from the same prior distribution that `sDM` uses for posterior inference. Specifically, the true action parameters $\theta_{*,a}$ for $a \in [K]$ and the true latent parameters $\psi_{*,\ell}$ for $\ell \in [L]$ are assumed to be sampled according to the generative process in Equation (B.1). As a consequence, the true parameters θ_* and the model parameters θ used in our derivations follow the same distribution, and we use them interchangeably throughout the proof to simplify notation.

B.5.1 Proof Sketch

We start with the following standard lemma upon which we build our analysis (Aouali et al., 2023b).

Lemma 4. *Assume that* $p(\theta_a \mid H_t) = \mathcal{N}(\theta_a; \check{\mu}_{t,a}, \check{\Sigma}_{t,a})$ *for any* $a \in [K]$, *then for any* $\delta \in (0,1)$,

$$\mathcal{BR}(T) \leq \sqrt{2T\log(1/\delta)}\sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \|X_t\|^2_{\check{\Sigma}_{t,A_t}}\right]} + cT\delta\,, \qquad \text{where } c > 0 \text{ is a constant}. \tag{B.14}$$

Applying Lemma 4 requires proving that the *marginal* action-posterior densities of $\theta_a \mid H_t$ in Equation (4.3) are Gaussian and computing their covariances, while we only know the *conditional* action-posteriors $p(\theta_a \mid \psi_1, H_t)$ and latent-posteriors $p(\psi_{\ell-1} \mid \psi_\ell, H_t)$. This is achieved by leveraging the preservation properties of the family of Gaussian distributions (Koller and Friedman, 2009) and the total covariance decomposition (Weiss, 2005) which leads to the next lemma.

Lemma 5. *Let* $t \in [T]$ *and* $a \in [K]$, *then the marginal covariance matrix* $\check{\Sigma}_{t,a}$ *reads*

$$\check{\Sigma}_{t,a} = \hat{\Sigma}_{t,a} + \sum_{\ell \in [L]} \mathrm{P}_{a,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{a,\ell}^\top\,, \qquad \text{where } \mathrm{P}_{a,\ell} = \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1 \prod_{i=1}^{\ell-1} \bar{\Sigma}_{t,i}\Sigma_{i+1}^{-1}\mathrm{W}_{i+1}. \tag{B.15}$$

The marginal covariance matrix $\check{\Sigma}_{t,a}$ in Equation (B.15) decomposes into $L+1$ terms. The first term corresponds to the posterior uncertainty of $\theta_a \mid \psi_1$. The remaining L terms capture the posterior uncertainties of ψ_L and $\psi_{\ell-1} \mid \psi_\ell$ for $\ell \in [L]/\{1\}$. These are then used to quantify the posterior information gain of latent parameters after one round as follows.

Lemma 6 (Posterior information gain)**.** *Let* $t \in [T]$ *and* $\ell \in [L]$, *then*

$$\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} \succeq \sigma^{-2}\sigma_{\scriptscriptstyle MAX}^{-2\ell} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}\,, \qquad \text{where } \sigma^2_{\scriptscriptstyle MAX} = \max_{\ell \in [L+1]} 1 + \frac{\sigma_\ell^2}{\sigma^2}\,. \tag{B.16}$$

Finally, Lemma 5 is used to decompose $\|X_t\|^2_{\check{\Sigma}_{t,A_t}}$ in Equation (B.14) into $L+1$ terms. Each term is bounded thanks to Lemma 6. This results in the Bayes regret bound in Theorem 7.

B.5.2 Proof of lemma 5

In this proof, we heavily rely on the total covariance decomposition (Weiss, 2005). Also, refer to (Hong et al., 2022b, Section 5.2) for a brief introduction to this decomposition. Now, from Equation (B.4), we have that

$$\begin{aligned}\operatorname{cov}\left[\theta_a \,|\, H_t, \psi_1\right] &= \hat{\Sigma}_{t,a} = \left(\hat{G}_{t,a} + \Sigma_1^{-1}\right)^{-1}, \\ \mathbb{E}\left[\theta_a \,|\, H_t, \psi_1\right] &= \hat{\mu}_{t,a} = \hat{\Sigma}_{t,a}\left(\hat{G}_{t,a}\hat{B}_{t,a} + \Sigma_1^{-1}\mathrm{W}_1\psi_1\right).\end{aligned}$$

First, given H_t, $\operatorname{cov}\left[\theta_a \,|\, H_t, \psi_1\right] = \left(\hat{G}_{t,a} + \Sigma_1^{-1}\right)^{-1}$ is constant. Thus

$$\mathbb{E}\left[\operatorname{cov}\left[\theta_a \,|\, H_t, \psi_1\right] | \, H_t\right] = \operatorname{cov}\left[\theta_a \,|\, H_t, \psi_1\right] = \left(\hat{G}_{t,a} + \Sigma_1^{-1}\right)^{-1} = \hat{\Sigma}_{t,a}\,.$$

In addition, given H_t, $\hat{\Sigma}_{t,a}$, $\hat{G}_{t,a}$ and $\hat{B}_{t,a}$ are constant. Thus

$$\begin{aligned}\operatorname{cov}\left[\mathbb{E}\left[\theta_a \,|\, H_t, \psi_1\right] | \, H_t\right] &= \operatorname{cov}\left[\hat{\Sigma}_{t,a}\left(\hat{G}_{t,a}\hat{B}_{t,a} + \Sigma_1^{-1}\mathrm{W}_1\psi_1\right) \,\middle|\, H_t\right], \\ &= \operatorname{cov}\left[\hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\psi_1 \,\middle|\, H_t\right], \\ &= \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\operatorname{cov}\left[\psi_1 \,|\, H_t\right]\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,a}, \\ &= \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\overline{\overline{\Sigma}}_{t,1}\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,a},\end{aligned}$$

where $\overline{\overline{\Sigma}}_{t,1} = \operatorname{cov}\left[\psi_1 \,|\, H_t\right]$ is the marginal posterior covariance of ψ_1. Finally, the total covariance decomposition (Weiss, 2005; Hong et al., 2022b) yields that

$$\begin{aligned}\check{\Sigma}_{t,a} = \operatorname{cov}\left[\theta_a \,|\, H_t\right] &= \mathbb{E}\left[\operatorname{cov}\left[\theta_a \,|\, H_t, \psi_1\right] | \, H_t\right] + \operatorname{cov}\left[\mathbb{E}\left[\theta_a \,|\, H_t, \psi_1\right] | \, H_t\right], \\ &= \hat{\Sigma}_{t,a} + \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\overline{\overline{\Sigma}}_{t,1}\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,a}, \qquad \text{(B.17)}\end{aligned}$$

However, $\overline{\overline{\Sigma}}_{t,1} = \operatorname{cov}\left[\psi_1 \,|\, H_t\right]$ is different from $\bar{\Sigma}_{t,1} = \operatorname{cov}\left[\psi_1 \,|\, H_t, \psi_2\right]$ that we already derived in Equation (B.5). Thus we do not know the expression of $\overline{\overline{\Sigma}}_{t,1}$. But we can use the same total covariance decomposition trick to find it. Precisely, let $\overline{\overline{\Sigma}}_{t,\ell} = \operatorname{cov}\left[\psi_\ell \,|\, H_t\right]$ for any $\ell \in [L]$. Then we have that

$$\begin{aligned}\bar{\Sigma}_{t,1} = \operatorname{cov}\left[\psi_1 \,|\, H_t, \psi_2\right] &= \left(\Sigma_2^{-1} + \bar{G}_{t,1}\right)^{-1}, \\ \bar{\mu}_{t,1} = \mathbb{E}\left[\psi_1 \,|\, H_t, \psi_2\right] &= \bar{\Sigma}_{t,1}\left(\Sigma_2^{-1}\mathrm{W}_2\psi_2 + \bar{B}_{t,1}\right).\end{aligned}$$

First, given H_t, $\operatorname{cov}\left[\psi_1 \,|\, H_t, \psi_2\right] = \left(\Sigma_2^{-1} + \bar{G}_{t,1}\right)^{-1}$ is constant. Thus

$$\mathbb{E}\left[\operatorname{cov}\left[\psi_1 \,|\, H_t, \psi_2\right] | \, H_t\right] = \operatorname{cov}\left[\psi_1 \,|\, H_t, \psi_2\right] = \bar{\Sigma}_{t,1}\,.$$

In addition, given H_t, $\bar{\Sigma}_{t,1}$, $\tilde{\Sigma}_{t,1}$ and $\bar{B}_{t,1}$ are constant. Thus

$$\begin{aligned}\operatorname{cov}\left[\mathbb{E}\left[\psi_1 \,|\, H_t, \psi_2\right] | \, H_t\right] &= \operatorname{cov}\left[\bar{\Sigma}_{t,1}\left(\Sigma_2^{-1}\mathrm{W}_2\psi_2 + \bar{B}_{t,1}\right) \,\middle|\, H_t\right], \\ &= \operatorname{cov}\left[\bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\psi_2 \,\middle|\, H_t\right], \\ &= \bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\operatorname{cov}\left[\psi_2 \,|\, H_t\right]\mathrm{W}_2^\top\Sigma_2^{-1}\bar{\Sigma}_{t,1}, \\ &= \bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\overline{\overline{\Sigma}}_{t,2}\mathrm{W}_2^\top\Sigma_2^{-1}\bar{\Sigma}_{t,1}.\end{aligned}$$

Finally, total covariance decomposition (Weiss, 2005; Hong et al., 2022b) leads to

$$\begin{aligned}\overline{\overline{\Sigma}}_{t,1} = \operatorname{cov}\left[\psi_1 \,|\, H_t\right] &= \mathbb{E}\left[\operatorname{cov}\left[\psi_1 \,|\, H_t, \psi_2\right] | \, H_t\right] + \operatorname{cov}\left[\mathbb{E}\left[\psi_1 \,|\, H_t, \psi_2\right] | \, H_t\right] , \\ &= \bar{\Sigma}_{t,1} + \bar{\Sigma}_{t,1}\Sigma_2^{-1}\mathrm{W}_2\overline{\overline{\Sigma}}_{t,2}\mathrm{W}_2^\top\Sigma_2^{-1}\bar{\Sigma}_{t,1} .\end{aligned}$$

Now using the techniques, this can be generalized using the same technique as above to

$$\overline{\overline{\Sigma}}_{t,\ell} = \bar{\Sigma}_{t,\ell} + \bar{\Sigma}_{t,\ell}\Sigma_{\ell+1}^{-1}\mathrm{W}_{\ell+1}\overline{\overline{\Sigma}}_{t,\ell+1}\mathrm{W}_{\ell+1}^\top\Sigma_{\ell+1}^{-1}\bar{\Sigma}_{t,\ell} , \qquad \forall \ell \in [L-1] .$$

Then, by induction, we get that

$$\overline{\overline{\Sigma}}_{t,1} = \sum_{\ell\in[L]} \bar{\mathrm{P}}_\ell\bar{\Sigma}_{t,\ell}\bar{\mathrm{P}}_\ell^\top , \qquad \forall \ell \in [L-1] ,$$

where we use that by definition $\overline{\overline{\Sigma}}_{t,L} = \operatorname{cov}\left[\psi_L \,|\, H_t\right] = \bar{\Sigma}_{t,L}$ and set $\bar{\mathrm{P}}_1 = I_d$ and $\bar{\mathrm{P}}_\ell = \prod_{i=1}^{\ell-1}\bar{\Sigma}_{t,i}\Sigma_{i+1}^{-1}\mathrm{W}_{i+1}$ for any $\ell \in [L]/\{1\}$. Plugging this in Equation (B.17) leads to

$$\begin{aligned}\check{\Sigma}_{t,a} &= \hat{\Sigma}_{t,a} + \sum_{\ell\in[L]} \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\bar{\mathrm{P}}_\ell\bar{\Sigma}_{t,\ell}\bar{\mathrm{P}}_\ell^\top\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,a} , \\ &= \hat{\Sigma}_{t,a} + \sum_{\ell\in[L]} \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\bar{\mathrm{P}}_\ell\bar{\Sigma}_{t,\ell}(\hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1)^\top , \\ &= \hat{\Sigma}_{t,a} + \sum_{\ell\in[L]} \mathrm{P}_{a,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{a,\ell}^\top ,\end{aligned}$$

where $\mathrm{P}_{a,\ell} = \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\bar{\mathrm{P}}_\ell = \hat{\Sigma}_{t,a}\Sigma_1^{-1}\mathrm{W}_1\prod_{i=1}^{\ell-1}\bar{\Sigma}_{t,i}\Sigma_{i+1}^{-1}\mathrm{W}_{i+1}$.

B.5.3 Proof of lemma 6

We prove this result by induction. We start with the base case when $\ell = 1$.

(I) Base case. Let $u = \sigma^{-1}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}X_t$ From the expression of $\bar{\Sigma}_{t,1}$ in Equation (B.5), we have that

$$\begin{aligned}\bar{\Sigma}_{t+1,1}^{-1} - \bar{\Sigma}_{t,1}^{-1} &= \mathrm{W}_1^\top\left(\Sigma_1^{-1} - \Sigma_1^{-1}(\hat{\Sigma}_{t,A_t}^{-1} + \sigma^{-2}X_tX_t^\top)^{-1}\Sigma_1^{-1} - (\Sigma_1^{-1} - \Sigma_1^{-1}\hat{\Sigma}_{t,A_t}\Sigma_1^{-1})\right)\mathrm{W}_1 , \\ &= \mathrm{W}_1^\top\left(\Sigma_1^{-1}(\hat{\Sigma}_{t,A_t} - (\hat{\Sigma}_{t,A_t}^{-1} + \sigma^{-2}X_tX_t^\top)^{-1})\Sigma_1^{-1}\right)\mathrm{W}_1 , \\ &= \mathrm{W}_1^\top\left(\Sigma_1^{-1}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}(I_d - (I_d + \sigma^{-2}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}X_tX_t^\top\hat{\Sigma}_{t,A_t}^{\frac{1}{2}})^{-1})\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}\Sigma_1^{-1}\right)\mathrm{W}_1 , \\ &= \mathrm{W}_1^\top\left(\Sigma_1^{-1}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}(I_d - (I_d + uu^\top)^{-1})\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}\Sigma_1^{-1}\right)\mathrm{W}_1 , \\ &\overset{(i)}{=} \mathrm{W}_1^\top\left(\Sigma_1^{-1}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}\frac{uu^\top}{1+u^\top u}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}}\Sigma_1^{-1}\right)\mathrm{W}_1 , \\ &\overset{(ii)}{=} \sigma^{-2}\mathrm{W}_1^\top\Sigma_1^{-1}\hat{\Sigma}_{t,A_t}\frac{X_tX_t^\top}{1+u^\top u}\hat{\Sigma}_{t,A_t}\Sigma_1^{-1}\mathrm{W}_1 .\end{aligned} \tag{B.18}$$

In (i) we use the Sherman-Morrison formula. Note that (ii) says that $\bar{\Sigma}_{t+1,1}^{-1} - \bar{\Sigma}_{t,1}^{-1}$ is one-rank which we will also need in induction step. Now, we have that $\|X_t\|^2 = 1$. Therefore,

$$1 + u^\top u = 1 + \sigma^{-2}X_t^\top\hat{\Sigma}_{t,A_t}X_t \leq 1 + \sigma^{-2}\lambda_1(\Sigma_1)\|X_t\|^2 = 1 + \sigma^{-2}\sigma_1^2 \leq \sigma_{\text{MAX}}^2 ,$$

where we use that by definition of σ^2_{MAX} in Lemma 6, we have that $\sigma^2_{\text{MAX}} \geq 1 + \sigma^{-2}\sigma_1^2$. Therefore, by taking the inverse, we get that $\frac{1}{1+u^\top u} \geq \sigma^{-2}_{\text{MAX}}$. Combining this with Equation (B.18) leads to

$$\bar{\Sigma}^{-1}_{t+1,1} - \bar{\Sigma}^{-1}_{t,1} \succeq \sigma^{-2}\sigma^{-2}_{\text{MAX}} \mathrm{W}_1^\top \Sigma_1^{-1} \hat{\Sigma}_{t,A_t} X_t X_t^\top \hat{\Sigma}_{t,A_t} \Sigma_1^{-1} \mathrm{W}_1$$

Noticing that $\mathrm{P}_{A_t,1} = \hat{\Sigma}_{t,A_t}\Sigma_1^{-1}\mathrm{W}_1$ concludes the proof of the base case when $\ell = 1$.

(II) Induction step. Let $\ell \in [L]/\{1\}$ and suppose that $\bar{\Sigma}^{-1}_{t+1,\ell-1} - \bar{\Sigma}^{-1}_{t,\ell-1}$ is one-rank and that it holds for $\ell - 1$ that

$$\bar{\Sigma}^{-1}_{t+1,\ell-1} - \bar{\Sigma}^{-1}_{t,\ell-1} \succeq \sigma^{-2}\sigma^{-2(\ell-1)}_{\text{MAX}} \mathrm{P}^\top_{A_t,\ell-1} X_t X_t^\top \mathrm{P}_{A_t,\ell-1}\,, \quad \text{where } \sigma^2_{\text{MAX}} = \max_{\ell\in[L+1]} 1 + \sigma^{-2}\sigma_\ell^2.$$

Then, we want to show that $\bar{\Sigma}^{-1}_{t+1,\ell} - \bar{\Sigma}^{-1}_{t,\ell}$ is also one-rank and that it holds that

$$\bar{\Sigma}^{-1}_{t+1,\ell} - \bar{\Sigma}^{-1}_{t,\ell} \succeq \sigma^{-2}\sigma^{-2\ell}_{\text{MAX}} \mathrm{P}^\top_{A_t,\ell} X_t X_t^\top \mathrm{P}_{A_t,\ell}\,, \qquad \text{where } \sigma^2_{\text{MAX}} = \max_{\ell\in[L+1]} 1 + \sigma^{-2}\sigma_\ell^2.$$

This is achieved as follows. Define the precision increment at level $\ell - 1$ by

$$\Delta_{t,\ell-1} := \bar{\Sigma}^{-1}_{t+1,\ell-1} - \bar{\Sigma}^{-1}_{t,\ell-1}.$$

By the induction hypothesis, $\Delta_{t,\ell-1}$ is rank-one PSD, hence there exists $u \in \mathbb{R}^d$ such that

$$\Delta_{t,\ell-1} = uu^\top.$$

Using Equation (B.8), we have for any $s \in \{t, t+1\}$:

$$\bar{\Sigma}^{-1}_{s,\ell} = \Sigma^{-1}_{\ell+1} + \bar{G}_{s,\ell} = \Sigma^{-1}_{\ell+1} + \mathrm{W}_\ell^\top \big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{s,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell.$$

Therefore,

$$\begin{aligned}\bar{\Sigma}^{-1}_{t+1,\ell} - \bar{\Sigma}^{-1}_{t,\ell} &= \bar{G}_{t+1,\ell} - \bar{G}_{t,\ell} \\ &= \mathrm{W}_\ell^\top \Sigma_\ell^{-1}\big(\bar{\Sigma}_{t,\ell-1} - \bar{\Sigma}_{t+1,\ell-1}\big)\Sigma_\ell^{-1}\mathrm{W}_\ell.\end{aligned}$$

Since $\bar{\Sigma}^{-1}_{t+1,\ell-1} = \bar{\Sigma}^{-1}_{t,\ell-1} + uu^\top$, Sherman–Morrison yields

$$\bar{\Sigma}_{t+1,\ell-1} = \big(\bar{\Sigma}^{-1}_{t,\ell-1} + uu^\top\big)^{-1} = \bar{\Sigma}_{t,\ell-1} - \frac{\bar{\Sigma}_{t,\ell-1}uu^\top\bar{\Sigma}_{t,\ell-1}}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}.$$

Hence,

$$\bar{\Sigma}_{t,\ell-1} - \bar{\Sigma}_{t+1,\ell-1} = \frac{\bar{\Sigma}_{t,\ell-1}uu^\top\bar{\Sigma}_{t,\ell-1}}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u},$$

and plugging this back gives

$$\bar{\Sigma}^{-1}_{t+1,\ell} - \bar{\Sigma}^{-1}_{t,\ell} = \mathrm{W}_\ell^\top\Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\frac{uu^\top}{1 + u^\top\bar{\Sigma}_{t,\ell-1}u}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\mathrm{W}_\ell.$$

In particular, this increment is rank-one PSD.

However, it follows from the induction hypothesis that

$$uu^\top = \bar{\Sigma}_{t+1,\ell-1}^{-1} - \bar{\Sigma}_{t,\ell-1}^{-1} \succeq \sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}\, \mathrm{P}_{A_t,\ell-1}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell-1}.$$

Therefore,

$$\begin{aligned}
\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} &= \mathrm{W}_\ell^\top \Sigma_\ell^{-1} \bar{\Sigma}_{t,\ell-1} \frac{uu^\top}{1 + u^\top \bar{\Sigma}_{t,\ell-1} u} \bar{\Sigma}_{t,\ell-1} \Sigma_\ell^{-1} \mathrm{W}_\ell \,, \\
&\succeq \mathrm{W}_\ell^\top \Sigma_\ell^{-1} \bar{\Sigma}_{t,\ell-1} \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)} \mathrm{P}_{A_t,\ell-1}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell-1}}{1 + u^\top \bar{\Sigma}_{t,\ell-1} u} \bar{\Sigma}_{t,\ell-1} \Sigma_\ell^{-1} \mathrm{W}_\ell \,, \\
&= \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + u^\top \bar{\Sigma}_{t,\ell-1} u} \mathrm{W}_\ell^\top \Sigma_\ell^{-1} \bar{\Sigma}_{t,\ell-1} \mathrm{P}_{A_t,\ell-1}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell-1} \bar{\Sigma}_{t,\ell-1} \Sigma_\ell^{-1} \mathrm{W}_\ell \,, \\
&= \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + u^\top \bar{\Sigma}_{t,\ell-1} u} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \,.
\end{aligned}$$

Finally, we use that $1 + u^\top \bar{\Sigma}_{t,\ell-1} u \le 1 + \|u\|_2^2 \lambda_1(\bar{\Sigma}_{t,\ell-1}) \le 1 + \sigma^{-2}\sigma_\ell^2$. Here we use that $\|u\|_2^2 \le \sigma^{-2}$, which can also be proven by induction, and that $\lambda_1(\bar{\Sigma}_{t,\ell-1}) \le \sigma_\ell^2$, which follows from the expression of $\bar{\Sigma}_{t,\ell-1}$ in Section B.1.2. Therefore, we have that

$$\begin{aligned}
\bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1} &\succeq \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + u^\top \bar{\Sigma}_{t,\ell-1} u} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \,, \\
&\succeq \frac{\sigma^{-2}\sigma_{\text{MAX}}^{-2(\ell-1)}}{1 + \sigma^{-2}\sigma_\ell^2} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \,, \\
&\succeq \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \,,
\end{aligned}$$

where the last inequality follows from the definition of $\sigma_{\text{MAX}}^2 = \max_{\ell \in [L+1]} 1 + \sigma^{-2}\sigma_\ell^2$. This concludes the proof.

B.5.4 Proof of theorem 7

We start with the following standard result which we borrow from (Hong et al., 2022a; Aouali et al., 2023b),

$$\mathcal{BR}(T) \le \sqrt{2T\log(1/\delta)} \sqrt{\mathbb{E}\left[\sum_{t=1}^T \|X_t\|_{\check{\Sigma}_{t,A_t}}^2\right]} + cT\delta \,, \qquad \text{where } c > 0 \text{ is a constant} \,. \tag{B.19}$$

Then we use Lemma 5 and express the marginal covariance $\check{\Sigma}_{t,A_t}$ as

$$\check{\Sigma}_{t,a} = \hat{\Sigma}_{t,a} + \sum_{\ell \in [L]} \mathrm{P}_{a,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{a,\ell}^\top \,, \qquad \text{where } \mathrm{P}_{a,\ell} = \hat{\Sigma}_{t,a} \Sigma_1^{-1} \mathrm{W}_1 \prod_{i=1}^{\ell-1} \bar{\Sigma}_{t,i} \Sigma_{i+1}^{-1} \mathrm{W}_{i+1}. \tag{B.20}$$

Therefore, we can decompose $\|X_t\|^2_{\check{\Sigma}_{t,A_t}}$ as

$$\|X_t\|^2_{\check{\Sigma}_{t,A_t}} = \sigma^2 \frac{X_t^\top \check{\Sigma}_{t,A_t} X_t}{\sigma^2} \overset{(i)}{=} \sigma^2 \big(\sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t + \sigma^{-2} \sum_{\ell \in [L]} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t\big)\,,$$

$$\overset{(ii)}{\leq} c_0 \log(1 + \sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t) + \sum_{\ell \in [L]} c_\ell \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t)\,, \tag{B.21}$$

where (i) follows from Equation (B.20), and we use the following inequality in (ii)

$$x = \frac{x}{\log(1+x)} \log(1+x) \leq \left(\max_{x \in [0,u]} \frac{x}{\log(1+x)} \right) \log(1+x) = \frac{u}{\log(1+u)} \log(1+x)\,,$$

which holds for any $x \in [0, u]$, where constants c_0 and c_ℓ are derived as

$$c_0 = \frac{\sigma_1^2}{\log(1 + \frac{\sigma_1^2}{\sigma^2})}\,, \quad c_\ell = \frac{\sigma_{\ell+1}^2}{\log(1 + \frac{\sigma_{\ell+1}^2}{\sigma^2})}\,.$$

The derivation of c_0 uses that

$$X_t^\top \hat{\Sigma}_{t,A_t} X_t \leq \lambda_1(\hat{\Sigma}_{t,A_t}) \|X_t\|^2 \leq \lambda_d^{-1}(\Sigma_1^{-1} + G_{t,A_t}) \leq \lambda_d^{-1}(\Sigma_1^{-1}) = \lambda_1(\Sigma_1) = \sigma_1^2\,.$$

The derivation of c_ℓ follows from

$$X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t \leq \lambda_1(\mathrm{P}_{A_t,\ell} \mathrm{P}_{A_t,\ell}^\top) \lambda_1(\bar{\Sigma}_{t,\ell}) \|X_t\|^2 \leq \sigma_{\ell+1}^2\,.$$

Therefore, from Equation (B.21) and Equation (B.19), we get that

$$\mathcal{BR}(T) \leq \sqrt{2T \log(1/\delta)} \Big(\mathbb{E}\Big[c_0 \sum_{t=1}^T \log(1 + \sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t) + \sum_{\ell \in [L]} c_\ell \sum_{t=1}^T \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) \Big] \Big)^{\frac{1}{2}} + cT\delta \tag{B.22}$$

Now we focus on bounding the logarithmic terms in Equation (B.22).

(I) First term in Equation (B.22) We first rewrite this term as

$$\begin{aligned}
\log(1 + \sigma^{-2} X_t^\top \hat{\Sigma}_{t,A_t} X_t) &\overset{(i)}{=} \log\det(I_d + \sigma^{-2} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \hat{\Sigma}_{t,A_t}^{\frac{1}{2}})\,, \\
&= \log\det(\hat{\Sigma}_{t,A_t}^{-1} + \sigma^{-2} X_t X_t^\top) - \log\det(\hat{\Sigma}_{t,A_t}^{-1}) = \log\det(\hat{\Sigma}_{t+1,A_t}^{-1}) - \log\det(\hat{\Sigma}_{t,A_t}^{-1})\,,
\end{aligned}$$

where (i) follows from the Weinstein-Aronszajn identity. Then we sum over all rounds $t \in [T]$, and get a telescoping

$$\begin{aligned}
\sum_{t=1}^T \log\det(I_d + \sigma^{-2} \hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \hat{\Sigma}_{t,A_t}^{\frac{1}{2}}) &= \sum_{t=1}^T \log\det(\hat{\Sigma}_{t+1,A_t}^{-1}) - \log\det(\hat{\Sigma}_{t,A_t}^{-1})\,, \\
&= \sum_{t=1}^T \sum_{a=1}^K \log\det(\hat{\Sigma}_{t+1,a}^{-1}) - \log\det(\hat{\Sigma}_{t,a}^{-1}) = \sum_{a=1}^K \sum_{t=1}^T \log\det(\hat{\Sigma}_{t+1,a}^{-1}) - \log\det(\hat{\Sigma}_{t,a}^{-1})\,, \\
&= \sum_{a=1}^K \log\det(\hat{\Sigma}_{T+1,a}^{-1}) - \log\det(\hat{\Sigma}_{1,a}^{-1}) \overset{(i)}{=} \sum_{a=1}^K \log\det(\Sigma_1^{\frac{1}{2}} \hat{\Sigma}_{T+1,a}^{-1} \Sigma_1^{\frac{1}{2}})\,,
\end{aligned}$$

where (i) follows from the fact that $\hat{\Sigma}_{1,a} = \Sigma_1$. Now we use the inequality of arithmetic and geometric means and get

$$\begin{aligned}\sum_{t=1}^{T} \log\det(I_d + \sigma^{-2}\hat{\Sigma}_{t,A_t}^{\frac{1}{2}} X_t X_t^\top \hat{\Sigma}_{t,A_t}^{\frac{1}{2}}) &= \sum_{a=1}^{K} \log\det(\Sigma_1^{\frac{1}{2}} \hat{\Sigma}_{T+1,a}^{-1} \Sigma_1^{\frac{1}{2}}) \,, \\ &\leq \sum_{a=1}^{K} d\log\left(\frac{1}{d}\operatorname{Tr}(\Sigma_1^{\frac{1}{2}} \hat{\Sigma}_{T+1,a}^{-1} \Sigma_1^{\frac{1}{2}})\right) \,, \qquad \text{(B.23)} \\ &\leq \sum_{a=1}^{K} d\log\left(1 + \frac{T}{d}\frac{\sigma_1^2}{\sigma^2}\right) = Kd\log\left(1 + \frac{T}{d}\frac{\sigma_1^2}{\sigma^2}\right) \,.\end{aligned}$$

(II) Remaining terms in Equation (B.22) Let $\ell \in [L]$. Then we have that

$$\begin{aligned}\log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) &= \sigma_{\text{MAX}}^{2\ell}\sigma_{\text{MAX}}^{-2\ell} \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) \,, \\ &\leq \sigma_{\text{MAX}}^{2\ell} \log(1 + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) \,, \\ &\overset{(i)}{=} \sigma_{\text{MAX}}^{2\ell} \log\det(I_d + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell} \bar{\Sigma}_{t,\ell}^{\frac{1}{2}} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell}^{\frac{1}{2}}) \,, \\ &= \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{t,\ell}^{-1} + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell}) - \log\det(\bar{\Sigma}_{t,\ell}^{-1})\Big) \,,\end{aligned}$$

where we use the Weinstein-Aronszajn identity in (i). Now we know from Lemma 6 that the following inequality holds $\sigma^{-2}\sigma_{\text{MAX}}^{-2\ell} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \preceq \bar{\Sigma}_{t+1,\ell}^{-1} - \bar{\Sigma}_{t,\ell}^{-1}$. As a result, we get that $\bar{\Sigma}_{t,\ell}^{-1} + \sigma^{-2}\sigma_{\text{MAX}}^{-2\ell} \mathrm{P}_{A_t,\ell}^\top X_t X_t^\top \mathrm{P}_{A_t,\ell} \preceq \bar{\Sigma}_{t+1,\ell}^{-1}$. Thus,

$$\log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) \leq \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{t+1,\ell}^{-1}) - \log\det(\bar{\Sigma}_{t,\ell}^{-1})\Big) \,,$$

Then we sum over all rounds $t \in [T]$, and get a telescoping

$$\begin{aligned}\sum_{t=1}^{T} \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) &\leq \sigma_{\text{MAX}}^{2\ell} \sum_{t=1}^{T} \log\det(\bar{\Sigma}_{t+1,\ell}^{-1}) - \log\det(\bar{\Sigma}_{t,\ell}^{-1}) \,, \\ &= \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{T+1,\ell}^{-1}) - \log\det(\bar{\Sigma}_{1,\ell}^{-1})\Big) \,, \\ &\overset{(i)}{=} \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\bar{\Sigma}_{T+1,\ell}^{-1}) - \log\det(\Sigma_{\ell+1}^{-1})\Big) \,, \\ &= \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}} \bar{\Sigma}_{T+1,\ell}^{-1} \Sigma_{\ell+1}^{\frac{1}{2}})\Big) \,,\end{aligned}$$

where we use that $\bar{\Sigma}_{1,\ell} = \Sigma_{\ell+1}$ in (i). Finally, we use the inequality of arithmetic and geometric means and get that

$$\begin{aligned}\sum_{t=1}^{T} \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell} \bar{\Sigma}_{t,\ell} \mathrm{P}_{A_t,\ell}^\top X_t) &\leq \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}} \bar{\Sigma}_{T+1,\ell}^{-1} \Sigma_{\ell+1}^{\frac{1}{2}})\Big) \,, \\ &\leq d\sigma_{\text{MAX}}^{2\ell} \log\left(\frac{1}{d}\operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}} \bar{\Sigma}_{T+1,\ell}^{-1} \Sigma_{\ell+1}^{\frac{1}{2}})\right) \,, \qquad \text{(B.24)} \\ &\leq d\sigma_{\text{MAX}}^{2\ell} \log\left(1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\right) \,,\end{aligned}$$

The last inequality follows from the expression of $\bar{\Sigma}_{T+1,\ell}^{-1}$ in Equation (B.5) that leads to

$$\begin{aligned}\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}} &= I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\bar{G}_{T+1,\ell}\Sigma_{\ell+1}^{\frac{1}{2}}\,,\\ &= I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{T+1,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}}\,, \qquad \text{(B.25)}\end{aligned}$$

since $\bar{G}_{T+1,\ell} = \mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{T+1,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell$. This allows us to bound $\frac{1}{d}\operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})$ as

$$\begin{aligned}\frac{1}{d}\operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}}) &= \frac{1}{d}\operatorname{Tr}(I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{T+1,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}})\,,\\ &= \frac{1}{d}(d + \operatorname{Tr}(\Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{T+1,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}})\,,\\ &\le 1 + \frac{1}{d}\sum_{i=1}^{d}\lambda_1(\Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{T+1,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}}\,,\\ &\le 1 + \frac{1}{d}\sum_{i=1}^{d}\lambda_1(\Sigma_{\ell+1})\lambda_1(\mathrm{W}_\ell^\top\mathrm{W}_\ell)\lambda_1\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{T+1,\ell-1}\Sigma_\ell^{-1}\big)\,,\\ &\le 1 + \frac{1}{d}\sum_{i=1}^{d}\lambda_1(\Sigma_{\ell+1})\lambda_1(\mathrm{W}_\ell^\top\mathrm{W}_\ell)\lambda_1\big(\Sigma_\ell^{-1}\big)\,,\\ &\le 1 + \frac{1}{d}\sum_{i=1}^{d}\frac{\sigma_{\ell+1}^2}{\sigma_\ell^2} = 1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\,, \qquad \text{(B.26)}\end{aligned}$$

where we use the assumption that $\lambda_1(\mathrm{W}_\ell^\top\mathrm{W}_\ell) = 1$ **(A4)** and that $\lambda_1(\Sigma_{\ell+1}) = \sigma_{\ell+1}^2$ and $\lambda_1(\Sigma_\ell^{-1}) = 1/\sigma_\ell^2$. This is because $\Sigma_\ell = \sigma_\ell^2 I_d$ for any $\ell \in [L+1]$. Finally, plugging Equations (B.23) and (B.24) in Equation (B.22) concludes the proof.

B.5.5 Proof of proposition 7

We use exactly the same proof in Section B.5.4, with one change to account for the sparsity assumption **(A5)**. The change corresponds to Equation (B.24). First, recall that Equation (B.24) writes

$$\sum_{t=1}^{T}\log(1 + \sigma^{-2}X_t^\top\mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t) \le \sigma_{\text{MAX}}^{2\ell}\Big(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})\Big)\,,$$

where

$$\begin{aligned}\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}} &= I_d + \Sigma_{\ell+1}^{\frac{1}{2}}\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\Sigma_{\ell+1}^{\frac{1}{2}}\,,\\ &= I_d + \sigma_{\ell+1}^2\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell\,, \qquad \text{(B.27)}\end{aligned}$$

where the second equality follows from the assumption that $\Sigma_{\ell+1} = \sigma_{\ell+1}^2 I_d$. But notice that in our assumption, **(A5)**, we assume that $\mathrm{W}_\ell = (\bar{\mathrm{W}}_\ell, 0_{d,d-d_\ell})$, where $\bar{\mathrm{W}}_\ell \in \mathbb{R}^{d\times d_\ell}$ for any $\ell \in [L]$. Therefore, we have that for any $d\times d$ matrix $\mathrm{B} \in \mathbb{R}^{d\times d}$, the following holds, $\mathrm{W}_\ell^\top\mathrm{B}\mathrm{W}_\ell = \begin{pmatrix}\bar{\mathrm{W}}_\ell^\top\mathrm{B}\bar{\mathrm{W}}_\ell & 0_{d_\ell,d-d_\ell}\\ 0_{d-d_\ell,d_\ell} & 0_{d-d_\ell,d-d_\ell}\end{pmatrix}$. In particular, we have that

$$\mathrm{W}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\mathrm{W}_\ell = \begin{pmatrix}\bar{\mathrm{W}}_\ell^\top\big(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\big)\bar{\mathrm{W}}_\ell & 0_{d_\ell,d-d_\ell}\\ 0_{d-d_\ell,d_\ell} & 0_{d-d_\ell,d-d_\ell}\end{pmatrix}. \qquad \text{(B.28)}$$

Therefore, plugging this in Equation (B.27) yields that

$$\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}} = \begin{pmatrix} I_{d_\ell} + \sigma_{\ell+1}^2 \bar{\mathrm{W}}_\ell^\top \left(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\right)\bar{\mathrm{W}}_\ell & 0_{d_\ell,d-d_\ell} \\ 0_{d-d_\ell,d_\ell} & I_{d-d_\ell} \end{pmatrix}. \tag{B.29}$$

As a result, $\det(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}}) = \det(I_{d_\ell} + \sigma_{\ell+1}^2 \bar{\mathrm{W}}_\ell^\top \left(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\right)\bar{\mathrm{W}}_\ell)$. This allows us to move the problem from a d-dimensional one to a d_ℓ-dimensional one. Then we use the inequality of arithmetic and geometric means and get that

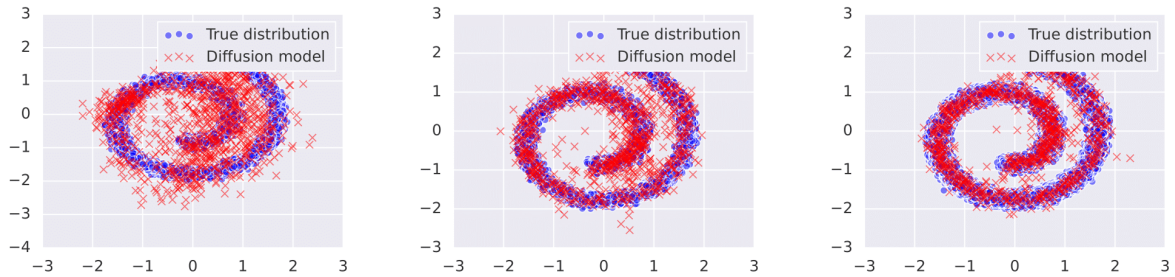
$$\begin{aligned} \sum_{t=1}^{T} \log(1 + \sigma^{-2} X_t^\top \mathrm{P}_{A_t,\ell}\bar{\Sigma}_{t,\ell}\mathrm{P}_{A_t,\ell}^\top X_t) &\leq \sigma_{\text{MAX}}^{2\ell}\left(\log\det(\Sigma_{\ell+1}^{\frac{1}{2}}\bar{\Sigma}_{T+1,\ell}^{-1}\Sigma_{\ell+1}^{\frac{1}{2}})\right), \\ &= \sigma_{\text{MAX}}^{2\ell}\log\det(I_{d_\ell} + \sigma_{\ell+1}^2 \bar{\mathrm{W}}_\ell^\top \left(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\right)\bar{\mathrm{W}}_\ell), \\ &\leq d_\ell\sigma_{\text{MAX}}^{2\ell}\log\left(\frac{1}{d_\ell}\mathrm{Tr}(I_{d_\ell} + \sigma_{\ell+1}^2 \bar{\mathrm{W}}_\ell^\top \left(\Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}\right)\bar{\mathrm{W}}_\ell)\right), \\ &\leq d_\ell\sigma_{\text{MAX}}^{2\ell}\log\left(1 + \frac{\sigma_{\ell+1}^2}{\sigma_\ell^2}\right). \end{aligned} \tag{B.30}$$

To get the last inequality, we use derivations similar to the ones we used in Equation (B.26). Finally, the desired result in obtained by replacing Equation (B.24) by Equation (B.30) in the previous proof in Section B.5.4.

B.6 Additional Experiments

B.6.1 Swiss roll data

Figure B.1 shows samples from the Swiss roll data and samples from generated by the pre-trained diffusion model for different pre-training sample sizes.



(a) Diffusion pre-trained on 50 samples from the Swiss roll dataset.

(b) Diffusion pre-trained on 10^3 samples from the Swiss roll dataset.

(c) Diffusion pre-trained on 10^4 samples from the Swiss roll dataset.

Figure B.1: True distribution of action parameters (blue) vs. distribution of pre-trained diffusion model (red).

B.6.2 Diffusion models pre-training

We used JAX for diffusion model pre-training, summarized as follows:

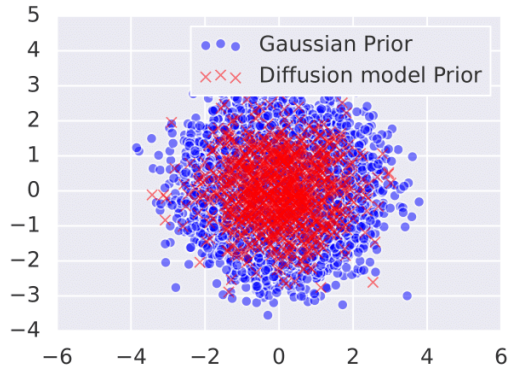
- **Parameterization:** Functions f_ℓ are parameterized with a fully connected 2-layer neural network (NN) with ReLU activation. The step ℓ is provided as input to capture the current sampling stage. Covariances are fixed (not learned) as $\Sigma_\ell = \sigma_\ell^2 I_d$ with σ_ℓ increasing with ℓ.
- **Loss:** Offline data samples are progressively noised over steps $\ell \in [L]$, creating increasingly noisy versions of the data following a predefined noise schedule (Ho et al., 2020). The NN is trained to reverse this noise (i.e., denoise) by predicting the noise added at each step. The loss function measures the L_2 norm difference between the predicted and actual noise at each step, as explained in Ho et al. (2020).
- **Optimization:** Adam optimizer with a 10^{-3} learning rate was used. The NN was trained for 20,000 epochs with a batch size of min(2048, pre-training sample size). We used CPUs for pre-training, which was efficient enough to conduct multiple ablation studies.
- **After pre-training:** The pre-trained diffusion model is used as a prior for `sDM` and compared to `LinTS` as the reference baseline. In our ablation study, we plot the cumulative regret of `LinTS` in the last round divided by that of `sDM`. A ratio greater than 1 indicates that `sDM` outperforms `LinTS`, with higher values representing a larger performance gap.

B.6.3 Quality of our posterior approximation

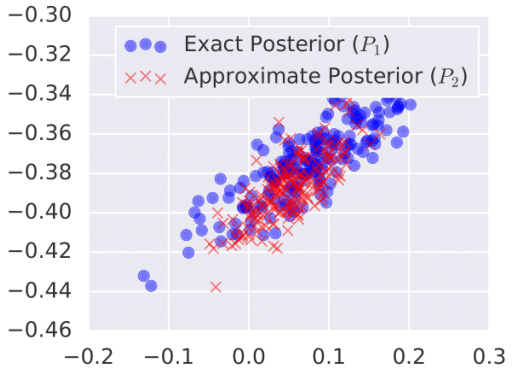
To assess the quality of our posterior approximation, we consider the scenario where the true distribution of action parameters is $\mathcal{N}(0_d, I_d)$ with $d = 2$ and rewards are linear. We pre-train a diffusion model using samples drawn from $\mathcal{N}(0_d, I_d)$. We then consider two priors: the true prior $\mathcal{N}(0_d, I_d)$ and the pre-trained diffusion model prior. This yields two posteriors:

- P_1 : Uses $\mathcal{N}(0_d, I_d)$ as the prior. P_1 is an exact posterior since the prior is Gaussian and rewards are linear-Gaussian.
- P_2 : Uses the pre-trained diffusion model as the prior. P_2 is our approximate posterior.

The learned diffusion model prior matches the true Gaussian prior (as seen in Figure B.2a). Thus, if our approximation is accurate, their posteriors P_1 and P_2 should also be similar. This is observed in Figure B.2b where the approximate posterior P_2 nearly matches the exact posterior P_1.



(a) Gaussian distribution vs. diffusion model pre-trained on 10^3 samples drawn from it.



(b) Exact posterior P_1 vs. approximate posterior P_2 after $T = 100$ rounds of interactions.

Figure B.2: Assessing the quality of our posterior approximation.

B.6.4 CIFAR Ablation

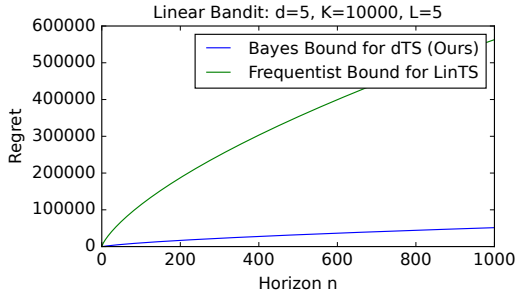
CIFAR. In Figure 4.3a in Section 4.4.2, we showed that with only 10 pre-training samples, `sDM` outperforms `LinTS` on the Swiss-roll benchmark. We now extend this analysis to the vision dataset CIFAR (Krizhevsky et al., 2009) (similar results were obtained on MNIST (Zhu, 2018)). Our setting is similar to that in Hong et al. (2022a) and we use `sDM`'s variant that uses a single shared parameter $\theta \in \mathbb{R}^d$ (Remark 5 and Section 4.2.2) because it is more suited for this setting. These additional ablations on CIFAR confirm that `sDM` consistently benefits from offline pre-training, even when the true prior is not a diffusion model. Specifically, we vary the percentage of offline data used to train the prior and compare against both `HierTS` and `LinTS`.

Table B.1: Regret improvement (%) of `sDM` on CIFAR.

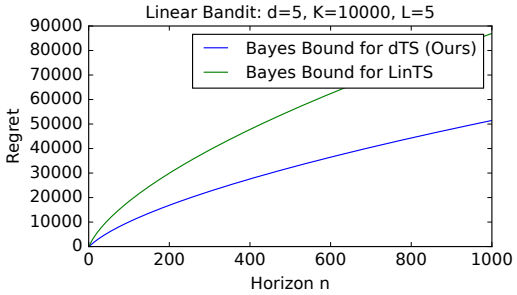
Offline Data (%)	**vs. `HierTS`**	**vs. `LinTS`**
1%	69.11%	87.74%
5%	79.56%	92.18%
25%	80.65%	92.48%
50%	81.67%	92.88%

B.6.5 Bound comparison

Here, we compare our bound in Theorem 7 to bounds of `LinTS` from the literature.



(a) Our bound vs. the frequentist bound of `LinTS` in Abeille and Lazaric (2017).



(b) Our bound vs. the standard Bayesian bound of `LinTS`.

Figure B.3: Comparing our Bayesian regret bound of `dTS` to the frequentist and Bayesian bounds of `LinTS`.

Chapter C

Supplementary Materials for Chapter 6

Contents

C.1 Posterior Derivations Under Standard Priors

Here we derive the posterior under the standard prior in Equation (6.2). These are standard derivations and we present them here for the sake of completeness. But first, we state the following standard assumption that allows posterior derivations.

Assumption 5 (Independence)**.** (X, A) *is independent of* θ*, and the* θ_a*, for* $a \in \mathcal{A}$ *are independent.*

Derivation of $p(\theta_a \mid \mathcal{D}_n)$ *for the standard prior in Equation* (6.2)*.* We start by recalling the standard prior in Equation (6.2)

$$\begin{aligned} \theta_a &\sim \mathcal{N}(\mu_a, \Sigma_a), && \forall a \in \mathcal{A}, \\ R \mid \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2), \end{aligned} \tag{C.1}$$

where $\mathcal{N}(\mu_a, \Sigma_a)$ is the prior on the action parameter θ_a. Let $\theta = (\theta_a)_{a\in\mathcal{A}} \in \mathbb{R}^{dK}$, $\Sigma_\mathcal{A} = \text{diag}(\Sigma_a)_{a\in\mathcal{A}} \in \mathbb{R}^{dK\times dK}$ and $\mu_\mathcal{A} = (\mu_a)_{a\in\mathcal{A}} \in \mathbb{R}^{dK}$. Also, let $u_a \in \{0,1\}^K$ be the binary vector representing the action a. That is, $u_{a,a} = 1$ and $u_{a,a'} = 0$ for all $a' \neq a$. Then we can rewrite the model in Equation (C.1) as

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu_\mathcal{A}, \Sigma_\mathcal{A}), \\ R \mid \theta, X, A &\sim \mathcal{N}((u_A \otimes \phi(X))^\top \theta, \sigma^2). \end{aligned} \tag{C.2}$$

Then the *joint* action posterior $p(\theta \mid \mathcal{D}_n)$ decomposes as

$$\begin{aligned} p(\theta \mid \mathcal{D}_n) &= p(\theta \mid (X_i, A_i, R_i)_{i\in[n]}) \overset{(i)}{\propto} p((R_i)_{i\in[n]} \mid \theta, (X_i, A_i)_{i\in[n]}) p(\theta \mid (X_i, A_i)_{i\in[n]}), \\ &\overset{(ii)}{=} p((R_i)_{i\in[n]} \mid \theta, (X_i, A_i)_{i\in[n]}) p(\theta) \overset{(iii)}{=} \prod_{i\in[n]} p(R_i \mid \theta, X_i, A_i) p(\theta), \\ &\overset{(iv)}{=} \prod_{i\in[n]} \mathcal{N}(R_i; (u_{A_i} \otimes \phi(X_i))^\top \theta, \sigma^2) \mathcal{N}(\theta; \mu_\mathcal{A}, \Sigma_\mathcal{A}), \\ &\overset{(v)}{\propto} \mathcal{N}\left(\theta; \hat{\mu}_\mathcal{A}, \left(\hat{\Lambda}_\mathcal{A}\right)^{-1}\right). \end{aligned}$$

In (i), we apply Bayes rule. In (ii), we use that θ is independent of (X, A), and (iii) follows from the assumption that $R_i \mid \theta, X_i, A_i$ are independent. Finally, in (iv), we replace the distribution by their Gaussian form, and in (v), we set $\hat{\Lambda}_\mathcal{A} = v \sum_{i=1}^n (u_{A_i} u_{A_i}^\top \otimes \phi(X_i)\phi(X_i)^\top) + \Lambda_\mathcal{A}$, and $\hat{\mu}_\mathcal{A} = \hat{\Lambda}_\mathcal{A}^{-1}\left(v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i)) R_i + \Lambda_\mathcal{A}\mu_\mathcal{A}\right)$ where $v = \sigma^{-2}$ and $\Lambda_\mathcal{A} = \Sigma_\mathcal{A}^{-1} = \text{diag}(\Sigma_a^{-1})_{a\in\mathcal{A}}$. Now notice that $\hat{\Lambda}_\mathcal{A} = \text{diag}(\Sigma_a^{-1} + G_a)_{a\in\mathcal{A}}$. Thus, $p(\theta \mid \mathcal{D}_n) = \mathcal{N}(\theta; \hat{\mu}_\mathcal{A}, \hat{\Lambda}_\mathcal{A}^{-1})$ where $\hat{\mu}_\mathcal{A} = (\hat{\mu}_a)_{a\in\mathcal{A}}$ and $\hat{\Lambda}_\mathcal{A} = \text{diag}(\hat{\Lambda}_a)_{a\in\mathcal{A}}$, with

$$\hat{\Lambda}_a = \Sigma_a^{-1} + G_a, \qquad \hat{\Lambda}_a \hat{\mu}_a = \Sigma_a^{-1}\mu_a + B_a.$$

Since the covariance matrix of $p(\theta \mid \mathcal{D}_n)$ is diagonal by block, we know that the marginals $\theta_a \mid \mathcal{D}_n$ also have a Gaussian density $p(\theta_a \mid \mathcal{D}_n) = \mathcal{N}(\theta_a; \hat{\mu}_a, \hat{\Sigma}_a)$ where $\hat{\Sigma}_a = \hat{\Lambda}_a^{-1}$.

□

C.2 Posterior Derivations Under Structured Priors

Here we derive the posteriors under the structured prior in Equation (6.9). Precisely, we derive the latent posterior density of $\psi \mid \mathcal{D}_n$, the conditional posterior density of $\theta \mid \mathcal{D}_n, \psi$. Then, we derive the marginal posterior $\theta \mid \mathcal{D}_n$. Posterior derivations rely on the following assumption.

Assumption 6 (Structured Independence)**.** *(i)* (X, A) *is independent of* ψ *and given* ψ*,* (X, A) *is independent of* θ*. (ii) Given* ψ*, the* θ_a*, for all* $a \in \mathcal{A}$ *are independent.*

C.2.1 Latent Posterior

Derivation of $p(\psi \mid \mathcal{D}_n)$*.* First, recall that our model in Equation (6.9) reads

$$\begin{aligned} \psi &\sim \mathcal{N}(\mu, \Sigma), \\ \theta_a \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_a\psi, \Sigma_a\Big), && \forall a \in \mathcal{A}, \\ R \mid \psi, \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2). \end{aligned} \tag{C.3}$$

Then we first rewrite it as

$$\begin{aligned} \psi &\sim \mathcal{N}(\mu, \Sigma), \\ \theta \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_{\mathcal{A}}\psi, \Sigma_{\mathcal{A}}\Big), \\ R \mid \psi, \theta, X, A &\sim \mathcal{N}((u_A \otimes \phi(X))^\top \theta, \sigma^2). \end{aligned} \tag{C.4}$$

Then the latent posterior is

$$\begin{aligned} p(\psi \mid (X_i, A_i, R_i)_{i\in[n]}) &\propto p((R_i)_{i\in[n]} \mid \psi, (X_i, A_i)_{i\in[n]}) p(\psi \mid (X_i, A_i)_{i\in[n]}), \\ &\overset{(i)}{=} p((R_i)_{i\in[n]} \mid \psi, (X_i, A_i)_{i\in[n]}) q(\psi), \\ &= \int_\theta p((R_i)_{i\in[n]}, \theta \mid \psi, (X_i, A_i)_{i\in[n]}) \,\mathrm{d}\theta q(\psi), \\ &= \int_\theta p((R_i)_{i\in[n]} \mid \psi, \theta, (X_i, A_i)_{i\in[n]}) p(\theta \mid \psi, (X_i, A_i)_{i\in[n]}) \,\mathrm{d}\theta q(\psi), \\ &\overset{(ii)}{=} \int_\theta p((R_i)_{i\in[n]} \mid \psi, \theta, (X_i, A_i)_{i\in[n]}) p(\theta \mid \psi) \,\mathrm{d}\theta q(\psi), \end{aligned}$$

In (i), we use that (X, A) is independent of ψ, which follows from Assumption 6. Similarly, in (ii), we use that θ is conditionally independent of (X, A) given ψ. Now we know that given θ, $R_i \mid X_i, A_i$ are i.i.d. and hence $p((R_i)_{i\in[n]} \mid \psi, \theta, (X_i, A_i)_{i\in[n]}) = \prod_{a\in\mathcal{A}} \mathcal{L}_a(\theta_a)$. Moreover, θ_a for $a \in \mathcal{A}$ are conditionally independent given ψ. Thus $p(\theta \mid \psi) = \prod_{a\in\mathcal{A}} p_a\left(\theta_a; f_a(\psi)\right)$, where we also used that $\theta_a \mid \psi \sim p_a\left(\cdot; f_a(\psi)\right)$. This leads to

$$\begin{aligned} p(\psi \mid (X_i, A_i, R_i)_{i\in[n]}) &\propto \int_\theta \prod_{a\in\mathcal{A}} \mathcal{L}_a(\theta_a) p_a\left(\theta_a; f_a(\psi)\right) \mathrm{d}\theta q(\psi), \\ &\overset{(i)}{=} \prod_{a\in\mathcal{A}} \int_{\theta_a} \mathcal{L}_a(\theta_a) \mathcal{N}\left(\theta_a; \mathrm{W}_a\psi, \Sigma_a\right) \mathrm{d}\theta_a \mathcal{N}(\psi; \mu, \Sigma), \\ &\overset{(ii)}{=} \prod_{a\in\mathcal{A}} \int_{\theta_a} \Big(\prod_{i\in I_a} \mathcal{N}(R_i; \phi(X_i)^\top \theta_a, \sigma^2) \Big) \mathcal{N}\left(\theta_a; \mathrm{W}_a\psi, \Sigma_a\right) \mathrm{d}\theta_a \mathcal{N}(\psi; \mu, \Sigma). \end{aligned}$$

In (i), we notice that $\theta = (\theta_a)_{a\in\mathcal{A}}$ and apply Fubini's Theorem. In (ii), we let $I_a = \{i \in [n]; A_i = a\}$ as the rounds where action a appears in the sample set $\mathcal{D}_n$. Now let $h_a(\psi) = \int_{\theta_a} \left(\prod_{i\in I_a} \mathcal{N}(R_i; \phi(X_i)^\top \theta_a, \sigma^2)\right) \mathcal{N}\left(\theta_a; \mathrm{W}_a\psi, \Sigma_a\right) \mathrm{d}\theta_a$. Then we have that

$$p(\psi \mid \mathcal{D}_n) \propto \prod_{a\in\mathcal{A}} h_a(\psi)\mathcal{N}(\psi; \mu, \Sigma). \tag{C.5}$$

We start by computing h_a. To reduce clutter, let $v = \sigma^{-2}$ and $\Lambda_a = \Sigma_a^{-1}$. Then we compute h_a as

$$\begin{aligned}
h_a(\psi) &= \int_{\theta_a} \left(\prod_{i\in I_a} \mathcal{N}(R_i; \phi(X_i)^\top \theta_a, \sigma^2)\right) \mathcal{N}(\theta_a; \mathrm{W}_a\psi, \Sigma_a)\, \mathrm{d}\theta_a, \\
&\propto \int_{\theta_a} \exp\left[-\frac{1}{2} v \sum_{i\in I_a} (R_i - \phi(X_i)^\top \theta_a)^2 - \frac{1}{2}(\theta_a - \mathrm{W}_a\psi)^\top \Lambda_a (\theta_a - \mathrm{W}_a\psi)\right] \mathrm{d}\theta_a, \\
&= \int_{\theta_a} \exp\Big[-\frac{1}{2}\Big(v \sum_{i\in I_a} (R_i^2 - 2R_i\theta_a^\top \phi(X_i) + (\theta_a^\top \phi(X_i))^2) + \theta_a^\top \Lambda_a \theta_a - 2\theta_a^\top \Lambda_a \mathrm{W}_a\psi \\
&\qquad\qquad + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi)\Big)\Big] \mathrm{d}\theta_a, \\
&\propto \int_{\theta_a} \exp\Big[-\frac{1}{2}\Big(\theta_a^\top \left(v \sum_{i\in I_a} \phi(X_i)\phi(X_i)^\top + \Lambda_a\right) \theta_a - 2\theta_a^\top \left(v \sum_{i\in I_a} R_i \phi(X_i) + \Lambda_a \mathrm{W}_a\psi\right) \\
&\qquad\qquad + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi)\Big)\Big] \mathrm{d}\theta_a.
\end{aligned}$$

Now recall that $G_a = v \sum_{i\in I_a} \phi(X_i)\phi(X_i)^\top$ and $B_a = v \sum_{i\in I_a} R_i \phi(X_i)$ and let $V_a = (G_a + \Lambda_a)^{-1}$, $U_a = V_a^{-1}$, and $\beta_a = V_a(B_a + \Lambda_a \mathrm{W}_a\psi)$. Then have that $U_aV_a = V_aU_a = I_d$, and thus

$$\begin{aligned}
h_a(\psi) &\propto \int_{\theta_a} \exp\left[-\frac{1}{2}\left(\theta_a^\top U_a \theta_a - 2\theta_a^\top U_a V_a \left(B_a + \Lambda_a \mathrm{W}_a\psi\right) + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi)\right)\right] \mathrm{d}\theta_a, \\
&= \int_{\theta_a} \exp\left[-\frac{1}{2}\left(\theta_a^\top U_a \theta_a - 2\theta_a^\top U_a \beta_a + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi)\right)\right] \mathrm{d}\theta_a, \\
&= \int_{\theta_a} \exp\left[-\frac{1}{2}\left((\theta_a - \beta_a)^\top U_a (\theta_a - \beta_a) - \beta_a^\top U_a \beta_a + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi)\right)\right] \mathrm{d}\theta_a, \\
&\propto \exp\left[-\frac{1}{2}\left(-\beta_a^\top U_a \beta_a + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi)\right)\right], \\
&= \exp\left[-\frac{1}{2}\left(-\left(B_a + \Lambda_a \mathrm{W}_a\psi\right)^\top V_a \left(B_a + \Lambda_a \mathrm{W}_a\psi\right) + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi)\right)\right], \\
&\propto \exp\left[-\frac{1}{2}\left(\psi^\top \mathrm{W}_a^\top \left(\Lambda_a - \Lambda_a V_a \Lambda_a\right) \mathrm{W}_a\psi - 2\psi^\top \left(\mathrm{W}_a^\top \Lambda_a V_a B_a\right)\right)\right], \\
&\propto \exp\left[-\frac{1}{2}\left(\psi^\top \bar{\Lambda}_a \psi - 2\psi^\top \bar{\Lambda}_a \bar{\mu}_a\right)\right],
\end{aligned}$$

where

$$\begin{aligned}
\bar{\Lambda}_a &= \mathrm{W}_a^\top \left(\Lambda_a - \Lambda_a V_a \Lambda_a\right) \mathrm{W}_a = \mathrm{W}_a^\top \left(\Sigma_a^{-1} - \Sigma_a^{-1}(G_a + \Sigma_a^{-1})^{-1}\Sigma_a^{-1}\right) \mathrm{W}_a, \\
\bar{\Lambda}_a \bar{\mu}_a &= \mathrm{W}_a^\top \Lambda_a V_a B_a = \mathrm{W}_a^\top \Sigma_a^{-1} (G_a + \Sigma_a^{-1})^{-1} B_a.
\end{aligned} \tag{C.6}$$

However, we know from Equation (C.5) that $p(\psi \mid \mathcal{D}_n) \propto \prod_{a\in\mathcal{A}} h_a(\psi)\mathcal{N}(\psi;\mu,\Sigma)$. But $h_a(\psi)$ is proportional to $\exp\left[-\frac{1}{2}\left(\psi^\top \bar{\Lambda}_a\psi - 2\psi^\top\bar{\Lambda}_a\bar{\mu}_a\right)\right]$ for any a. Thus $p(\psi \mid \mathcal{D}_n)$ can be seen as the product of $K+1$ Gaussian kernels. Thus, $p(\psi \mid \mathcal{D}_n)$ is a multivariate Gaussian distribution $\mathcal{N}(\bar{\mu},\bar{\Sigma})$, with

$$\bar{\Sigma}^{-1} = \Sigma^{-1} + \sum_{a\in\mathcal{A}} \bar{\Lambda}_a = \Sigma^{-1} + \sum_{a\in\mathcal{A}} \mathrm{W}_a^\top\left(\Sigma_a^{-1} - \Sigma_a^{-1}(G_a+\Sigma_a^{-1})^{-1}\Sigma_a^{-1}\right)\mathrm{W}_a, \tag{C.7}$$

$$\bar{\Sigma}^{-1}\bar{\mu} = \Sigma^{-1}\mu + \sum_{a\in\mathcal{A}} \bar{\Lambda}_a\bar{\mu}_a = \Sigma^{-1}\mu + \sum_{a\in\mathcal{A}} \mathrm{W}_a^\top\Sigma_a^{-1}(G_a+\Sigma_a^{-1})^{-1}B_a. \tag{C.8}$$

□

C.2.2 Conditional Posterior

Derivation of $p(\theta_a \mid \psi, \mathcal{D}_n)$. Let $v = \sigma^{-2}$, $\Lambda_a = \Sigma_a^{-1}$. We consider the model rewritten in Equation (C.4), then the *joint* conditional action posterior $p(\theta \mid \psi, \mathcal{D}_n)$ decomposes as

$$\begin{aligned}
p(\theta \mid \psi, \mathcal{D}_n) &= p(\theta \mid \psi, (X_i, A_i, R_i)_{i\in[n]}) \overset{(i)}{\propto} p((R_i)_{i\in[n]} \mid \theta, \psi, (X_i, A_i)_{i\in[n]})p(\theta \mid \psi, (X_i, A_i)_{i\in[n]}),\\
&\overset{(ii)}{=} p((R_i)_{i\in[n]} \mid \theta, (X_i, A_i)_{i\in[n]})p(\theta \mid \psi) \overset{(iii)}{=} \prod_{i\in[n]} p(R_i \mid \theta, X_i, A_i)p(\theta \mid \psi),\\
&\overset{(iv)}{=} \prod_{i\in[n]} \mathcal{N}(R_i; (u_{A_i}\otimes\phi(X_i))^\top\theta, \sigma^2)\mathcal{N}(\theta; \mathrm{W}_\mathcal{A}\psi, \Sigma_\mathcal{A}),\\
&= \exp\Big[-\frac{1}{2}\Big(v\sum_{i=1}^n (R_i^2 - 2R_i(u_{A_i}\otimes\phi(X_i))^\top\theta + ((u_{A_i}\otimes\phi(X_i))^\top\theta)^2) + \theta^\top\Lambda_\mathcal{A}\theta - 2\theta^\top\Lambda_\mathcal{A}\mathrm{W}_\mathcal{A}\psi\\
&\qquad\qquad + \psi^\top\mathrm{W}_\mathcal{A}^\top\Lambda_\mathcal{A}\mathrm{W}_\mathcal{A}\psi\Big)\Big],\\
&\propto \exp\Big[-\frac{1}{2}\Big(\theta^\top(v\sum_{i=1}^n (u_{A_i}\otimes\phi(X_i))(u_{A_i}\otimes\phi(X_i))^\top + \Lambda_\mathcal{A})\theta\\
&\qquad\qquad - 2\theta^\top\left(v\sum_{i=1}^n (u_{A_i}\otimes\phi(X_i))R_i + \Lambda_\mathcal{A}\mathrm{W}_\mathcal{A}\psi\right)\Big)\Big],\\
&= \exp\Big[-\frac{1}{2}\Big(\theta^\top(v\sum_{i=1}^n (u_{A_i}u_{A_i}^\top\otimes\phi(X_i)\phi(X_i)^\top) + \Lambda_\mathcal{A})\theta -\\
&\qquad\qquad 2\theta^\top\left(v\sum_{i=1}^n (u_{A_i}\otimes\phi(X_i))R_i + \Lambda_\mathcal{A}\mathrm{W}_\mathcal{A}\psi\right)\Big)\Big],\\
&\overset{(v)}{\propto} \mathcal{N}\left(\theta; \tilde{\mu}_\mathcal{A}, \left(\tilde{\Lambda}_\mathcal{A}\right)^{-1}\right),
\end{aligned}$$

where we use Bayes rule in (i), (ii) uses two assumptions. First, Given θ, X, A, R is independent of ψ. Second, given ψ, θ is independent of (X, A). Moreover, (iii) follows from the assumption that $R_i \mid \theta, X_i, A_i$ are independent. Finally, in iv, we replace the distribution by their Gaussian form, and in (v), we set $\tilde{\Lambda}_\mathcal{A} = v\sum_{i=1}^n (u_{A_i}u_{A_i}^\top\otimes\phi(X_i)\phi(X_i)^\top) + \Lambda_\mathcal{A}$, and $\tilde{\mu}_\mathcal{A} = \tilde{\Lambda}_\mathcal{A}^{-1}\left(v\sum_{i=1}^n (u_{A_i}\otimes\phi(X_i))R_i + \Lambda_\mathcal{A}\mathrm{W}_\mathcal{A}\psi\right)$, where $\Lambda_\mathcal{A} = \Sigma_\mathcal{A}^{-1} = \mathrm{diag}(\Sigma_a^{-1})_{a\in\mathcal{A}}$.

Now notice that $\tilde{\Lambda}_{\mathcal{A}} = \text{diag}(\Sigma_a^{-1} + G_a)_{a\in\mathcal{A}}$. Thus, $p(\theta \mid \psi, \mathcal{D}_n) = \mathcal{N}(\theta; \tilde{\mu}_{\mathcal{A}}, \tilde{\Lambda}_{\mathcal{A}}^{-1})$ where $\tilde{\mu}_{\mathcal{A}} = (\tilde{\mu}_a)_{a\in\mathcal{A}}$ and $\tilde{\Lambda}_{\mathcal{A}} = \text{diag}(\tilde{\Lambda}_a)_{a\in\mathcal{A}}$, with

$$\begin{aligned}\tilde{\Lambda}_a &= \Sigma_a^{-1} + G_a,\\ \tilde{\Lambda}_a\tilde{\mu}_a &= \Sigma_a^{-1}\mathrm{W}_a\psi + B_a.\end{aligned}$$

The covariance matrix of $p(\theta \mid \psi, \mathcal{D}_n)$ is diagonal by block. Thus $\theta_a \mid \psi, \mathcal{D}_n$ for $a \in \mathcal{A}$ are independent and have a Gaussian density $p(\theta_a \mid \psi, \mathcal{D}_n) = \mathcal{N}(\theta_a; \tilde{\mu}_a, \tilde{\Sigma}_a)$ where $\tilde{\Sigma}_a = \tilde{\Lambda}_a^{-1}$. □

C.2.3 Action Posterior

Derivation of $p(\theta_a \mid \mathcal{D}_n)$. We know that $\theta_a \mid \mathcal{D}_n, \psi \sim \mathcal{N}(\tilde{\mu}_a, \tilde{\Sigma}_a)$ and $\psi \mid \mathcal{D}_n \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$. Thus the posterior density of $\theta_a \mid \mathcal{D}_n$ is also Gaussian since Gaussianity is preserved after marginalization (Koller and Friedman, 2009). We let $\theta_a \mid \mathcal{D}_n \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$. Then, we can compute $\hat{\mu}_a$ and $\hat{\Sigma}_a$ using the total expectation and total covariance decompositions. Let $\Lambda_a = \Sigma_a^{-1}$. Then we have that

$$\begin{aligned}\tilde{\Sigma}_a &= (G_a + \Lambda_a)^{-1}\\ \mathbb{E}\left[\theta_a \,|\, \psi, \mathcal{D}_n\right] &= \tilde{\Sigma}_a\left(B_a + \Lambda_a\mathrm{W}_a\psi\right)\end{aligned}$$

First, given $\mathcal{D}_n$, $\tilde{\Sigma}_a = (G_a + \Lambda_a)^{-1}$ and B_a are constant (do not depend on ψ). Thus

$$\begin{aligned}\hat{\mu}_a = \mathbb{E}\left[\theta_a \,|\, \mathcal{D}_n\right] = \mathbb{E}\left[\mathbb{E}\left[\theta_a \,|\, \psi, \mathcal{D}_n\right] \,|\, \mathcal{D}_n\right] &= \mathbb{E}_{\psi\sim\mathcal{N}(\bar{\mu},\bar{\Sigma})}\left[\tilde{\Sigma}_a\left(B_a + \Lambda_a\mathrm{W}_a\psi\right)\right]\\ &= \tilde{\Sigma}_a\left(B_a + \Lambda_a\mathrm{W}_a\mathbb{E}_{\psi\sim\mathcal{N}(\bar{\mu},\bar{\Sigma})}\left[\psi\right]\right),\\ &= \tilde{\Sigma}_a\left(B_a + \Lambda_a\mathrm{W}_a\bar{\mu}\right).\end{aligned}$$

This concludes the computation of $\hat{\mu}_a$. Similarly, given $\mathcal{D}_n$, $\tilde{\Sigma}_a = (G_a + \Lambda_a)^{-1}$ and B_a are constant (do not depend on ψ), yields two things. First,

$$\mathbb{E}\left[\text{cov}\left[\theta_a \,|\, \psi, \mathcal{D}_n\right] \,|\, \mathcal{D}_n\right] = \mathbb{E}\left[\tilde{\Sigma}_a \,\middle|\, \mathcal{D}_n\right] = \tilde{\Sigma}_a.$$

Second,

$$\begin{aligned}\text{cov}\left[\mathbb{E}\left[\theta_a \,|\, \psi, \mathcal{D}_n\right] \,|\, \mathcal{D}_n\right] &= \text{cov}\left[\tilde{\Sigma}_a\Lambda_a\mathrm{W}_a\psi \,\middle|\, \mathcal{D}_n\right]\\ &= \tilde{\Sigma}_a\Lambda_a\mathrm{W}_a\text{cov}\left[\psi \,|\, \mathcal{D}_n\right]\mathrm{W}_a^\top\Lambda_a\tilde{\Sigma}_a\\ &= \tilde{\Sigma}_a\Lambda_a\mathrm{W}_a\bar{\Sigma}\mathrm{W}_a^\top\Lambda_a\tilde{\Sigma}_a.\end{aligned}$$

Finally, the total covariance decomposition (Weiss, 2005) yields that

$$\begin{aligned}\hat{\Sigma}_a = \text{cov}\left[\theta_a \,|\, \mathcal{D}_n\right] &= \mathbb{E}\left[\text{cov}\left[\theta_a \,|\, \psi, \mathcal{D}_n\right] \,|\, \mathcal{D}_n\right] + \text{cov}\left[\mathbb{E}\left[\theta_a \,|\, \psi, \mathcal{D}_n\right] \,|\, \mathcal{D}_n\right]\\ &= \tilde{\Sigma}_a + \tilde{\Sigma}_a\Lambda_a\mathrm{W}_a\bar{\Sigma}\mathrm{W}_a^\top\Lambda_a\tilde{\Sigma}_a.\end{aligned}$$

This concludes the proof.

□

C.3 Proofs

C.3.1 Main Result

In this section, we prove Theorem 2. Recall that we make the following well-specified prior assumption.

Assumption 7 (Well-specified priors)**.** *Action parameters $\theta_{*,a}$ and rewards are drawn from Equation (6.9).*

Assumption 8 (Diagonal covariances for simplicity)**.** *We assume $\Sigma_a = \sigma_0^2 I_d$, $\Sigma = \tau^2 I_{d'}$, $\|\phi(x)\|_2 \leq 1$, and the matrices W_a are normalized such that $\lambda_1(\mathrm{W}_a \mathrm{W}_a^\top) = \lambda_d(\mathrm{W}_a \mathrm{W}_a^\top) = 1$.*

Proof. First, given $x \in \mathcal{X}$, by definition of the optimal policy, we know that it is deterministic. That is, there exists $a_{x,\theta_*} \in [K]$ such that $\pi_*(a_{x,\theta_*} \mid x) = 1$. To simplify the notation and since π_* is deterministic, we let $\pi_*(x) = a_{x,\theta_*}$. Also, we know that the greedy policy is deterministic in $\hat{a}_x = \mathrm{argmax}_{b \in \mathcal{A}} \hat{r}(x, b)$. That is $\hat{\pi}_{\mathrm{G}}(\hat{a}_x \mid x) = 1$. Similarly, we let $\hat{\pi}_{\mathrm{G}}(x) = \hat{a}_x$. Moreover, we let $\Phi(x, a) = e_a \otimes \phi(x) \in \mathbb{R}^{dK}$ where $e_a \in \mathbb{R}^K$ is the indicator vector of action a, such that $e_{a,b} = 0$ for any $b \in \mathcal{A}/\{a\}$ and $e_{a,a} = 1$. Also, recall that $\hat{\mu} = (\hat{\mu}_a)_{a \in \mathcal{A}}$ is the concatenation of the posterior means.

$$
\begin{aligned}
\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) &= \mathbb{E}\left[V(\pi_*; \theta_*) - V(\hat{\pi}_{\mathrm{G}}; \theta_*)\right], \\
&= \mathbb{E}\left[r(X, \pi_*(X); \theta_*) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*)\right], \\
&= \mathbb{E}\left[r(X, \pi_*(X); \theta_*) - r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) + r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*)\right], \\
&\leq \mathbb{E}\left[r(X, \pi_*(X); \theta_*) - r(X, \pi_*(X); \hat{\mu}) + r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*)\right], \\
&\leq \mathbb{E}\left[r(X, \pi_*(X); \theta_*) - r(X, \pi_*(X); \hat{\mu})\right] + \mathbb{E}\left[r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*)\right].
\end{aligned}
$$

Now we start by proving that $\mathbb{E}\left[r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*)\right] = 0$. This is achieved as follows

$$
\begin{aligned}
\mathbb{E}\left[r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*)\right] &= \mathbb{E}\left[\mathbb{E}\left[r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*) \mid X, \mathcal{D}_n\right]\right], \\
&= \mathbb{E}\left[\mathbb{E}\left[\phi(X)^\top \hat{\mu}_{\hat{\pi}_{\mathrm{G}}(X)} - \phi(X)^\top \theta_{*,\hat{\pi}_{\mathrm{G}}(X)} \mid X, \mathcal{D}_n\right]\right], \\
&\overset{(i)}{=} \mathbb{E}\left[\mathbb{E}\left[\Phi(X, \hat{\pi}_{\mathrm{G}}(X))^\top \hat{\mu} - \Phi(X, \hat{\pi}_{\mathrm{G}}(X))^\top \theta_* \mid X, \mathcal{D}_n\right]\right], \\
&\overset{(ii)}{=} \mathbb{E}\left[\Phi(X, \hat{\pi}_{\mathrm{G}}(X))^\top \mathbb{E}\left[\hat{\mu} - \theta_* \mid X, \mathcal{D}_n\right]\right], \\
&\overset{(iii)}{=} \mathbb{E}\left[\Phi(X, \hat{\pi}_{\mathrm{G}}(X))^\top (\hat{\mu} - \mathbb{E}\left[\theta_* \mid X, \mathcal{D}_n\right])\right], \\
&\overset{(iv)}{=} 0.
\end{aligned}
$$

In (i), we used that by definition of $\Phi(x, a) = e_a \otimes \phi(X) \in \mathbb{R}^{dK}$, we have $\phi(x)^\top \hat{\mu}_a = \Phi(x, a)^\top \hat{\mu}$ for any (x, a), and the same holds for θ_*. In (ii), we used that $\Phi(X, \hat{\pi}_{\mathrm{G}}(X))$ is deterministic given X and $\mathcal{D}_n$. In (iii), we used that $\hat{\mu}$ is deterministic given X and $\mathcal{D}_n$. Finally, in (iv), we used that $\mathbb{E}\left[\theta_* \mid X, \mathcal{D}_n\right] = \mathbb{E}\left[\theta_* \mid \mathcal{D}_n\right] = \hat{\mu}$, which follows from the assumption that θ_* does not depend on X and the assumption that θ_* is drawn from the prior, and hence when conditioned on $\mathcal{D}_n$, it is drawn from the posterior whose mean is $\hat{\mu}$. Therefore, $\mathbb{E}\left[r(X, \hat{\pi}_{\mathrm{G}}(X); \hat{\mu}) - r(X, \hat{\pi}_{\mathrm{G}}(X); \theta_*)\right] = 0$ which leads to

$$
\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) \leq \mathbb{E}\left[r(X, \pi_*(X); \theta_*) - r(X, \pi_*(X); \hat{\mu})\right].
$$

Now let $\delta \in (0,1)$ and define the *joint parameter event*

$$\mathcal{E}_\alpha = \Big\{ \forall a \in \mathcal{A} : \ \|\theta_{*,a} - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}} \le \alpha \Big\}.$$

Recall $Z_a = r(X,a;\theta_*) - r(X,a;\hat{\mu}) = \phi(X)^\top(\theta_{*,a} - \hat{\mu}_a)$. By Cauchy-Schwarz, on $\mathcal{E}_\alpha$ we have for all a,

$$|Z_a| = |\phi(X)^\top(\theta_{*,a} - \hat{\mu}_a)| \le \|\phi(X)\|_{\hat{\Sigma}_a} \|\theta_{*,a} - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}} \le \alpha \|\phi(X)\|_{\hat{\Sigma}_a},$$

hence in particular $|Z_{\pi_*(X)}| \le \alpha \|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}$ on $\mathcal{E}_\alpha$. Therefore,

$$\begin{aligned}
\mathrm{BSO}(\hat{\pi}_{\mathrm{G}}) &\le \mathbb{E}\left[|Z_{\pi_*(X)}|\right] \\
&= \mathbb{E}\left[|Z_{\pi_*(X)}| \mathbb{1}\{\mathcal{E}_\alpha\}\right] + \mathbb{E}\left[|Z_{\pi_*(X)}| \mathbb{1}\{\bar{\mathcal{E}}_\alpha\}\right] \\
&\le \alpha \mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] + \mathbb{E}\left[|Z_{\pi_*(X)}| \mathbb{1}\{\bar{\mathcal{E}}_\alpha\}\right].
\end{aligned}$$

We now bound $\mathbb{P}(\bar{\mathcal{E}}_\alpha \mid \mathcal{D}_n)$. Under the well-specified Bayes assumption, conditional on $\mathcal{D}_n$ each marginal posterior satisfies $\theta_{*,a} \mid \mathcal{D}_n \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$. This means that $\theta_{*,a} - \hat{\mu}_a \mid \mathcal{D}_n \sim \mathcal{N}(0, \hat{\Sigma}_a)$. Thus, $\hat{\Sigma}_a^{-\frac{1}{2}}(\theta_{*,a} - \hat{\mu}_a) \sim \mathcal{N}(0, I_d)$. But notice that $\|\theta_{*,a} - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}} = \|\hat{\Sigma}_a^{-\frac{1}{2}}(\theta_{*,a} - \hat{\mu}_a)\|$. Thus we apply Laurent and Massart (2000, Lemma 1) and get that

$$\mathbb{P}\left(\|\theta_{*,a} - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}} \le \alpha \,\middle|\, \mathcal{D}_n \right) \ge 1 - \frac{\delta}{K},$$

where $\alpha = \sqrt{d + 2\sqrt{d \log \frac{K}{\delta}} + 2 \log \frac{K}{\delta}}$. This means that for every a, $\mathbb{P}(\|\theta_{*,a} - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}} > \alpha \mid \mathcal{D}_n) \le \delta/K$, and by a union bound,

$$\mathbb{P}(\bar{\mathcal{E}}_\alpha \mid \mathcal{D}_n) \le \delta. \tag{C.9}$$

and therefore $\mathbb{P}(\bar{\mathcal{E}}_\alpha) = \mathbb{E}[\mathbb{P}(\bar{\mathcal{E}}_\alpha \mid \mathcal{D}_n)] \le \delta$. Finally, control the bad-event contribution by Cauchy-Schwarz:

$$\mathbb{E}\left[|Z_{\pi_*(X)}| \mathbb{1}\{\bar{\mathcal{E}}_\alpha\}\right] \le \sqrt{\mathbb{E}\left[Z^2_{\pi_*(X)}\right]} \sqrt{\mathbb{P}(\bar{\mathcal{E}}_\alpha)} \le \sqrt{\mathbb{E}\left[Z^2_{\pi_*(X)}\right]} \sqrt{\delta}. \tag{C.10}$$

Putting the pieces together gives

$$\mathrm{BSO}(\hat{\pi}_{\mathrm{G}}) \le \alpha \mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] + \sqrt{\delta} \sqrt{\mathbb{E}\left[Z^2_{\pi_*(X)}\right]}, \tag{C.11}$$

with $\alpha^2 = d + 2\sqrt{d \log(K/\delta)} + 2\log(K/\delta)$.

We now upper bound $\mathbb{E}\left[Z^2_{\pi_*(X)}\right]$ in (C.11). We have that

$$Z^2_{\pi_*(X)} \le \max_{a \in \mathcal{A}} Z_a^2, \qquad \text{hence} \qquad \mathbb{E}\left[Z^2_{\pi_*(X)}\right] \le \mathbb{E}\left[\max_{a \in \mathcal{A}} Z_a^2\right]. \tag{C.12}$$

Fix $(X, \mathcal{D}_n)$. Under the well-specified Bayes assumption, the conditional joint posterior $(\theta_{*,a})_{a\in\mathcal{A}} \mid \mathcal{D}_n$ is Gaussian, and each marginal satisfies $\theta_{*,a} \mid \mathcal{D}_n \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$ (with $\hat{\Sigma}_a$ as derived in Section C.2.3). Therefore for each fixed a,

$$Z_a \mid X, \mathcal{D}_n \sim \mathcal{N}\big(0, s_a^2\big), \qquad s_a^2 = \|\phi(X)\|_{\hat{\Sigma}_a}^2.$$

Let $s_{\max}^2 = \max_{a\in\mathcal{A}} s_a^2$. Then for any $t \geq 0$, by a union bound and the Gaussian tail bound,

$$\mathbb{P}\left(\max_{a\in\mathcal{A}} |Z_a| \geq t \middle| X, \mathcal{D}_n\right) \leq \sum_{a\in\mathcal{A}} \mathbb{P}\left(|Z_a| \geq t | X, \mathcal{D}_n\right) \leq 2K \exp\left(-\frac{t^2}{2s_{\max}^2}\right).$$

Using the identity $\mathbb{E}[W^2] = \int_0^\infty \mathbb{P}(W^2 \geq u)\mathrm{d}u$ for $W \geq 0$ and setting $W = \max_a |Z_a|$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\max_{a\in\mathcal{A}} Z_a^2 \middle| X, \mathcal{D}_n\right] &= \int_0^\infty \mathbb{P}\left(\max_{a\in\mathcal{A}} |Z_a| \geq \sqrt{u} \middle| X, \mathcal{D}_n\right) \mathrm{d}u \\
&\leq \int_0^\infty \min\left\{1,\ 2K \exp\left(-\frac{u}{2s_{\max}^2}\right)\right\} \mathrm{d}u \\
&= \int_0^{2s_{\max}^2 \log(2K)} 1 \mathrm{d}u + \int_{2s_{\max}^2 \log(2K)}^\infty 2K \exp\left(-\frac{u}{2s_{\max}^2}\right) \mathrm{d}u \\
&= 2s_{\max}^2 \log(2K) + 2s_{\max}^2 = \big(2\log(2K) + 2\big) s_{\max}^2.
\end{aligned}$$

Taking expectation over $(X, \mathcal{D}_n)$ yields

$$\mathbb{E}\left[\max_{a\in\mathcal{A}} Z_a^2\right] \leq \big(2\log(2K) + 2\big)\mathbb{E}\left[\max_{a\in\mathcal{A}} \|\phi(X)\|_{\hat{\Sigma}_a}^2\right]. \tag{C.13}$$

Combining (C.12) and (C.13) gives

$$\mathbb{E}\left[Z_{\pi_*(X)}^2\right] \leq \big(2\log(2K) + 2\big)\mathbb{E}\left[\max_{a\in\mathcal{A}} \|\phi(X)\|_{\hat{\Sigma}_a}^2\right]. \tag{C.14}$$

Using the assumption that $\|\phi(X)\|_2 \leq 1$ yields that $\|\phi(X)\|_{\hat{\Sigma}_a}^2 \leq \lambda_1(\hat{\Sigma}_a)$, hence

$$\mathbb{E}\left[Z_{\pi_*(X)}^2\right] \leq \big(2\log(2K) + 2\big) \max_{a\in\mathcal{A}} \lambda_1(\hat{\Sigma}_a). \tag{C.15}$$

Now recall that to simplify, we also assumed that $\Sigma_a = \sigma_0^2 I_d$ for any $a \in \mathcal{A}$ and that $\Sigma = \tau^2 I_{d'}$. As a result, we have that:

$$\hat{\Sigma}_a = \tilde{\Sigma}_a + \sigma_0^{-4} \tilde{\Sigma}_a \mathrm{W}_a \bar{\Sigma} \mathrm{W}_a^\top \tilde{\Sigma}_a,$$

and we also have that $\lambda_1(\tilde{\Sigma}_a) \leq \sigma_0^2$ and that

$$\lambda_1(\hat{\Sigma}_a) \leq \sigma_0^2 + \tau^2, \qquad \forall a \in \mathcal{A}, \tag{C.16}$$

Plugging (C.14) (or (C.15)) into (C.11) yields

$$\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) \leq \alpha\mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] + \sqrt{(2\log(2K) + 2)(\sigma_0^2 + \tau^2)\delta}, \tag{C.17}$$

with $\alpha^2 = d + 2\sqrt{d\log(K/\delta)} + 2\log(K/\delta)$ as defined above. Choosing $\delta = 1/n$ in (C.17) yields

$$\textsc{Bso}(\hat{\pi}_{\text{G}}) \leq \alpha_n \mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] + \sqrt{\frac{(2\log(2K)+2)(\sigma_0^2+\tau^2)}{n}}, \tag{C.18}$$

where

$$\alpha_n^2 = d + 2\sqrt{d\log(Kn)} + 2\log(Kn). \tag{C.19}$$

This concludes the proof. □

C.3.2 Explicit Bound

Additional simplification. To further simplify the exposition, we just set $\phi(x) = x$ for any $x \in \mathcal{X}$.

Assumption 9. *Let $G = \mathbb{E}[XX^\top]$ with $g = \lambda_d(G)$. We assume that $g > 0$.*

Assumption 10 (Context-independent logging policy)**.** *A is independent of X, i.e., $\pi_0(a \mid x) = p_a$ for all x and a. Equivalently, (X_i) are i.i.d. $\sim \nu$ and independent of (A_i), with $\mathbb{P}(A = a) = p_a$.*

Theorem 8 (Explicit Bound)**.** *Let $\pi_*(x)$ be the optimal action for context x. Then, the BSO of* `sDM` *under the structured prior Equation (6.9) satisfies*

$$\begin{aligned}\textit{Bso}(\hat{\pi}_{\text{G}}) \leq \alpha_n \sqrt{\mathbb{E}\left[\frac{1}{\lambda_X} + \frac{\tau^2}{\sigma_0^4\lambda_X^2} + (\sigma_0^2+\tau^2)\left((d+1)e^{-n\rho_X^2/2} + d\left(\tfrac{2}{e}\right)^{gm_X/2}\right)\right]} \\ + \sqrt{\frac{(2\log(2K)+2)(\sigma_0^2+\tau^2)}{n}},\end{aligned}$$

where

$$\alpha_n = \sqrt{d + 2\sqrt{d\log(Kn)} + 2\log(Kn)},$$

and

$$\rho_X = p_{\pi_*(X)}, \qquad m_X = \left\lfloor \frac{n\rho_X}{2} \right\rfloor, \qquad \lambda_X = \sigma^{-2} g \frac{m_X}{2} + \sigma_0^{-2}.$$

Scaling with n. Recall $m_X = \lfloor n\rho_X/2 \rfloor$ and $\lambda_X = \sigma^{-2} g \frac{m_X}{2} + \sigma_0^{-2}$, where $\rho_X = \pi_0(\pi_*(X))$. For n large enough (so that m_X is roughly $n\rho_X$), we have $\lambda_X = \Theta(n\rho_X + 1)$ and the exponentially small tail terms can be ignored at the level of leading-order scaling. Consequently, up to absolute constants and polylogarithmic factors,

$$\textsc{Bso}(\hat{\pi}_{\text{G}}) = \tilde{\mathcal{O}}\left(\alpha_n \sqrt{\mathbb{E}\left[\frac{1}{n\rho_X + 1}\right]} + \sqrt{\frac{\log K}{n}}\right),$$

and using $\alpha_n = \tilde{\Theta}(\sqrt{d})$ this can be summarized as

$$\textsc{Bso}(\hat{\pi}_{\text{G}}) = \tilde{\mathcal{O}}\left(\sqrt{\mathbb{E}\left[\frac{d}{n\rho_X + 1}\right]} + \sqrt{\frac{\log K}{n}}\right),$$

Proof. Let's focus on the main term of Theorem 2, and we start by bounding

$$\mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}}\right], \tag{C.20}$$

Now recall that to simplify, we also assumed that $\Sigma_a = \sigma_0^2 I_d$ for any $a \in \mathcal{A}$ and that $\Sigma = \tau^2 I_{d'}$. As a result, we have that:

$$\hat{\Sigma}_a = \tilde{\Sigma}_a + \sigma_0^{-4}\tilde{\Sigma}_a \mathrm{W}_a \bar{\Sigma} \mathrm{W}_a^\top \tilde{\Sigma}_a,$$

and we also have that $\lambda_1(\tilde{\Sigma}_a) \le \sigma_0^2$ and that

$$\lambda_1(\hat{\Sigma}_a) \le \sigma_0^2 + \tau^2, \qquad \forall a \in \mathcal{A}, \tag{C.21}$$

since $\lambda_1(\mathrm{W}_a \mathrm{W}_a^\top) = 1$. These are obtained using Weyl's inequalities.

Now let

$$N_a = \sum_{i=1}^n \mathbb{1}\{A_i = a\}, \qquad p_a = \mathbb{P}(A = a) = \pi_0(a), \qquad \rho_x = p_{\pi_*(x)}, \qquad \rho_X = p_{\pi_*(X)}.$$

Define

$$m_x = \left\lfloor \frac{n\rho_x}{2} \right\rfloor, \qquad m_X = \left\lfloor \frac{n\rho_X}{2} \right\rfloor.$$

Then, under Assumption 10, $N_a \sim \mathrm{Bin}(n, p_a)$ and it is independent of context X. Therefore, Hoeffding gives, for $t = n\rho_X/2$,

$$\mathbb{P}\left(N_{\pi_*(X)} < \frac{n\rho_X}{2} \middle| X, \theta_*\right) \le \exp\left(-\frac{n\rho_X^2}{2}\right). \tag{C.22}$$

Let

$$\Omega_{X,1} = \left\{N_{\pi_*(X)} \ge m_X\right\}.$$

Then we have that

$$\mathbb{P}\left(\bar{\Omega}_{X,1} \middle| X, \theta_*\right) \le \exp\left(-\frac{n\rho_X^2}{2}\right). \tag{C.23}$$

Lemma 7 (Matrix Chernoff, Tropp (2012, Theorem 1.1))**.** *Let $(Y_k)_{k=1}^m$ be independent PSD matrices with $\lambda_1(Y_k) \le R$ a.s. Let $\mu_{\min} = \lambda_d(\sum_{k=1}^m \mathbb{E}[Y_k])$. Then for $\delta \in [0,1]$,*

$$\mathbb{P}\left(\lambda_d\Big(\sum_{k=1}^m Y_k\Big) \le (1-\delta)\mu_{\min}\right) \le d\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R}.$$

Define

$$\Omega_{X,2} = \left\{\lambda_d\Big(\sum_{i=1}^n \mathbb{1}\{A_i = \pi_*(X)\} X_i X_i^\top\Big) \ge \frac{1}{2} N_{\pi_*(X)} g\right\}, \qquad \Omega_X = \Omega_{X,1} \cap \Omega_{X,2}.$$

We now bound $\mathbb{P}(\bar{\Omega}_{X,2} \mid X, \theta_*)$. Recall that

$$\Omega_{X,1} = \left\{N_{\pi_*(X)} \ge m_X\right\}, \qquad m_X = \left\lfloor \frac{n\rho_X}{2} \right\rfloor.$$

Under Assumption 10, (A_i) is independent of (X_i), hence conditional on the index set $S_{\pi_*(X)} = \{i \in [n] : A_i = \pi_*(X)\}$, the matrices $\{X_i X_i^\top : i \in S_{\pi_*(X)}\}$ are i.i.d. with the same law as $XX^\top$. Moreover, since the Chernoff bound below depends on $S_{\pi_*(X)}$ only through $|S_{\pi_*(X)}| = N_{\pi_*(X)}$, the same bound holds when conditioning on $N_{\pi_*(X)}$. Moreover, we have that $\phi(x) = x$ and that $\|\phi(x)\| \le 1$. Thus, $\|x\| \le 1$ and hence $\lambda_1(X_i X_i^\top) \le 1$. Applying Lemma 7 with $R = 1$, $\mathbb{E}[XX^\top] = G$, and

$$\mu_{\min} = \lambda_d \left(N_{\pi_*(X)} G \right) = N_{\pi_*(X)} g,$$

and choosing $\delta = \frac{1}{2}$, we obtain the conditional bound

$$\mathbb{P}(\bar{\Omega}_{X,2} \mid X, N_{\pi_*(X)}, \theta_*) \le d \left(\sqrt{2/e} \right)^{g N_{\pi_*(X)}}.$$

Taking conditional expectation $N_{\pi_*(X)}$ given X and θ_*,

$$\mathbb{P}(\bar{\Omega}_{X,2} \mid X, \theta_*) = \mathbb{E}\left[\mathbb{P}(\bar{\Omega}_{X,2} \mid X, N_{\pi_*(X)}, \theta_*) \mid X, \theta_* \right] \le d \mathbb{E}\left[\left(\sqrt{2/e} \right)^{g N_{\pi_*(X)}} \mid X, \theta_* \right].$$

Splitting on $\Omega_{X,1}$ yields

$$\begin{aligned} \mathbb{E}\left[\left(\sqrt{2/e} \right)^{g N_{\pi_*(X)}} \mid X, \theta_* \right] &= \mathbb{E}\left[\left(\sqrt{2/e} \right)^{g N_{\pi_*(X)}} \mathbb{1}\{\Omega_{X,1}\} \mid X, \theta_* \right] + \mathbb{E}\left[\left(\sqrt{2/e} \right)^{g N_{\pi_*(X)}} \mathbb{1}\{\bar{\Omega}_{X,1}\} \mid X, \theta_* \right] \\ &\le \left(\sqrt{2/e} \right)^{g m_X} + \mathbb{P}(\bar{\Omega}_{X,1} \mid X, \theta_*), \end{aligned}$$

since on $\Omega_{X,1}$ we have $N_{\pi_*(X)} \ge m_X$ and thus $\left(\sqrt{2/e} \right)^{g N_{\pi_*(X)}} \le \left(\sqrt{2/e} \right)^{g m_X}$, while on $\bar{\Omega}_{X,1}$ we have $\left(\sqrt{2/e} \right)^{g N_{\pi_*(X)}} \le 1$. Therefore,

$$\mathbb{P}(\bar{\Omega}_{X,2} \mid X, \theta_*) \le d \left(\sqrt{2/e} \right)^{g m_X} + d \mathbb{P}(\bar{\Omega}_{X,1} \mid X, \theta_*). \tag{C.24}$$

Combining (C.24) with (C.22) yields

$$\mathbb{P}(\bar{\Omega}_X \mid X, \theta_*) \le \mathbb{P}(\bar{\Omega}_{X,1} \mid X, \theta_*) + \mathbb{P}(\bar{\Omega}_{X,2} \mid X, \theta_*) \le (d+1) \exp\left(-\frac{n \rho_X^2}{2} \right) + d \left(\sqrt{2/e} \right)^{g m_X}. \tag{C.25}$$

Finally, let

$$I_1 = \mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{1}\{\Omega_X\} \mid X, \theta_* \right], \qquad I_2 = \mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{1}\{\bar{\Omega}_X\} \mid X, \theta_* \right].$$

Then

$$\mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \right] = \mathbb{E}[I_1] + \mathbb{E}[I_2].$$

Using $\|X\|_2 \le 1$ and (C.16),

$$\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} = X^\top \hat{\Sigma}_{\pi_*(X)} X \le \lambda_1(\hat{\Sigma}_{\pi_*(X)}) \|X\|_2^2 \le \sigma_0^2 + \tau^2.$$

Let $c_1 = \sigma_0^2 + \tau^2$. Then

$$I_2 \leq c_1 \mathbb{P}(\bar{\Omega}_X \mid X, \theta_*) \leq c_1 \left((d+1) \exp\left(-\frac{n\rho_X^2}{2} \right) + d \left(\sqrt{2/e} \right)^{gm_X} \right). \tag{C.26}$$

Moreover, fix X and θ_*, on Ω_X,

$$\lambda_d\Big(\sum_{i=1}^n \mathbb{1}\{A_i = \pi_*(X)\} X_i X_i^\top \Big) \geq \frac{1}{2} N_{\pi_*(X)} g \geq \frac{1}{2} m_X g,$$

so

$$\lambda_d(\hat{G}_{\pi_*(X)}) = \sigma^{-2} \lambda_d\Big(\sum_{i=1}^n \mathbb{1}\{A_i = \pi_*(X)\} X_i X_i^\top \Big) \geq \sigma^{-2} \frac{1}{2} m_X g.$$

Hence

$$\lambda_d(\hat{G}_{\pi_*(X)} + \sigma_0^{-2} I_d) \geq \sigma^{-2} \frac{1}{2} m_X g + \sigma_0^{-2}.$$

Let

$$\lambda_X = \sigma^{-2} g \frac{m_X}{2} + \sigma_0^{-2}.$$

Since $\tilde{\Sigma}_{\pi_*(X)} = (\hat{G}_{\pi_*(X)} + \sigma_0^{-2} I_d)^{-1}$, we obtain on Ω_X:

$$\lambda_1(\tilde{\Sigma}_{\pi_*(X)}) \leq \frac{1}{\lambda_X}.$$

Moreover, on Ω_X we have that

$$\lambda_1(\hat{\Sigma}_{\pi_*(X)}) \leq \lambda_1(\tilde{\Sigma}_{\pi_*(X)}) + \sigma_0^{-4} \tau^2 \lambda_1(\tilde{\Sigma}_{\pi_*(X)})^2 \leq \frac{1}{\lambda_X} + \frac{\sigma_0^{-4} \tau^2}{\lambda_X^2}.$$

Therefore, since $\|X\|_2 \leq 1$,

$$I_1 \leq \left(\frac{1}{\lambda_X} + \frac{\sigma_0^{-4} \tau^2}{\lambda_X^2} \right). \tag{C.27}$$

Combining (C.26) and (C.27),

$$\mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \right] \leq \mathbb{E}\left[\frac{1}{\lambda_X} + \frac{\sigma_0^{-4} \tau^2}{\lambda_X^2} + (\sigma_0^2 + \tau^2) \left((d+1) \exp\left(-\frac{n\rho_X^2}{2} \right) + d \left(\sqrt{2/e} \right)^{gm_X} \right) \right].$$

But from Theorem 2, we know that

$$\textsc{Bso}(\hat{\pi}_{\textsc{g}}) \leq \alpha_n \mathbb{E}\left[\|X\|_{\hat{\Sigma}_{\pi_*(X)}} \right] + \sqrt{\frac{(2\log(2K) + 2)(\sigma_0^2 + \tau^2)}{n}}, \tag{C.28}$$

where $\alpha_n = \sqrt{d + 2\sqrt{d \log(Kn)} + 2\log(Kn)}$.

Finally, by Jensen's inequality, $\mathbb{E}\left[\|X\|_{\hat{\Sigma}_{\pi_*(X)}}\right] \leq \sqrt{\mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}}\right]}$. Substituting into (C.28) and combining with (C.26) and (C.27), we obtain

$$\begin{aligned}\textsc{Bso}(\hat{\pi}_{\text{G}}) \leq \alpha_n \sqrt{\mathbb{E}\left[\frac{1}{\lambda_X} + \frac{\tau^2}{\sigma_0^4 \lambda_X^2} + (\sigma_0^2 + \tau^2)\left((d+1)e^{-n\rho_X^2/2} + d\left(\tfrac{2}{e}\right)^{gm_X/2}\right)\right]} \\ + \sqrt{\frac{(2\log(2K)+2)(\sigma_0^2+\tau^2)}{n}},\end{aligned}$$

where $\alpha_n = \sqrt{d + 2\sqrt{d\log(Kn)} + 2\log(Kn)}$.

□

C.3.3 Optimality of Greedy Policies

Here, we show that Greedy policy $\hat{\pi}_{\text{G}}$ should be preferred to any other choice of policies when considering the BSO as our performance metric. This is because $\hat{\pi}_{\text{G}}$ minimizes the BSO. To see this, note that by definition the Greedy policy $\hat{\pi}_{\text{G}}$ is deterministic, that is for any context $x \in \mathcal{X}$, there exists $\hat{a}_{\text{G}}$, such that $\hat{\pi}_{\text{G}}(\hat{a}_{\text{G}} \mid x) = 1$. Thus, for any context $x \in \mathcal{X}$, we simplify the notation by letting $\hat{\pi}_{\text{G}}(x)$ denote the action that has a mass equal to 1. Then, we have that

$$\begin{aligned}\mathbb{E}_{A\sim\hat{\pi}_{\text{G}}(\cdot|x)}\left[\mathbb{E}_{\theta_*}\left[r(x,A;\theta_*) \mid \mathcal{D}_n\right]\right] &= \mathbb{E}_{\theta_*}\left[r(x,\hat{\pi}_{\text{G}}(x);\theta_*) \mid \mathcal{D}_n\right], && \text{(C.29)}\\ &\geq \mathbb{E}\left[r(x,a;\theta_*) \mid \mathcal{D}_n\right] \qquad \forall x,a \in \mathcal{X}\times\mathcal{A}. && \text{(C.30)}\end{aligned}$$

where this follows from the definition of $\hat{r}(x,a) = \mathbb{E}\left[r(x,a;\theta) \mid \mathcal{D}_n\right]$, the definition of $\hat{\pi}_{\text{G}}$ and the fact that θ_* is sampled from the prior, which leads to $\mathbb{E}\left[r(x,a;\theta) \mid \mathcal{D}_n\right] = \mathbb{E}\left[r(x,a;\theta_*) \mid \mathcal{D}_n\right]$. Now Equation (C.29) holds for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$, and hence it holds in expectation under $X \sim \nu$ and $A \sim \pi(\cdot \mid X)$ for any policy π. That is,

$$\mathbb{E}_{X\sim\nu,A\sim\hat{\pi}_{\text{G}}(\cdot|X)}\left[\mathbb{E}_{\theta_*}\left[r(x,A;\theta_*) \mid \mathcal{D}_n\right]\right] \geq \mathbb{E}_{X\sim\nu,A\sim\pi(\cdot|X)}\left[\mathbb{E}_{\theta_*}\left[r(x,A;\theta_*) \mid \mathcal{D}_n\right]\right]. \qquad \text{(C.31)}$$

Taking another expectation w.r.t. the sample set $\mathcal{D}_n$ and using Fubini's theorem and the tower rule leads to $\mathbb{E}\left[V(\hat{\pi}_{\text{G}};\theta_*)\right] \geq \mathbb{E}\left[V(\pi;\theta_*)\right]$ for any stationary policy π. Then, subtracting $\mathbb{E}\left[V(\pi_*;\theta_*)\right]$ from both sides of the previous inequality yields that the BSO is minimized by $\hat{\pi}_{\text{G}}$ compared to any stationary policy π, in particular, compared to the policy π_{P} induced by pessimism.

C.4 Additional Experiments

As mentioned in Section C.4, our experiments were conducted on internal machines with 30 CPUs and thus they required a moderate amount of computation. These experiments are also reproducible with minimal computational resources.

C.4.1 Implementation Details of Baselines

We implement the baselines as follows.

- **IPS.**

$$\operatorname*{argmax}_{\pi} \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i|X_i)}{\max\{\pi_0(A_i|X_i), \tau\}} R_i, \tag{C.32}$$

  where $\tau \in [0, 1]$ is a hyper-parameter.
- **snIPS.**

$$\operatorname*{argmax}_{\pi} \frac{1}{\sum_{i=1}^{n} \frac{\pi(A_i|X_i)}{\pi_0(A_i|X_i)}} \sum_{i=1}^{n} \frac{\pi(A_i|X_i)}{\pi_0(A_i|X_i)} R_i, \tag{C.33}$$

- **MIPS.** We cluster actions into L groups and let $h(a)$ be the cluster of action a. Let C_i be the cluster of action A_i, then we use

$$\operatorname*{argmax}_{\pi} \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(C_i|X_i)}{\pi_0(C_i|X_i)} R_i, \tag{C.34}$$

  where $C_i = h(A_i)$ for any $i \in [n]$, and $\pi(c|x) = \sum_{a \in \mathcal{A}} \mathbb{1}\left[h(a) = c\right] \pi(a|x)$.
- **PC.** We use the Knn implementation of `PC`. Let $N(a, k)$ be the set of k-nearest neighbors of a, then

$$\operatorname*{argmax}_{\pi} \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{a \in \mathcal{A}} \mathbb{1}\left[a \in N(A_i, k)\right] \pi(a|x)}{\sum_{a \in \mathcal{A}} \mathbb{1}\left[a \in N(A_i, k)\right] \pi_0(a|x)} R_i. \tag{C.35}$$

- **DM (Freq).** This DM uses the linear-Gaussian likelihood model $R \mid \theta, X, A \sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2)$ and learn the parameters θ_a using the maximum likelihood principle leading to

$$\hat{r}(x, a) = \phi(x)^\top \hat{\mu}_a, \tag{C.36}$$

  where the MLE is $\hat{\mu}_a = (G_a + \lambda I_d)^{-1} B_a$, with $G_a = \sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} \phi(X_i) \phi(X_i)^\top$ and $B_a = \sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} R_i \phi(X_i)$, and λ is a regularization hyper-parameter.
- **DM (Bayes).** This DM uses the linear-Gaussian likelihood model combined with Gaussian priors as

$$\begin{aligned} \theta_a &\sim \mathcal{N}(\mu_a, \Sigma_a), && \forall a \in \mathcal{A}, \\ R \mid \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2), \end{aligned} \tag{C.37}$$

  Under this prior, each action a has an associated parameter θ_a. Given the prior in Equation (6.2), the posterior distribution of an action parameter follows a multivariate Gaussian: $\theta_a \mid \mathcal{D}_n \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$, where $\hat{\Sigma}_a^{-1} = \Sigma_a^{-1} + G_a$ and $\hat{\Sigma}_a^{-1} \hat{\mu}_a = \Sigma_a^{-1} \mu_a + B_a$. Here, $G_a = \sigma^{-2} \sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} \phi(X_i) \phi(X_i)^\top$ and $B_a = \sigma^{-2} \sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} R_i \phi(X_i)$. Then, the reward estimate is

$$\hat{r}(x, a) = \phi(x)^\top \hat{\mu}_a, \tag{C.38}$$

- **DR.**

$$\operatorname*{argmax}_{\pi} \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i|X_i)}{\max\{\pi_0(A_i|X_i), \tau\}} \left(R_i - \hat{r}(X_i, A_i)\right) + \mathbb{E}_{A \sim \pi(\cdot|X_i)}\left[\hat{r}(X_i, A)\right], \tag{C.39}$$

  with $\tau \in [0, 1]$ and $\hat{r}$ is the reward model obtained using `DM (Freq)`.

C.4.2 Robustness to Likelihood Misspecification

We strengthened our evaluation by assessing `sDM`'s robustness to likelihood misspecification below (robustness to prior misspecification is provided in Section C.4.3). In these experiments, the true data-generating process (same as the synthetic experiments in Section 6.5) differed from `sDM`'s assumptions in two different ways: either the likelihood is misspecified

Misspecified likelihood (Figure C.1). We also simulate when the true reward distribution differed from the likelihood assumed by `sDM`. For example, we simulated binary rewards using a Bernoulli-logistic model while `sDM` used a linear-Gaussian likelihood. Other DMs: `DM (Bayes)` and `DM (Freq)` also use a misspecified likelihood model and to emphasize this we add the suffix `Lin` to all DMs names. Overall, `sDM` still outperforms all methods by a large margin despite misspecification.

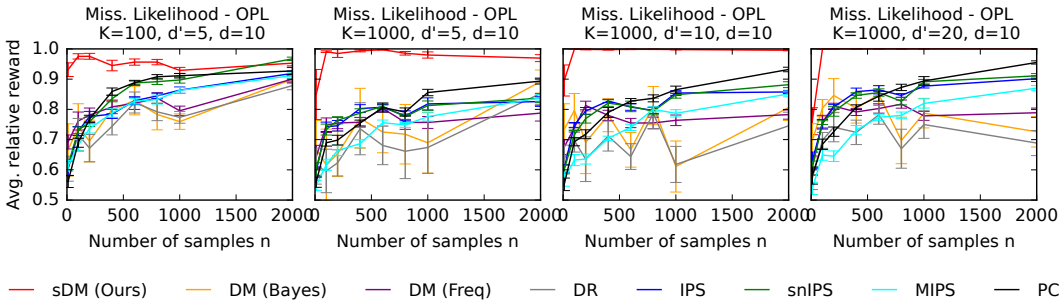


Figure C.1: Effect of likelihood misspecification: The relative reward of the learned policy on synthetic problems using misspecified likelihood with varying n and K.

C.4.3 Robustness to Prior Misspecification

Misspecified prior means and covariances(Figure C.2). This is achieved by adding uniformly sampled noise from $[v, v + 0.5]$ to both the true prior mean and covariance parameters $\mu, \Sigma, \mathrm{W}_a, \Sigma_a$, with v controlling the level of misspecification. We varied $v \in \{0.5, 1, 1.5\}$ and analyzed its impact on `sDM`'s performance. For comparison, we included the well-specified `sDM` and the most competitive baseline, `DM (Bayes)`, while omitting other baselines to reduce clutter. `sDM`'s performance decreases with increasing misspecification, yet `sDM` with misspecification still outperforms the most competitive baseline, especially when K is large. We also observe that the impact of prior covariance misspecification is less significant compared to prior mean misspecification.

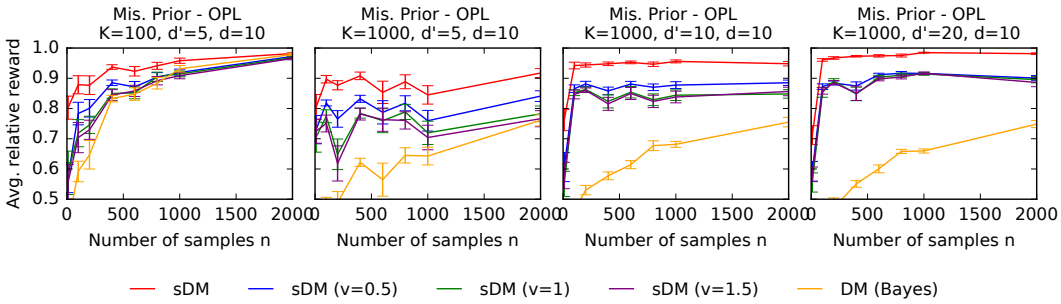


Figure C.2: Effect of prior mean and covariance misspecification: The average relative reward of the learned policy on synthetic problems using both misspecified prior means and covariances with varying n and K and d'.

C.4.4 Comparison of Greedy and Pessimistic Policies

To validate our theory that a greedy policy should be preferred over the commonly adopted pessimistic policy in our Bayesian setting, we used a performance metric averaged over multiple bandit problems sampled from the prior. To verify this, we considered the same OPL synthetic setting as in Section 6.5 and compared `sDM` with a greedy policy to `sDM` with a pessimistic policy. Recall that a greedy policy with respect to our reward estimate writes

$$\hat{\pi}_{\mathrm{G}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b \in \mathcal{A}} \hat{r}(x, b)\}, \tag{C.40}$$

while a pessimistic one writes

$$\hat{\pi}_{\mathrm{P}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b \in \mathcal{A}} \hat{r}(x, b) - u(x, a)\}, \tag{C.41}$$

where $u(x, a) = \alpha(d, \delta)\|\phi(X)\|_{\hat{\Sigma}_a}$ with $\alpha(d, \delta) = \sqrt{d + 2\sqrt{d \log \frac{1}{\delta}} + 2 \log \frac{1}{\delta}}$. As predicted by our theory, the results show that the greedy policy has better average performance over multiple bandit instances sampled from the prior.

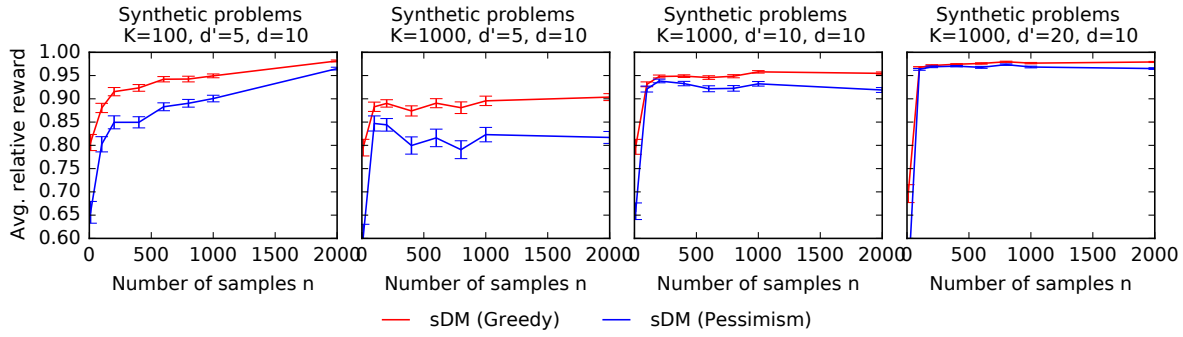


Figure C.3: Comparison of `sDM` with greedy policy and `sDM` with pessimistic policy in OPL: The average MSE of an ϵ-greedy policy on synthetic problems with varying n, K, and d'.

Chapter D

Supplementary Materials for Chapter 7

Contents

D.1 Proofs for Oracle Policies

D.1.1 Oracle Policies for IPS-Based Objectives

(IPS), cIPS and ES. Recall the definition of the (logging propensity) clipped IPS estimator with $\tau \in [0, 1]$:

$$\hat{V}_{\text{CIPS}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(A_i|X_i)}{\max\{\pi_0(A_i|X_i), \tau\}} R_i \,.$$

Taking $n \to \infty$, one obtains:

$$V_{\text{CIPS}}(\pi) = \mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\pi(A|X)}{\max\{\pi_0(A|X),\tau\}} r(X,A)\right]$$
$$= \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[\frac{\pi_0(A|X)}{\max\{\pi_0(A|X),\tau\}} r(X,A)\right].$$

As the objective is linear in the policy π, the optimal policy should put for any $x \in \mathcal{X}$, all the mass on the action a that maximizes the weighted reward, giving:

$$\pi_*^{\texttt{cIPS}}(a|x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a'\in\mathcal{A}} \frac{\pi_0(a'|x) r(x,a')}{\max\{\pi_0(a'|x),\tau\}}\right].$$

We recover the solution for `IPS` when we let $\tau \to 0$:

$$\pi_*^{\texttt{IPS}}(a|x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a'\in\mathcal{A}} r(x,a')\mathbb{1}\left[\pi_0(a'|x) > 0\right]\right].$$

We also recover the solution of `ES` just by replacing the clipping function by an exponential function of factor α, obtaining:

$$\pi_*^{\texttt{ES}}(a|x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a'\in\mathcal{A}} r(x,a')\pi_0(a'|x)^{1-\alpha}\right].$$

Doubly Robust (DR). The doubly robust estimator converges to the following quantity:

$$V_{\text{DR}}(\pi) = \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[(r(X,A) - \hat{r}(X,A))\frac{\pi_0(A|X)}{\max\{\pi_0(A|X),\tau\}} + \hat{r}(X,A)\right].$$

The objective is linear in π and is thus maximized by the following deterministic decision rule:

$$\pi_*^{\texttt{DR}}(a|x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a'\in\mathcal{A}} \hat{r}(x,a') + (r(x,a') - \hat{r}(x,a'))\frac{\pi_0(a'|x)}{\max\{\pi_0(a'|x),\tau\}}\right]$$

Marginalized IPS (MIPS) with clusters. We adopt the same approach to look for the maximizer of `MIPS`. We generalize the clustering function h to also account for context. We write down the estimator:

$$\hat{V}_{\text{MIPS}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{a'} \mathbb{1}\left[h(a',X_i) = h(A_i,X_i)\right]\pi(a'|X_i)}{\sum_{a''} \mathbb{1}\left[h(a'',X_i) = h(A_i,X_i)\right]\pi_0(a''|X_i)} R_i = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(C_i|X_i)}{\pi_0(C_i|X_i)} R_i\,,$$

with which, we recover when $n \to \infty$:

$$V_{\text{MIPS}}(\pi) = \mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\sum_{a'} \mathbb{1}\left[h(a',X) = h(A,X)\right]\pi(a'|X)}{\sum_{a''} \mathbb{1}\left[h(a'',X) = h(A,X)\right]\pi_0(a''|X)} r(X,A)\right]$$
$$= \mathbb{E}_{X\sim\nu}\left[\sum_a \pi_0(a|X)\frac{\sum_{a'} \mathbb{1}\left[h(a',X) = h(a,X)\right]\pi(a'|X)}{\sum_{a''} \mathbb{1}\left[h(a'',X) = h(a,X)\right]\pi_0(a''|X)} r(X,a)\right]$$
$$= \mathbb{E}_{X\sim\nu}\left[\sum_{a'} \pi(a'|X)\sum_a \pi_0(a|X)\frac{\mathbb{1}\left[h(a',X) = h(a,X)\right]}{\sum_{a''} \mathbb{1}\left[h(a'',X) = h(a,X)\right]\pi_0(a''|X)} r(X,a)\right]$$
$$= \mathbb{E}_{X\sim\nu}\left[\sum_{a'} \pi(a'|X)\mathbb{E}_{A\sim\pi_0(\cdot|X)}\left[\frac{\mathbb{1}\left[h(a',X) = h(A,X)\right] r(X,A)}{\mathbb{E}_{A''\sim\pi_0(\cdot|X)}\left[\mathbb{1}\left[h(A'',X) = h(A,X)\right]\right]}\right]\right].$$

The objective is linear in π, and depends on the action a' through its cluster $h(a', \cdot)$ alone. This means that multiple solutions are maximizers as long as the policy chooses the best cluster c. We thus write down the oracle policy for `MIPS` in the cluster level, giving:

$$\begin{aligned}
\pi_*^{\texttt{MIPS}}(c|x) &= \mathbb{1}\Big[c = \operatorname*{argmax}_{c' \in \mathcal{C}} \Big\{ \mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[\frac{r(x,A)\mathbb{1}[h(A,x) = c']}{\mathbb{E}_{A'' \sim \pi_0(\cdot|x)} \left[\mathbb{1}\left[h(A'',x) = h(A,x)\right]\right]} \right] \Big\} \Big] \\
&= \mathbb{1}\Big[c = \operatorname*{argmax}_{c' \in \mathcal{C}} \Big\{ \mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[\frac{r(x,A)\mathbb{1}[h(A,x) = c']}{\mathbb{E}_{A'' \sim \pi_0(\cdot|x)} \left[\mathbb{1}\left[h(A'',x) = c'\right]\right]} \right] \Big\} \Big] \\
&= \mathbb{1}\Big[c = \operatorname*{argmax}_{c' \in \mathcal{C}} \Big\{ \frac{\mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[r(x,A)\mathbb{1}[h(A,x) = c']\right]}{\mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[\mathbb{1}\left[h(A,x) = c'\right]\right]} \Big\} \Big],
\end{aligned}$$

which ends the proof.

Conjunct Effect Modeling (OffCEM). This estimator can be seen as the natural, doubly robust extension of the `MIPS` estimator. Combining similar techniques to the ones employed for `MIPS` and `DR` yields

$$\pi_*^{\texttt{OffCEM}}(a|x) = \mathbb{1}\Big[a = \operatorname*{argmax}_{a' \in \mathcal{A}} \Big\{ \hat{r}(x,a') + \frac{\mathbb{E}_{\bar{A} \sim \pi_0(\cdot|x)} \left[(r(x,\bar{A}) - \hat{r}(x,\bar{A}))\mathbb{1}[h(x,\bar{A}) = h(x,a')]\right]}{\pi_0(h(x,a')|x)} \Big\} \Big].$$

Two Stage Decomposition (POTEC). This is an *optimization strategy* for `OffCEM`. It restricts the policy to a cluster-informed form,

$$\pi(a \mid x) = \sum_{c \in \mathcal{C}} \pi^{\mathrm{RM}}(a \mid x, c)\pi^{\mathrm{CL}}(c \mid x),$$

where $\pi^{\mathrm{RM}}(a \mid x, c) = \mathbb{1}[a = \operatorname{argmax}_{a' \in c} \hat{r}(x,a')]$ is fixed, model-based policy that deterministically selects the best action within each cluster. Learning is then simplified to finding the optimal cluster-level policy π^{CL} that maximizes the `OffCEM` objective:

$$\hat{V}_{\mathrm{POTEC}}(\pi^{\mathrm{CL}}) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\pi^{\mathrm{CL}}(C_i \mid X_i)}{\pi_0(C_i \mid X_i)} \left(R_i - \hat{r}(X_i, A_i)\right) + \sum_{c \in \mathcal{C}} \pi^{\mathrm{CL}}(c \mid X_i)\hat{r}_c^*(X_i) \right),$$

where $\hat{r}_c^*(x) = \max_{a \in c} \hat{r}(x,a)$ is the estimated reward of the best action in cluster c. This is exactly the Doubly Robust version of `MIPS` on the cluster level, the oracle policy on the cluster level can be followed in the same fashion:

$$\pi_*^{\mathrm{CL}}(c \mid x) = \mathbb{1}\Big[c = \operatorname*{argmax}_{c' \in \mathcal{C}} \Big\{ \frac{\mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[(r(x,A) - \hat{r}(x,A))\mathbb{1}[h(A,x) = c']\right]}{\mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[\mathbb{1}\left[h(A,x) = c'\right]\right]} + \hat{r}_{c'}^*(x) \Big\} \Big].$$

The optimal policy for the `POTEC` optimization strategy unfolds as:

$$\pi_*^{\texttt{POTEC}}(a|x) = \sum_{c \in \mathcal{C}} \pi^{\mathrm{RM}}(a \mid x, c)\pi_*^{\mathrm{CL}}(c \mid x).$$

At first glance, it might be hard to see the connection between `POTEC` and `OffCEM` solutions, but they are equivalent. For ease of notation, let us denote by $D_{\hat{r},x}(c)$:

$$D_{\hat{r},x}(c) = \frac{\mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[(r(x,A) - \hat{r}(x,A))\mathbb{1}[h(A,x) = c]\right]}{\mathbb{E}_{A \sim \pi_0(\cdot|x)} \left[\mathbb{1}\left[h(A,x) = c\right]\right]}.$$

and recall that the optimal policy of `OffCEM` finds the action a that maximizes:

$$\tilde{V}(x,a) = \hat{r}(x,a) + D_{\hat{r},x}(h(a,x))\,.$$

For any context x, the optimal action a^* of `POTEC` verifies:

- a^* is in the optimal cluster: $h(a^*,x) = c_*(x)$ with $c_*(x) = \operatorname{argmax}_{c\in\mathcal{C}} D_{\hat{r},x}(c) + \hat{r}^*_c(x)$.
- a^* is optimal within that cluster: $a = \operatorname{argmax}_{a\in c_*(x)} \hat{r}(x,a)$.

This means that for all actions a with $h(a,x) \neq c_*(x)$, we have:

$$\begin{aligned}
\tilde{V}(x,a) &= D_{\hat{r},x}(h(a,x)) + \hat{r}(x,a) \\
&\leq D_{\hat{r},x}(h(a,x)) + \hat{r}^*_{h(a,x)}(x) \\
&\leq D_{\hat{r},x}(c_*(x)) + \hat{r}^*_{c_*(x)}(x) \\
&= D_{\hat{r},x}(h(x,a^*)) + \hat{r}(x,a^*) = \tilde{V}(x,a^*)\,.
\end{aligned}$$

In addition, for all actions a with $h(a,x) = c_*(x)$, we have:

$$\begin{aligned}
\tilde{V}(x,a) &= D_{\hat{r},x}(h(a,x)) + \hat{r}(x,a) \\
&= D_{\hat{r},x}(c_*(x)) + \hat{r}(x,a) \\
&\leq D_{\hat{r},x}(c_*(x)) + \hat{r}^*_{c_*(x)}(x) = \tilde{V}(x,a^*)\,.
\end{aligned}$$

This means that the optimal action a^* for `POTEC` is the maximizer of $\tilde{V}(x,a)$, which is exactly the solution of `OffCEM`.

Policy Convolution (PC). This estimator uses a nearest neighbors function to aggregate the propensities of similar actions, making the hypothesis that similar actions will result in similar reward signal. The estimator writes:

$$\hat{V}_{\mathrm{PC}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(N_\epsilon(A_i) \mid X_i)}{\pi_0(N_\epsilon(A_i) \mid X_i)} R_i, \quad \text{with } \pi(N_\epsilon(a) \mid x) = \sum_{a'\in N_\epsilon(a)} \pi(a' \mid x).$$

This estimator is equivalent to the following when $n \to \infty$:

$$\begin{aligned}
V^{\mathtt{PC}}(\pi) &= \mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\sum_{a'} \pi(a'|X)\mathbb{1}\left[a' \in N_\epsilon(A)\right]}{\pi_0(N_\epsilon(A)|X)} r(X,A)\right] \\
&= \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[\mathbb{E}_{\bar{A}\sim\pi_0(\cdot|X)}\left[\frac{r(x,\bar{A})\mathbb{1}\left[A \in N_\epsilon(\bar{A})\right]}{\pi_0(N_\epsilon(\bar{A})|X)}\right]\right].
\end{aligned}$$

The same argument of linearity applies here, giving us the corresponding oracle policy:

$$\pi^{\mathtt{PC}}_*(a|x) = \mathbb{1}\left[a = \operatorname*{argmax}_{a'\in\mathcal{A}} \left\{\mathbb{E}_{\bar{A}\sim\pi_0(\cdot|x)}\left[\frac{r(x,\bar{A})\mathbb{1}[a' \in N_\epsilon(\bar{A})]}{\pi_0(N_\epsilon(\bar{A})|x)}\right]\right\}\right].$$

D.1.2 Oracle Policies for PWLL-Based Objectives

Our objectives can be written in the same form, only choosing for each a different function g:

$$\hat{U}_g(\pi) = \frac{1}{n} \sum_{i=1}^{n} g(X_i, A_i, R_i) \log \pi(A_i \mid X_i) \,.$$

Since we are looking at oracle policies, we consider the expectation

$$U_g(\pi) = \mathbb{E}_{X \sim \nu,\, A \sim \pi_0(\cdot|X),\, R \sim p(\cdot|X,A)} \left[g(X, A, R) \log \pi(A \mid X) \right].$$

The maximization decomposes over contexts. Fix x and define the nonnegative weights

$$w_x(a) = \mathbb{E}_{R \sim p(\cdot|x,a)} \left[g(x, a, R) \right] \geq 0.$$

For each x, we thus consider

$$\begin{aligned} \max_{\pi(\cdot|x)} \quad & \sum_{a \in \mathcal{A}} \pi_0(a \mid x)\, w_x(a)\, \log \pi(a \mid x) \\ \text{s.t.} \quad & \sum_{a \in \mathcal{A}} \pi(a \mid x) = 1, \qquad \forall a \in \mathcal{A},\ \pi(a \mid x) \geq 0. \end{aligned}$$

Let $v_x(a) = \pi_0(a \mid x)\, w_x(a) \geq 0$. The Lagrangian (with equality multiplier $\lambda \in \mathbb{R}$ and inequality multipliers $\{\mu(a)\}_{a \in \mathcal{A}}$, $\mu(a) \geq 0$) is

$$\mathcal{L}(\pi, \lambda, \mu) = \sum_{a \in \mathcal{A}} v_x(a) \log \pi(a \mid x) + \lambda \Big(\sum_{a \in \mathcal{A}} \pi(a \mid x) - 1 \Big) + \sum_{a \in \mathcal{A}} \mu(a)\, \pi(a \mid x).$$

By KKT conditions, at an optimum $\pi_*^g(\cdot \mid x)$ we have for all $a \in \mathcal{A}$:

$$\frac{\partial \mathcal{L}}{\partial \pi(a \mid x)} = \frac{v_x(a)}{\pi(a \mid x)} + \lambda + \mu(a) = 0, \qquad \text{and} \quad \mu(a)\, \pi(a \mid x) = 0.$$

For any action with $\pi_*^g(a \mid x) > 0$, we get that $\mu(a) = 0$, and hence

$$\pi_*^g(a \mid x) = -\frac{v_x(a)}{\lambda}.$$

Normalizing with $\sum_a \pi_*^g(a \mid x) = 1$ gives $\lambda = -\sum_{a'} v_x(a')$ and therefore

$$\pi_*^g(a \mid x) = \frac{v_x(a)}{\sum_{a' \in \mathcal{A}} v_x(a')} = \frac{\pi_0(a \mid x)\, \mathbb{E}_{R \sim p(\cdot|x,a)} \left[g(x, a, R) \right]}{\sum_{a' \in \mathcal{A}} \pi_0(a' \mid x)\, \mathbb{E}_{R \sim p(\cdot|x,a')} \left[g(x, a', R) \right]}.$$

This concludes the proof.

D.2 Proofs for Optimization Properties

In this section, we prove the propositions about the optimization landscape of IPS-based and PWLL learning approaches. We start by stating the following lemmas, that will be helpful to prove our propositions.

Lemma 8. *(Mei et al., 2020b, Lemma 2) Consider the single context case. With a slight abuse of notation, we drop the dependence on x and write $r(a)$ instead of $r(x,a)$, $\pi_\theta(a)$ instead of $\pi_\theta(a|x)$, and $\hat{r}(a)$ instead of $\hat{r}(x,a)$. Let π_θ be a softmax policy parameterized by θ. Then, for any $\hat{\boldsymbol{r}} \in [0,1]^K$, and any estimator $\hat{V}$ linear in π_θ, the mapping $\theta \mapsto \hat{V}(\pi_\theta) = \langle \hat{\boldsymbol{r}}, \pi_\theta \rangle$ is 5/2-smooth.*

Lemma 9. *All the action level estimators `EST` in (`IPS`, `cIPS`, `DR`, `PC`) can be written, for any policy π, in the form:*

$$\hat{V}_{\texttt{EST}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{A\sim\pi(\cdot|X_i)}\left[\hat{r}_{\texttt{EST},i}(A, X_i)\right] , \tag{D.1}$$

For the cluster level estimators/approaches `EST-C` in (`MIPS`, `OffCEM`, `POTEC`), we also have

$$\hat{V}_{\texttt{EST-C}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{C\sim\pi(\cdot|X_i)}\left[\hat{r}_{\texttt{EST-C},i}(C, X_i)\right] , \tag{D.2}$$

meaning that all these estimators are linear in π.

Proof. This is straightforward to prove. We begin by the action level estimators and take `DR` as a representative. For `DR`, we have the following:

$$\hat{r}_{\texttt{DR},i}(a, X_i) = \hat{r}(a, X_i) + \mathbb{I}[a = A_i]\frac{R_i - \hat{r}(A_i, X_i)}{\max(\tau, \pi_0(A_i|X_i))}$$

verifies the equation. Solutions for `cIPS` and `IPS` can be recovered directly, and `PC` follows the same construction. For the cluster level approaches, we take `POTEC` as a representative, and we have:

$$\hat{r}_{\texttt{POTEC},i}(c, X_i) = \hat{r}^\star_c(X_i) + \mathbb{I}[c = C_i]\frac{R_i - \hat{r}(A_i, X_i)}{\pi_0(C_i|X_i)} ,$$

The $\hat{r}_{\texttt{MIPS},i}$ follows as a special case when $\hat{r} = 0$. □

Lemma 10. *Consider the single-context case and assume a finite action set $\mathcal{A}$. For any estimator `EST` in (`IPS`, `cIPS`, `DR`, `OffCEM`, `MIPS`, `PC`), there exists a problem instance (i.e., a choice of r and π_0; and when relevant, a choice of auxiliary objects such as $\hat{r}$, h, N_ϵ) such that, in the large-n limit,*

$$\hat{r}_{\texttt{EST}}(a) = \mathbb{1}[a = a_K]$$

for some optimal action a_K. Similarly, for cluster-based approaches (e.g., `POTEC` and `MIPS`), there exists an instance such that

$$\hat{r}_{\texttt{EST-C}}(c) = \mathbb{1}[c = c_{|\mathcal{C}|}]$$

for some optimal cluster $c_{|\mathcal{C}|}$.

Proof. We give explicit constructions for `cIPS` (action-level) and `POTEC` (cluster-level). The other estimators follow by the same idea: choose a setting where the estimator becomes linear in π with some deterministic coefficient, and pick r (and possibly $\hat{r}$, h, N_ϵ) so that the resulting linearized reward is one-hot.

Action-level: `cIPS`. Fix $\tau \in [0, 1)$.[1] Choose a logging policy π_0 with full support and such that

$$\max_{a \in \mathcal{A}} \pi_0(a) \;\geq\; \tau, \tag{D.3}$$

Let

$$a_K \in \arg\max_{a \in \mathcal{A}} \frac{\pi_0(a)}{\max\{\pi_0(a), \tau\}}.$$

Under Equation (D.3), there exists at least one action with $\pi_0(a) \geq \tau$, for which the ratio equals 1, hence the maximizer satisfies $\pi_0(a_K) \geq \tau$ and therefore

$$\frac{\pi_0(a_K)}{\max\{\pi_0(a_K), \tau\}} = 1.$$

Now define the reward function

$$r(a) \;=\; \mathbb{1}[a = a_K] \frac{\max\{\pi_0(a), \tau\}}{\pi_0(a)}.$$

This satisfies $r(a) \in [0, 1]$ for all a because $r(a) = 0$ for $a \neq a_K$, and

$$r(a_K) = \frac{\max\{\pi_0(a_K), \tau\}}{\pi_0(a_K)} = 1 \quad (\text{since } \pi_0(a_K) \geq \tau).$$

For `cIPS`, the large-n linearized reward is

$$\hat{r}_{\text{CIPS}}(a) = \frac{\pi_0(a)}{\max\{\pi_0(a), \tau\}} r(a),$$

hence

$$\hat{r}_{\text{CIPS}}(a) = \frac{\pi_0(a)}{\max\{\pi_0(a), \tau\}} \mathbb{1}[a = a_K] \frac{\max\{\pi_0(a), \tau\}}{\pi_0(a)} = \mathbb{1}[a = a_K],$$

as desired.

Cluster-level: `POTEC`. We work in the single-context case and consider a clustering map $h : \mathcal{A} \to \mathcal{C}$. Choose h so that a_K forms a singleton cluster:

$$c_{|\mathcal{C}|} = h(a_K) = \{a_K\}, \qquad h(a) \neq c_{|\mathcal{C}|} \;\; \forall a \neq a_K.$$

Let rewards be $r(a_K) = 1$ and $r(a) = 0$ for $a \neq a_K$. Pick any $\varepsilon \in (0, 1/2]$ and define a reward model

$$\hat{r}(a_K) = 1 - \varepsilon, \qquad \hat{r}(a) = \varepsilon \;\; \forall a \neq a_K.$$

For `POTEC`, the induced (cluster-level) linearized reward takes the form

$$\hat{r}_{\texttt{POTEC}}(c) = \max_{a \in c} \hat{r}(a) + \frac{\sum_{a \in c} \pi_0(a) \big(r(a) - \hat{r}(a)\big)}{\pi_0(c)}.$$

[1] If $\tau = 1$ and $|\mathcal{A}| > 1$, the simplifying assumption $\pi_0(a) > 0$ for all a is incompatible with having some $\pi_0(a) \geq \tau$. In practice $\tau \ll 1$.

For the singleton cluster $c_{|\mathcal{C}|} = \{a_K\}$, we get

$$\hat{r}_{\texttt{POTEC}}(c_{|\mathcal{C}|}) = (1-\varepsilon) + \frac{\pi_0(a_K)\big(1-(1-\varepsilon)\big)}{\pi_0(a_K)} = (1-\varepsilon) + \varepsilon = 1.$$

For any other cluster $c \neq c_{|\mathcal{C}|}$, all its actions satisfy $r(a) = 0$ and $\hat{r}(a) = \varepsilon$, hence

$$\hat{r}_{\texttt{POTEC}}(c) = \varepsilon + \frac{\sum_{a\in c}\pi_0(a)(-\varepsilon)}{\pi_0(c)} = \varepsilon - \varepsilon = 0.$$

Therefore $\hat{r}_{\texttt{POTEC}}(c) = \mathbb{1}[c = c_{|\mathcal{C}|}]$.

This concludes the constructions for `cIPS` and `POTEC`. The remaining estimators can be handled analogously by choosing π_0 (and when relevant, h or N_ϵ) so that the estimator's linear coefficient on $r(a)$ equals 1 at a chosen a_K (and equals something finite elsewhere), and then defining r (and possibly $\hat{r}$) to make the resulting $\hat{r}_{\text{EST}}$ one-hot. □

Now we restate Proposition 3 and proceed to its proof.

Proposition 8 (Plateau for linear-in-π objectives under softmax)**.** *Consider the single-context case. Let $\hat{V}(\pi)$ be any objective linear in π and let $\pi^\star \in \arg\max_\pi \hat{V}(\pi)$ denote a maximizer over the probability simplex on the effective action space $\mathcal{A}_{\text{eff}}$ (of size $K_{\text{eff}} = |\mathcal{A}_{\text{eff}}|$). Let $\{\pi_{\theta_t}\}_{t\geq 1}$ be the iterates of gradient ascent on $\theta \mapsto \hat{V}(\pi_\theta)$ with a linear softmax policy $\pi_\theta(a) = \exp(\theta_a)/\sum_{a'\in\mathcal{A}_{\text{eff}}}\exp(\theta_{a'})$ and step sizes $\eta_t \in (0,1]$. Then there exists a problem instance such that gradient ascent cannot escape a suboptimal region before $t_0 = C\,K_{\text{eff}} = \mathcal{O}(K_{\text{eff}})$ iterations, in the sense that*

$$\forall t \leq t_0 : \qquad \hat{V}(\pi^\star) - \hat{V}(\pi_{\theta_t}) \geq 0.9.$$

Proof. The proof follows the same technique as (Mei et al., 2020a, Theorem 1). By Lemma 10, there exists an instance (single context) for which the linearized reward is one-hot:

$$\hat{r}_{\text{EST}}(a) = \mathbb{1}[a = a_K] \qquad \text{for some } a_K \in \mathcal{A}_{\text{eff}}.$$

Hence, for any policy π supported on $\mathcal{A}_{\text{eff}}$,

$$\hat{V}(\pi) = \sum_{a\in\mathcal{A}_{\text{eff}}} \pi(a)\hat{r}_{\text{EST}}(a) = \pi(a_K).$$

The maximizer over the simplex is therefore $\pi^\star = \delta_{a_K}$, and

$$\hat{V}(\pi^\star) = 1, \qquad \hat{V}(\pi^\star) - \hat{V}(\pi_\theta) = 1 - \pi_\theta(a_K).$$

(Notice that $\sup_\theta \hat{V}(\pi_\theta) = 1$ as well, although the supremum is not attained by any finite θ when $K_{\text{eff}} \geq 2$). We now upper bound the gradient norm. For the softmax parametrization,

$$\Big\|\nabla_\theta \hat{V}(\pi_\theta)\Big\|_2 \leq \sqrt{2}\,\pi_\theta(a_K)\big(1-\pi_\theta(a_K)\big),$$

where the bound follows by a direct computation (as in (Mei et al., 2020a)).

Define the update $\theta_{t+1} = \theta_t + \eta_t \nabla_\theta \hat{V}(\pi_{\theta_t})$ and split iterations into

$$t_{\text{good}} = \{t \geq 1 : \ \pi_{\theta_{t+1}}(a_K) > \pi_{\theta_t}(a_K)\}, \qquad t_{\text{bad}} = \{t \geq 1 : \ \pi_{\theta_{t+1}}(a_K) \leq \pi_{\theta_t}(a_K)\}.$$

For $t \in t_{\text{bad}}$,

$$\frac{1}{\pi_{\theta_t}(a_K)} - \frac{1}{\pi_{\theta_{t+1}}(a_K)} \leq 0.$$

For $t \in t_{\text{good}}$, using Lemma 8 (the 5/2-smoothness of $\theta \mapsto \hat{V}(\pi_\theta)$) and $\eta_t \in (0, 1]$, we obtain

$$\pi_{\theta_{t+1}}(a_K) - \pi_{\theta_t}(a_K) \leq \frac{9}{2}\, \pi_{\theta_t}(a_K)^2,$$

and therefore (since $\pi_{\theta_{t+1}}(a_K) \geq \pi_{\theta_t}(a_K) > 0$),

$$\frac{1}{\pi_{\theta_t}(a_K)} - \frac{1}{\pi_{\theta_{t+1}}(a_K)} = \frac{\pi_{\theta_{t+1}}(a_K) - \pi_{\theta_t}(a_K)}{\pi_{\theta_{t+1}}(a_K)\pi_{\theta_t}(a_K)} \leq \frac{9}{2}.$$

Summing over $s = 1, \ldots, t-1$ yields

$$\frac{1}{\pi_{\theta_1}(a_K)} - \frac{1}{\pi_{\theta_t}(a_K)} = \sum_{s=1}^{t-1} \left(\frac{1}{\pi_{\theta_s}(a_K)} - \frac{1}{\pi_{\theta_{s+1}}(a_K)} \right) \leq \frac{9}{2}\, t.$$

Assume a standard symmetric initialization so that $\pi_{\theta_1}(a_K) = 1/K_{\text{eff}}$. Pick any constant $c \geq 11$ and take K_{eff} large enough so that $\pi_{\theta_1}(a_K) \leq 1/c$. If $t \leq \frac{2}{9c}\, K_{\text{eff}}$, then

$$\frac{1}{\pi_{\theta_t}(a_K)} \geq \frac{1}{\pi_{\theta_1}(a_K)} - \frac{9}{2}\, t \geq \frac{1}{\pi_{\theta_1}(a_K)} \left(1 - \frac{1}{c}\right) \geq c - 1 \geq 10,$$

hence $\pi_{\theta_t}(a_K) \leq 1/10$, and thus

$$\hat{V}(\pi^\star) - \hat{V}(\pi_{\theta_t}) = 1 - \pi_{\theta_t}(a_K) \geq 0.9.$$

This proves the claim with $t_0 = \frac{2}{9c} K_{\text{eff}}$. □

Proposition 9. *Even for a single context x, deterministic rewards, there is problem where IPS-based learning with a linear softmax policy $\pi_\theta(a) \propto \exp(\langle \theta, \phi(x, a) \rangle) \mathbb{I}[a \in \mathcal{A}_{\text{eff}}]$ can have a number of local maxima exponential in the number of effective actions K_{eff}.*

Proof. Let `EST` an off-policy estimators considered in the paper with an action-level policy. By Lemma 9, we have:

$$\hat{V}_{\text{EST}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{a \sim \pi(\cdot|x_i)} \left[\hat{r}_{\text{EST},i}(a, x_i) \right], \tag{D.4}$$

In a single context setting, it becomes:

$$\hat{V}_{\text{EST}}(\pi_\theta) = \mathbb{E}_{a \sim \pi_\theta(\cdot)} \left[\frac{1}{n} \sum_{i=1}^{n} \hat{r}_{\text{EST},i}(a) \right], \tag{D.5}$$

$$= \langle \frac{1}{n} \sum_{i=1}^{n} \hat{r}_{\text{EST},i}, \pi_\theta \rangle \,. \tag{D.6}$$

This also holds for estimators with policies in the cluster level, as we still have:

$$\hat{V}_{\text{EST-C}}(\pi_\theta) = \mathbb{E}_{c\sim\pi_\theta(\cdot)}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{r}_{\text{EST-C},i}(c)\right], \tag{D.7}$$

$$= \langle\frac{1}{n}\sum_{i=1}^{n}\hat{r}_{\text{EST-C},i}, \pi_\theta\rangle\,. \tag{D.8}$$

These softmax policies are all defined on the effective action space $\mathcal{A}_{\text{eff}}$, be it a subset of the action space $\mathcal{A}$ or the discrete cluster space $\mathcal{C}$. Using the linearity of the objective, we can directly apply Theorem 1 from Chen et al. (2019) and obtain our result. □

Finally, we also restate Proposition 5, and provide its proof.

Proposition 10. *For an ℓ_2 regularized (substituting $\frac{\lambda}{2}||\theta||^2$, with $\lambda > 0$), linear softmax policy π_θ, the PWLL objective $\hat{U}^g(\pi_\theta)$ defined as:*

$$\hat{U}^g(\pi) = \frac{1}{n}\sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i))\log\pi(A_i \mid X_i)\,,$$

is λ-strongly concave. Without regularization, the objective is concave.

Proof. For any x and $a \in \mathcal{A}_{\text{eff}}(x)$, we have:

$$\pi_\theta(a|x) = \frac{\exp(\langle\theta, \phi(x,a)\rangle)}{\sum_{a'\in\mathcal{A}_{\text{eff}}(x)}\exp(\langle\theta, \phi(x,a')\rangle)}\,,$$

optimizing an ℓ_2 regularized linear softmax, giving:

$$\hat{L}^{g,\lambda}(\pi) = \hat{U}^g(\pi) - \frac{\lambda}{2}||\theta||^2\,,$$

with $\lambda > 0$ and recall that $g \geq 0$. For strong concavity, we need to show that the Hessian $\nabla^2_\theta\hat{U}^g(\pi_\theta)$ is negative definite with eigenvalues bounded away from zero.

The gradient with respect to θ is: $\nabla_\theta\hat{U}^g(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i))\nabla_\theta\log\pi_\theta(A_i|X_i) - \lambda\theta$

For the softmax policy:

$$\nabla_\theta\log\pi_\theta(a|x) = \phi(x,a) - \sum_{a'}\pi_\theta(a'|x)\phi(x,a') = \phi(x,a) - \mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)]$$

Therefore: $\nabla_\theta\hat{U}^g(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i))\left(\phi(X_i,A_i) - \mathbb{E}_{A\sim\pi_\theta(\cdot|X_i)}[\phi(X_i,A)]\right) - \lambda\theta$

Taking the second derivative: $\nabla^2_\theta\hat{U}^g(\pi_\theta) = -\frac{1}{n}\sum_{i=1}^{n} g(R_i, \pi_0(A_i \mid X_i))\nabla_\theta\mathbb{E}_{A\sim\pi_\theta(\cdot|X_i)}[\phi(X_i,A)] - \lambda I_d$, where I_d is the $d\times d$ identity matrix. The gradient of the expectation is:

$$\nabla_\theta\mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)] = \sum_a \nabla_\theta\pi_\theta(a|x)\phi(x,a)$$

Using $\nabla_\theta \pi_\theta(a|x) = \pi_\theta(a|x)(\phi(x,a) - \mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)])$:

$$\nabla_\theta \mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)] = \sum_a \pi_\theta(a|x)(\phi(x,a) - \mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)])\phi(x,a)^\top$$

This simplifies to:

$$\nabla_\theta \mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)] = \mathrm{Cov}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)]$$

where $\mathrm{Cov}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)] = \mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)\phi(x,A)^\top] - \mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)]\mathbb{E}_{A\sim\pi_\theta(\cdot|x)}[\phi(x,A)]^\top$

Therefore:

$$\nabla^2_\theta \hat{U}^{\mathtt{g}}(\pi_\theta) = -\frac{1}{n}\sum_{i=1}^n g(R_i, \pi_0(A_i \mid X_i))\mathrm{Cov}_{A\sim\pi_\theta(\cdot|X_i)}[\phi(X_i,A)] - \lambda I_d$$

We can write this as: $\nabla^2_\theta \hat{U}^{\mathtt{g}}(\pi_\theta) = -H - \lambda I_d$

where $H = \frac{1}{n}\sum_{i=1}^n g(R_i, \pi_0(A_i \mid X_i))\mathrm{Cov}_{A\sim\pi_\theta(\cdot|X_i)}[\phi(X_i,A)]$ is positive semi-definite. To see this explicitly, for any vector $v \in \mathbb{R}^d$:

$$v^\top \mathrm{Cov}_{A\sim\pi_\theta(\cdot|X_i)}[\phi(X_i,A)]v = \mathrm{Var}_{A\sim\pi_\theta(\cdot|X_i)}[v^\top\phi(X_i,A)] \geq 0\,,$$

with the positivity of g, this ensures H is positive semi-definite. Then we have:

$$v^\top \nabla^2_\theta \hat{U}^{\mathtt{g}}(\pi_\theta)v = -v^\top Hv - \lambda v^\top v = -v^\top Hv - \lambda\|v\|^2\,,$$

meaning that when $v \neq 0$, we get $v^\top \nabla^2_\theta \hat{U}^{\mathtt{g}}(\pi_\theta)v \leq -\lambda\|v\|^2 < 0\,.$

This shows the Hessian is negative definite with all eigenvalues bounded above by $-\lambda < 0$. Therefore, ℓ_2 regularized $\hat{U}^{\mathtt{g}}(\pi_\theta)$ is λ-strongly concave. In addition, when $\lambda = 0$, the hessian is negative semi-definite, giving simple concavity. □

D.3 Stochastic Optimization Convergence Guarantees for PWLL

We analyze the convergence rates of stochastic gradient methods on the PWLL objective. We formulate this as the minimization of the finite-sum loss $f(\theta) = -\hat{U}_g(\pi_\theta)$:

$$f(\theta) = \frac{1}{n}\sum_{i=1}^n f_i(\theta), \quad \text{where } f_i(\theta) = -g_i \log \pi_\theta(A_i \mid X_i), \tag{D.9}$$

where $g_i = g(R_i, \pi_0(A_i|X_i))$. We adopt the linear softmax policy parametrization in Equation (7.16) with $s_\theta(x,a) = \phi(x,a)^\top\theta$ (lightweight parametrization in Equation (7.17)). We note that our analysis extends naturally to the heavyweight parametrization in Equation (7.17).

D.3.1 Assumptions and Regularity

To establish problem-dependent convergence bounds, we rely on the following structural assumptions regarding the feature space and the importance weights.

Assumption 11 (Bounded features)**.** *For all context-action pairs $(x, a) \in \mathcal{X} \times \mathcal{A}$, the feature representations are bounded in Euclidean norm:*

$$\|\phi(x, a)\|_2 \leq H.$$

Assumption 12 (Bounded weighting function)**.** *The weights $g_i = g(R_i, \pi_0(A_i|X_i))$ computed on the static dataset are strictly positive and bounded. That is, for all $i \in \{1, \ldots, n\}$:*

$$0 < g_i \leq G_{\max}.$$

Assumptions 11 and 12 are sufficient to establish the smoothness and bounded variance of the objective $f(\theta)$. We formally derive these properties in the following proposition.

Proposition 11 (Regularity and Variance Bounds)**.** *Under Assumptions 11 and 12, the objective $f(\theta)$ satisfies the following properties:*

1. ***Global Smoothness:*** *The objective is $\bar{L}$-smooth with $\bar{L} = G_{\max} H^2$.*
2. ***Bounded Single-Sample Variance:*** *The variance of the stochastic gradient for a single sample is bounded by $\bar{\sigma}^2 = 4G_{\max}^2 H^2$.*
3. ***Bounded Mini-Batch Variance:*** *For a mini-batch of size b, the variance is bounded by $\bar{\sigma}_b^2 = \frac{4G_{\max}^2 H^2}{b}$.*

Proof. 1. Smoothness: The Hessian of the objective is the weighted sum of the feature covariance matrices under the policy π_θ:

$$\nabla^2 f(\theta) = \frac{1}{n} \sum_{i=1}^{n} g_i \mathrm{Cov}_{A \sim \pi_\theta(\cdot|X_i)}[\phi(X_i, A)].$$

The spectral norm of a covariance matrix is bounded by the maximum squared norm of its random vectors. Thus, using Assumption 11 we get that $\|\nabla^2 f(\theta)\|_{\mathrm{op}} \leq \frac{1}{n} \sum_{i=1}^{n} g_i H^2 \leq G_{\max} H^2$.

2. Single-Sample Variance: We first bound the norm of the gradient for an arbitrary sample i. The gradient is $\nabla f_i(\theta) = -g_i(\phi(X_i, A_i) - \mathbb{E}_{A \sim \pi_\theta(\cdot|X_i)}[\phi(X_i, A)])$. Using the triangle inequality and Assumption 11:

$$\|\nabla f_i(\theta)\|_2 \leq g_i \left(\|\phi(X_i, A_i)\|_2 + \|\mathbb{E}_{A \sim \pi_\theta(\cdot|X_i)}[\phi(X_i, A)]\|_2 \right) \leq G_{\max}(H + H) = 2G_{\max} H.$$

Let $\xi = \nabla f_I(\theta)$ be the stochastic gradient sampled uniformly from the dataset. The variance is bounded by the second moment:

$$\mathrm{Var}(\xi) \leq \mathbb{E}[\|\xi\|^2] = \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\theta)\|^2 \leq (2G_{\max} H)^2 = 4G_{\max}^2 H^2.$$

3. Mini-Batch Variance: Let the mini-batch gradient be $\bar{g}_t = \frac{1}{b}\sum_{j=1}^{b} \nabla f_{i_j}(\theta)$, where indices are sampled independently with replacement. Using the standard variance reduction property for independent variables:

$$\mathbb{E}[\|\bar{g}_t - \nabla f(\theta)\|^2] = \frac{1}{b}\mathbb{E}[\|\nabla f_I(\theta) - \nabla f(\theta)\|^2] \leq \frac{4G_{\max}^2 H^2}{b}.$$

□

Based on Proposition 11, we define the following global problem-dependent constants on which our convergence rates depend:

- $\bar{L} = G_{\max} H^2$: Smoothness constant.
- $\bar{\sigma}^2 = 4G_{\max}^2 H^2$: Upper bound on the gradient variance for a single sample.
- $\bar{\sigma}_b^2 = \frac{4G_{\max}^2 H^2}{b}$: Upper bound on the gradient variance for a mini-batch of size b.

D.3.2 PWLL without ℓ_2 regularization

We begin by analyzing the standard unregularized PWLL objective. Here, the objective $f(\theta)$ is convex but not necessarily strongly convex. This implies the loss landscape may contain multiple minimizers rather than a unique global minimum. Consequently, we characterize convergence in terms of $\hat{U}_g(\pi_{\theta^{\text{opt}}}) - \hat{U}_g(\pi_{\bar{\theta}_T})$ (instead of $\|\theta_t - \theta_n^{\text{opt}}\|$). Here, $\theta^{\text{opt}} \in \arg\max_\theta \hat{U}_g(\pi_\theta)$ is an optimal parameter and $\bar{\theta}_T$ is the average of the SGA iterates.

Proposition 12. *Let $\theta^{opt} \in \arg\max_\theta \hat{U}_g(\pi_\theta)$ be an optimal parameter. If the learning rate satisfies $0 < \eta \leq \frac{1}{4\bar{L}}$, then by (Garrigos and Gower, 2023, Theorem 6.9), the iterates of mini-batch SGA satisfy:*

$$\mathbb{E}\left[\hat{U}_g(\pi_{\theta^{opt}}) - \hat{U}_g(\pi_{\bar{\theta}_T})\right] \leq \frac{\|\theta_0 - \theta^{opt}\|^2}{\eta T} + \frac{8\eta G_{\max}^2 H^2}{b}$$

where $\bar{\theta}_T$ is the average of the iterates.

Proposition 12 highlights the trade-off inherent to constant step-size SGA: a larger η accelerates the decay of the initial error (first term) but increases the asymptotic noise floor (second term). For a fixed horizon T, one can recover a convergence rate of $\mathcal{O}(1/\sqrt{T})$ by setting $\eta \propto 1/\sqrt{T}$, which balances both terms.

D.3.3 PWLL with ℓ_2 regularization

We now move to the ℓ_2-regularized case where the PWLL objective is strongly concave (Proposition 5). Precisely, we consider the regularized objective $\tilde{U}^\lambda(\theta) = \hat{U}_g(\pi_\theta) - \frac{\lambda}{2}\|\theta\|^2$.

Strong convexity implies the existence of a unique global minimizer. This allows us to guarantee convergence of the parameters θ_t themselves, which is a stronger condition than value convergence.

Proposition 13. *Let* $\theta^{opt}_{n,\lambda} = \arg\max_\theta \tilde{U}^\lambda(\theta)$ *be the unique optimal parameter. If the learning rate satisfies* $0 < \eta \leq \frac{1}{2(G_{\max}H^2+\lambda)}$*, then by (Garrigos and Gower, 2023, Theorem 6.12):*

$$\mathbb{E}\left[\|\theta_t - \theta^{opt}_{n,\lambda}\|^2\right] \leq (1-\eta\lambda)^t\|\theta_0 - \theta^{opt}_{n,\lambda}\|^2 + \frac{8\eta G^2_{\max}H^2}{\lambda b}$$

The regularized case demonstrates a convergence rate that is significantly faster than the rate of the unregularized case.

D.4 Additional Experiments

D.4.1 Detailed Experimental Setting

Experimental Setting.

Table D.1: Statistics of Post Processed Datasets

Dataset	Num. of actions	Num. of samples
`MovieLens`	60,000	132,744
`Twitch`	200,000	400,000
`GoodReads`	1,000,000	400,000

Our experimental setup is designed to study the behavior of the different policy learning paradigms in large action spaces. To this end, we use three large action spaces collaborative filtering datasets: Movielens (Lam and Herlocker, 2016), Twitch (Rappaz et al., 2021) and GoodReads (Wan et al., 2019) that are preprocessed to obtain a user-item interaction matrix. We follow the exact procedure of Sakhi et al. (2023) to pre-process the datasets. The statistics of the obtained datasets are described in Table D.1. For each user, we keep half of its history as the context x, and use the other half of the history as the products with positive reward, which align the learned policies to recommend new and relevant items. We direct the interested readers to Sakhi et al. (2023) for a detailed description of the experimental setup.

The large action space scenario restricts the policies used to the inner product parametrization (Aouali et al., 2022a). This parametrization is essential to leverage Maximum Inner Product Search algorithms (Shrivastava and Li, 2014) for fast query response. In particular, we adopt policies of the following form:

$$\pi_\theta(a|x) \propto \exp(\langle \phi_\Gamma(x), \beta_a \rangle)\,,$$

with the learnable parameter $\theta = [\Gamma, \beta]$, $\phi_\Gamma : \mathcal{X} \to \mathbb{R}^\ell$ defines the context embedding function in $\mathbb{R}^\ell$ and β the actions embeddings of size $K \times \ell$. To define our policies, we start by extracting action embeddings β_0 using an SVD decomposition of the user-item matrix. These embeddings help us define the context embedding function ϕ_Γ and our

logging policy π_0. ϕ_0 is set to the average embeddings of the observed actions in the contexts and is fixed for the logging policy π_0. Using the SVD action embeddings β_0, we define our logging policy π_0 as:

$$\pi_0(a|x) \propto \exp\left(\frac{1}{t}\langle \phi_0(x), \beta_{0,a}\rangle\right) \mathbb{I}\left[a \in \textsc{top}^{k_0}(x)\right] ,$$

with t the temperature of the logging policy, and k_0 define the support of the logging policy, concentrating on the top k_0 actions with: $\textsc{top}^{k_0}(x) = \text{argsort}_{a_1,\cdots,a_{k_0}} \langle \phi_0(x), \beta_{0,a}\rangle$.

If not explicitly stated, k_0 is set to 100 and the temperature at $t = 1$ in all experiments. This policy is used to collect the offline dataset $\mathcal{D}_n = \{X_i, A_i, R_i\}_{i\in[n]}$ on which all trainings are conducted. For each $i \in [n]$ in the processed dataset, X_i is the user history, A_i is the action played by the logging policy $\pi_0(\cdot|X_i)$ and $R_i = \mathbb{1}[A_i \in H_i]$ the observed reward, which is if the action played is in the hidden items of user i.

Trained Policies Parameterizations. We adopt two parameterizations of the trained policies. The first one is a **heavyweight** parametrization, and focuses on learning the embeddings of the actions β (be it $\mathcal{A}$ of size K or $\mathcal{C}$ of size $|\mathcal{C}|$), meaning that θ in this case is β. For action-level policies, this gives $\beta \in \mathbb{R}^{K\times\ell}$ and for any x:

$$\pi_\beta(a|x) = \frac{\exp(\langle \phi_0(x), \beta_a\rangle)}{\sum_{a'\in\mathcal{A}_{\text{eff}}(x)} \exp(\langle \phi_0(x), \beta_{a'}\rangle)} ,$$

with $\mathcal{A}_{\text{eff}}(x) \subset \mathcal{A}$, which depends on the choice of the practitioner, for example $\mathcal{A}_{\text{eff}}(x) = S_0(x)$, the support of π_0 for context x when we optimize IPS objectives. For cluster-level policies, this gives a $\beta \in \mathbb{R}^{|\mathcal{C}|\times\ell}$ and for any x:

$$\pi_\beta(c|x) = \frac{\exp(\langle \phi_0(x), \beta_c\rangle)}{\sum_{c'\in\mathcal{C}} \exp(\langle \phi_0(x), \beta_{c'}\rangle)} .$$

This is used by default if nothing is explicitly stated.

We have also define a **lightweight** parametrization, where only a small projection $W \in \mathbb{R}^{\ell\times\ell}$ is learned, giving in action level policies:

$$\pi_W(a|x) = \frac{\exp(\langle \phi_0(x)W, \beta_{a,0}\rangle)}{\sum_{a'\in\mathcal{A}_{\text{eff}}(x)} \exp(\langle \phi_0(x)W, \beta_{a',0}\rangle)} ,$$

using β_0, the embeddings of π_0. For cluster level policies, we first define $\bar{\beta}_0 \in \mathbb{R}^{|\mathcal{C}|\times\ell}$ with $\bar{\beta}_{0,c} = \frac{1}{|c|}\sum_{a\in c} \beta_{0,a}$, and use it to define the cluster level policy:

$$\pi_W(c|x) = \frac{\exp(\langle \phi_0(x)W, \bar{\beta}_{c,0}\rangle)}{\sum_{c'\in\mathcal{C}} \exp(\langle \phi_0(x)W, \bar{\beta}_{c',0}\rangle)} .$$

Reward Model. The reward model used $\hat{r}$ is learned using regularized linear regression the collected interaction data, with $\hat{r}(x, a) = \langle \phi(x), \theta_a\rangle$.

Clustering and ϵ used. We use the embeddings β_0, combined with K-means clustering to find our clusters. The number of clusters is set to 2000 for all datasets and experiments. For `PC`, the ℓ_2 threshold ϵ is set to 0.1.

D.4.2 Additional results

Benefits of objective-aware parametrization. Figure D.1 shows the effect of objective-aware policy parameterizations for two different objectives and three large action space datasets.

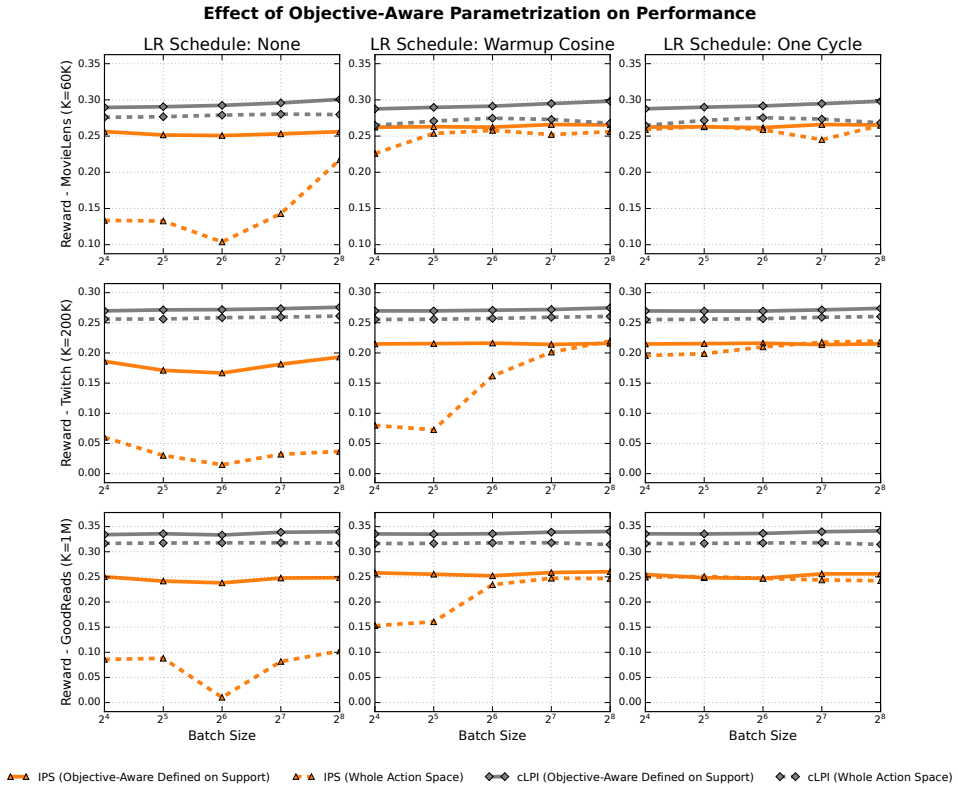


Figure D.1: The effect of objective-aware parametrization for `IPS` and `cLPI` on three large-scale datasets

Average MSE. Figure D.2 shows the average MSE by dataset and method. Several methods are excluded from the figure, as their high MSE values would distort the scale and obscure the comparison.

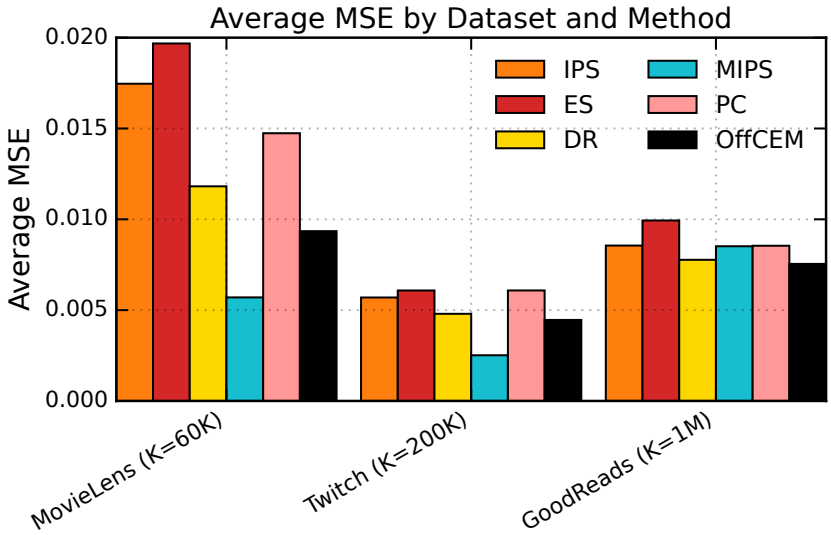
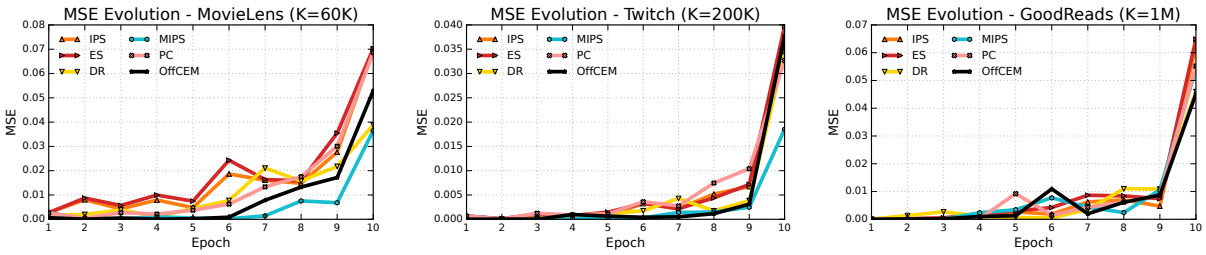


Figure D.2: Average MSE by Dataset and Method. Several methods are excluded from the figure, as their high MSE values would distort the scale and obscure the comparison.

MSE progress during training. Figures D.3a to D.3c show the progress of the MSE over 10 epochs on all three datasets. Several methods are excluded from the figure, as their high MSE values would distort the scale and obscure the comparison.



(a) `MovieLens` (b) `Twitch` (c) `GoodReads`

Figure D.3: MSE progression over 10 epochs across datasets.

D.4.3 Results Averaging Different Seeds

In this experiment, we analyze the reward evolution of representative PWLL and IPS-based methods on the three considered datasets. We compare two distinct optimization configurations: (i) a standard off-the-shelf Adam optimizer, and (ii) a carefully tuned setup using Adam with an optimized batch size and a one-cycle learning-rate scheduler. This comparison enables us to isolate the effect of optimization on stability and convergence. Each method is evaluated over 5 random seeds, and we report the mean reward along with a shaded standard deviation region to visualize sensitivity to optimization randomness.

In Figure D.4, across all datasets and optimization settings, we observe that IPS-based methods (`cIPS`, `IX`, and even `POTEC`) not only reach inferior performance but also suffer

from considerably higher variance. Their uncertainty bands are significantly wider, indicating unstable optimization. In contrast, PWLL-based methods exhibit near-invisible variance bands, with standard deviations roughly an order of magnitude smaller on average.

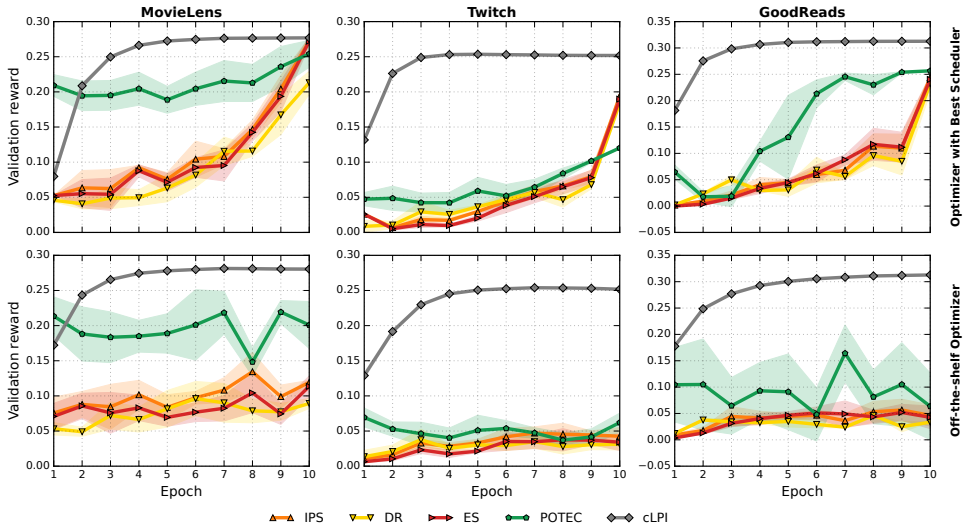


Figure D.4: `cLPI` vs IPS-Based methods: Evolution of rewards averaged over 5 different seeds. `cLPI` is more stable to optimize and reaches better policies.

Finally, in Figure D.5, we observe that adopting an Objective-Aware parametrization yields further performance and stability improvements. For example, `cIPS` with Objective-Aware parametrization surpasses `cIPS` while maintaining lower variability, and `cLPI` in its Objective-Aware form consistently achieves the best overall performance. These results demonstrate that the combination of PWLL objectives and clever parametrization leads to more robust and more effective learned policies, while being very simple to implement.

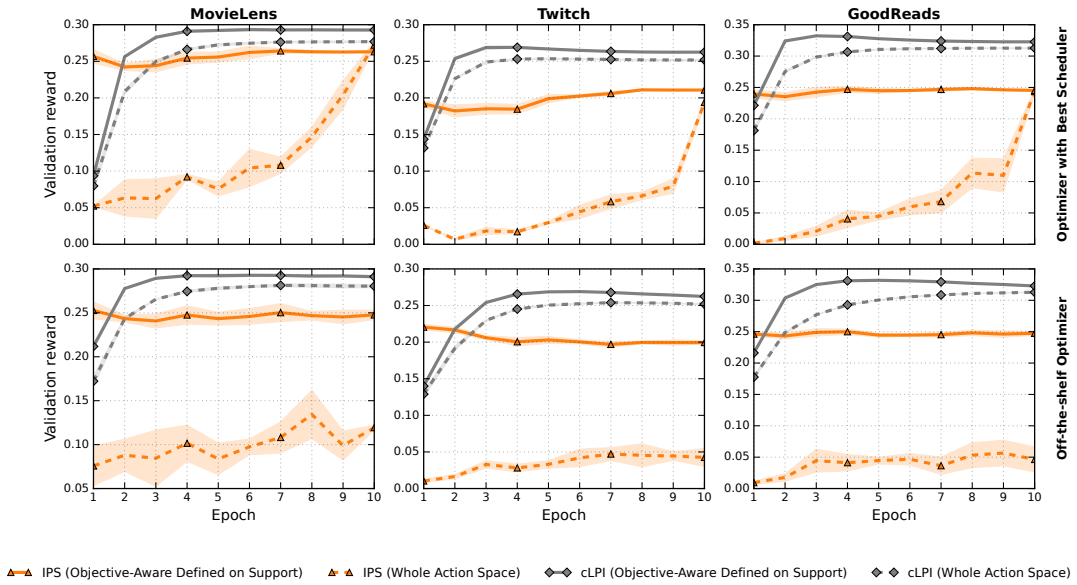


Figure D.5: Objective Aware Parametrisation: Evolution of rewards averaged over 5 different seeds. Objective Aware Parametrization stabilizes and improves performance for PWLL and IPS methods.

D.4.4 Ablation - Sensitivity to Reward Noise

In the original evaluation setup (see Section D.4), the observed reward is deterministic; for user i, we have $R_i = \mathbb{1}[A_i \in H_i]$, meaning that a positive reward is returned only when the selected action belongs to the user's hidden set H_i. In this section, we investigate robustness to reward noise by introducing stochasticity in the form:

$$R_i \sim \mathbb{1}[A_i \in H_i]\,(1 - B(\epsilon)) + B(\epsilon)\,s,$$

where $B(\epsilon)$ is a Bernoulli random variable with parameter ϵ, and $s \in [0, 1]$ is a shift. This results in noisy rewards supported on $[0, 1]$. Note that any reward scaling can be normalized to this range via $R/R_{\max}$ when $R_{\max} > 1$.

We evaluate six configurations defined by noise parameters $\epsilon \in \{0.1, 0.2, 0.3\}$ and reward shifts $s \in \{0, 0.5\}$. All methods are trained using the best-performing optimization schedule (one-cycle) to isolate the effect of noise. Results are reported in Figure D.6.

We observe that increasing the noise level when $s = 0$ consistently harms all methods, as expected from a more stochastic reward signal. In contrast, when $s = 0.5$, higher noise tends to increase the overall reward level, since the shift raises the baseline reward. Across all noise–shift conditions, PWLL-based objectives maintain a clear advantage over IPS-based methods. When $s = 0$, `RegKL` and `cLPI` perform similarly, confirming that both benefit from the logarithmic reparameterization. However, as both noise and shift increase, `RegKL` begins to outperform `cLPI`, suggesting that, with an appropriately chosen regularization weight β, `RegKL` remains highly competitive even under reward high stochasticity.

Conclusion. PWLL methods demonstrate robustness to reward noise, leading to improved stability and performance compared to traditional IPS-based objectives, even in

challenging noise regimes.

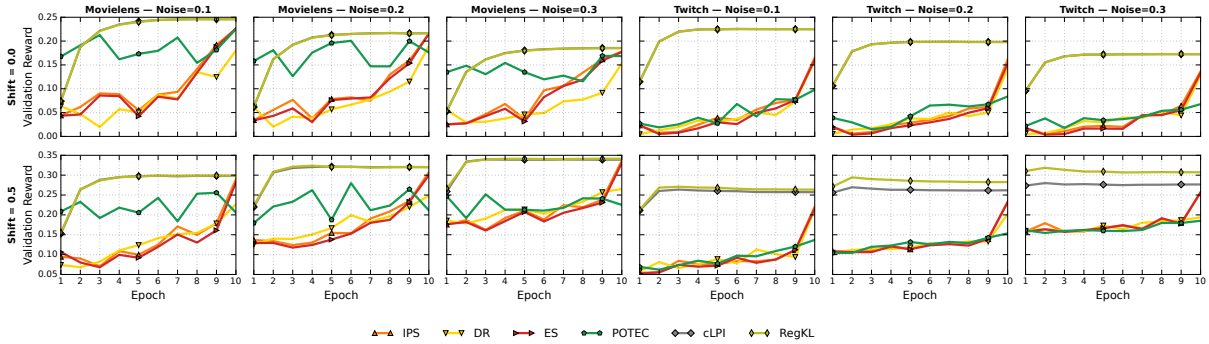


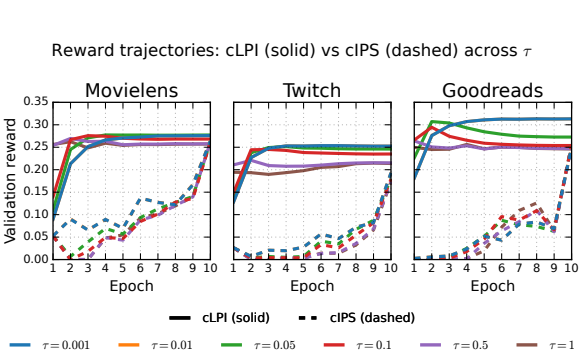
Figure D.6: Ablation - Sensitivity to reward noise

D.4.5 Ablation Study on Hyperparameters and Log Transform

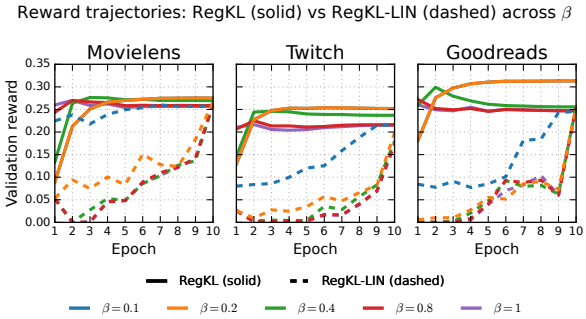
In this section, we evaluate the impact of hyperparameter choices on `cIPS`, `cLPI`, `RegKL`, and `RegKL-LIN` (the non-logarithmic variant of `RegKL`). All methods are run using the best-performing optimization configuration (optimizer + learning rate scheduler), ensuring that differences are driven solely by hyperparameter values and by whether the policy transformation is linear or logarithmic. The results are shown in Figure D.7.

- **cIPS** consistently fails to reach competitive performance across all values of τ, especially in large action spaces. Its PWLL counterpart, `cLPI`, dominates for every τ, converging faster and achieving superior results.
- For the KL-based objectives, we restrict to $\beta \geq 0.1$ in order to avoid numerical instability from the exponential term ($\exp(1/\beta) > 2 \cdot 10^5$ for $\beta < 0.1$). The same trend is observed: the PWLL variant (`RegKL`) reliably outperforms its linear analogue (`RegKL-LIN`) across all β, exhibiting more stable training dynamics, faster convergence and better performance.

PWLL dominates. Across both objective families, replacing linear weights with *log-transformed* policy weights (PWLL) consistently provides **greater robustness to hyperparameters**, **faster optimization**, and **higher final performance**, even more in challenging large-action-space settings.



(a) `cIPS` (linear) vs `cLPI` (PWLL) w.r.t τ



(b) `RegKL-LIN` (linear) vs `RegKL` (PWLL) w.r.t β

Figure D.7: Ablation Study on hyper-parameters and Log Transform

D.4.6 Ablation: PWLL in Smaller Action Spaces

We have shown that PWLL provides a more benign optimization landscape and yields stronger policies than IPS-based objectives in large action spaces. Here, we examine whether these benefits also extend to smaller action spaces. We construct a reduced version of `Movielens` by subsampling the action space to $K \in \{100, 500, 1000, 5000\}$ items.

Figure D.8 reports performance across varying K for `cIPS` (linear) and its PWLL-enhanced counterpart `cLPI` (log). In small action space settings ($K \leq 500$), `cIPS` convergences faster than `cLPI`, but `cLPI` identifies a better maxima by the end of the 10 epochs. For medium action spaces ($K \geq 500$), `cLPI` consistently outperforms `cIPS`, converging faster and identifying a better maximum. These results indicate that the optimization advantages of PWLL can still be beneficial in medium sized action space settings.

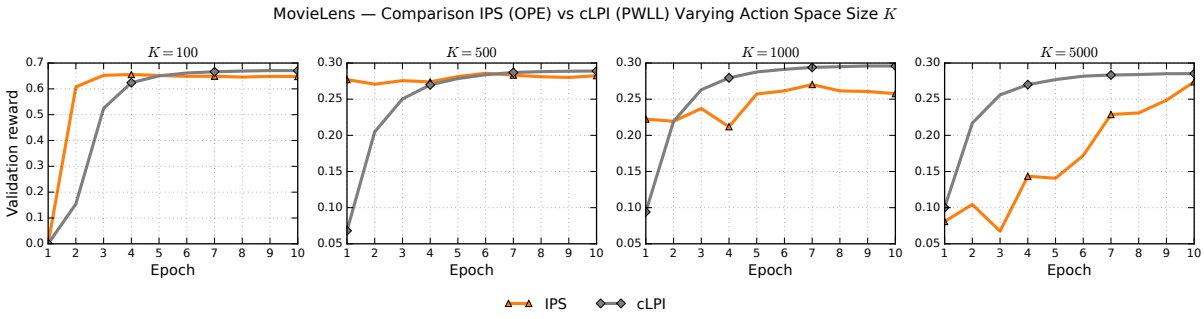


Figure D.8: PWLL (`cLPI`) vs IPS-based (`IPS`) in smaller action spaces.

D.4.7 Ablation - Sensitivity to the number of clusters

In this study, we compare our simple PWLL objective `cLPI` against `MIPS` and `POTEC`, two more complex IPS-based methods specifically designed for large action spaces. These baselines rely on a clustering function to reduce variance, and `POTEC` additionally leverages a reward model $\hat{r}$. We examine how the number of clusters affects their optimization performance. Figure D.9 reports the results.

POTEC generally outperforms MIPS for all numbers of clusters. However, both methods exhibit optimization instability across settings. While POTEC can occasionally match the final performance of cLPI on Movielens for a carefully selected number of clusters (1000), it consistently falls short on Twitch regardless of the cluster configuration.

Conclusion. These findings demonstrate that focusing on optimization properties pays off: despite its simplicity, the PWLL objective cLPI can consistently outperform intricate IPS-based approaches tailored to large action spaces, even with the best finetuning.

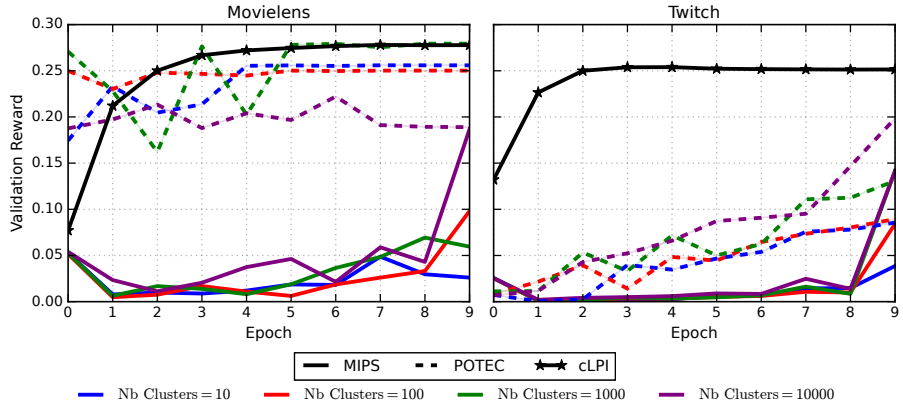


Figure D.9: PWLL (cLPI) vs POTEC and MIPS, changing the number of clusters.

D.4.8 Ablation - Different Logging Supports

We conduct experiments to quantify how increasing or restricting the support of the logging policy affects policy learning, comparing PWLL and IPS-based methods. Figure D.10 compiles the results and show that PWLL is still better than IPS-based approaches for different logging support sizes.

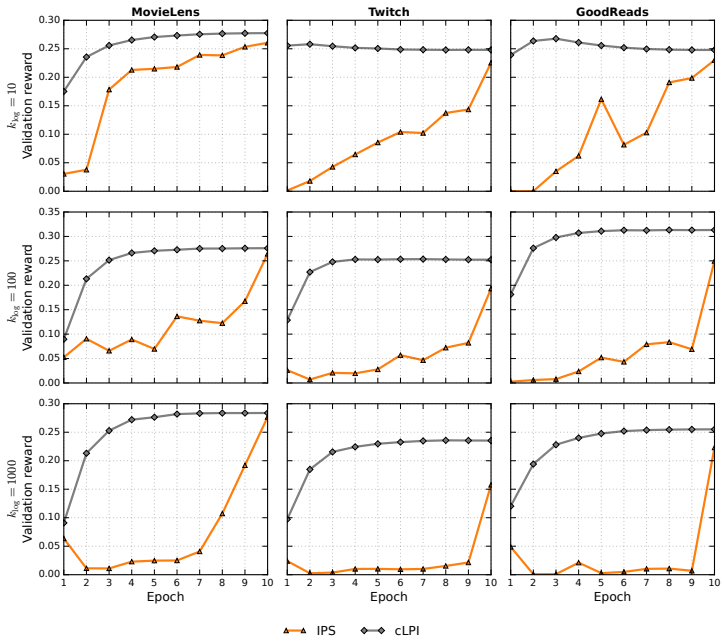


Figure D.10: PWLL vs IPS: Different Logging Support sizes $k_{\log}$

D.4.9 Pessimism Does not Solve Optimization Problems

Pessimism in face of uncertainty Jin et al. (2021) is motivated through a pure statistical learning rationale and is used to provide better statistical guarantees and more controlled excess risk. In the context of OPL, pessimistic strategies are derived combining concentration bounds with class complexity measures, be it VC dimension (Swaminathan and Joachims, 2015a) or PAC-Bayesian tools (London and Sandler, 2019; Aouali et al., 2023a; Sakhi et al., 2024). For example, in its PAC-Bayesian formulation, the pessimistic objectives are all written in the following form:

$$\arg\max_{\pi_\theta} \quad \hat{V}(\pi_\theta) - \frac{\lambda}{n}||\theta - \theta_0||_2^2\,,$$

Adding an ℓ_2 regularization term that pulls the parameters θ towards the behavior policy parameters θ_0 (defining π_0) induces pessimism by encouraging the learned policy to stay close to π_0 in parameter space. However, the optimization landscape of this objective becomes concave only when the regularization weight λ is sufficiently large for the ℓ_2 term to dominate. In that regime, the objective is indeed easier to optimize, but becomes overly conservative, yielding policies that remain too close to π_0 and under-exploit potential improvements. Figure D.11 confirms this empirically: pessimistic approaches, whether based on Sample Variance Penalisation (SVP) (Swaminathan and Joachims, 2015a), PAC-Bayesian learning with clipped IPS (London and Sandler, 2019), Exponential Smoothing (Aouali et al., 2023a), or Logarithmic Smoothing (Sakhi et al., 2024), fail to outperform `cLPI` for any value of λ.

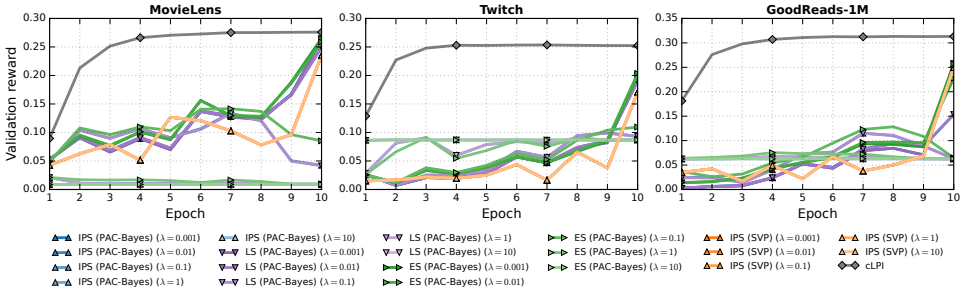


Figure D.11: `cLPI` outperforms the pessimistic approaches. Some methods do not appear in the plot because their curves overlap.

Chapter E

Supplementary Materials for Chapter 8

Contents

Notation Clarification: Value vs. Risk Formulation

Important: Throughout the main chapter, we present our results using the *value* formulation, where the goal is to maximize the expected reward $V(\pi) = \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}[r(X, A)]$

with rewards $R \in [0, 1]$. In contrast, the appendix uses the equivalent *risk* (or cost) formulation, where the goal is to minimize the expected cost $\mathcal{L}(\pi) = \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}[c(X, A)]$ with costs $C \in [-1, 0]$.

These two formulations are related by a simple sign change:

$$\mathcal{L}(\pi) = -V(\pi) \quad \text{and} \quad C = -R\,,$$

where $r(x, a) = -c(x, a)$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$. Consequently:

- Maximizing the value $V(\pi)$ is equivalent to minimizing the risk $\mathcal{L}(\pi)$.
- Upper bounds on $|V(\pi) - \hat{V}(\pi)|$ translate directly to upper bounds on $|\mathcal{L}(\pi) - \hat{\mathcal{L}}(\pi)|$.
- All theoretical guarantees derived in the appendix using the risk formulation apply equivalently to the value formulation presented in the corresponding chapter.

We adopt the risk formulation in the appendix as it aligns with the standard convention in statistical learning theory, where one typically minimizes a loss or risk function. The reader should keep this equivalence in mind when relating the appendix results to the main paper.

E.1 Bias and Variance Trade-Off

Notation reminder: This section uses the risk formulation with costs $C \in [-1, 0]$, which corresponds to the value formulation with rewards $R = -C \in [0, 1]$ in the main paper. The estimator $\hat{\mathcal{L}}_n^\alpha(\pi)$ here corresponds to $-\hat{V}^\alpha(\pi)$ in the main paper.

In this section, we provide additional results on how α controls the bias and variance of $\hat{\mathcal{L}}_n^\alpha(\cdot)$.

E.1.1 Bias and Variance of IPS-α

The following proposition states the bias-variance trade-off for $\hat{\mathcal{L}}_n^\alpha(\cdot)$.

Proposition 14 (Bias and variance of IPS-α)**.** *Let $\alpha \in [0, 1]$, the following holds for any evaluation policy $\pi \in \Pi$ that is absolutely continuous with respect to π_0*

$$|\mathbb{B}(\hat{\mathcal{L}}_n^\alpha(\pi))| \leq \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[1 - \pi_0(A|X)^{1-\alpha}\right]\,,$$
$$\mathbb{V}\left[\hat{\mathcal{L}}_n^\alpha(\pi)\right] \leq \frac{1}{n}\mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[\frac{\pi(A|X)}{\pi_0(A|X)^{2\alpha-1}}\right].$$

Proof. We first bound the bias as

$$\begin{aligned}
\mathbb{B}(\hat{\mathcal{L}}_n^\alpha(\pi)) &= \mathbb{E}\left[\hat{\mathcal{L}}_n^\alpha(\pi)\right] - \mathcal{L}(\pi)\,, \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{X_i\sim\nu, A_i\sim\pi_0(\cdot|X_i), C_i\sim p(\cdot|X_i,A_i)}\left[C_i \frac{\pi(A_i|X_i)}{\pi_0(A_i|X_i)^\alpha}\right] - \mathcal{L}(\pi)\,, \\
&\overset{(i)}{=} \mathbb{E}_{(X,A,C)\sim\mu_{\pi_0}}\left[C\frac{\pi(A|X)}{\pi_0(A|X)^\alpha}\right] - \mathcal{L}(\pi)\,, \\
&= \mathbb{E}_{X\sim\nu}\left[\sum_{a\in\mathcal{A}} c(X,a)\frac{\pi(a|X)}{\pi_0(a|X)^{\alpha-1}}\right] - \mathbb{E}_{X\sim\nu}\left[\sum_{a\in\mathcal{A}} c(X,a)\pi(a|X)\right]\,, \\
&= \mathbb{E}_{X\sim\nu}\left[\sum_{a\in\mathcal{A}} c(X,a)\pi(a|X)(\pi_0(a|X)^{1-\alpha}-1)\right]\,, \\
&= \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[c(X,A)(\pi_0(A|X)^{1-\alpha}-1)\right]\,,
\end{aligned}$$

where (i) follows from the i.i.d. assumption. Since $\pi_0(A|X)^{1-\alpha} \leq 1$ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$, we have that

$$\begin{aligned}
|\mathbb{B}(\hat{\mathcal{L}}_n^\alpha(\pi))| &\leq \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[|c(X,A)||\pi_0(A|X)^{1-\alpha}-1|\right]\,, \\
&\leq \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[1-\pi_0(A|X)^{1-\alpha}\right]\,.
\end{aligned}$$

The variance is bounded as

$$\begin{aligned}
\mathbb{V}\left[\hat{\mathcal{L}}_n^\alpha(\pi)\right] &= \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{V}_{X_i\sim\nu, A_i\sim\pi_0(\cdot|X_i), C_i\sim p(\cdot|X_i,A_i)}\left[C_i\frac{\pi(A_i|X_i)}{\pi_0(A_i|X_i)^\alpha}\right]\,, \\
&= \frac{1}{n}\mathbb{V}_{(X,A,C)\sim\mu_{\pi_0}}\left[C\frac{\pi(A|X)}{\pi_0(A|X)^\alpha}\right]\,, \\
&\leq \frac{1}{n}\mathbb{E}_{(X,A,C)\sim\mu_{\pi_0}}\left[C^2\frac{\pi(A|X)^2}{\pi_0(A|X)^{2\alpha}}\right]\,, \\
&\leq \frac{1}{n}\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\pi(A|X)^2}{\pi_0(A|X)^{2\alpha}}\right]\,, \\
&= \frac{1}{n}\mathbb{E}_{X\sim\nu}\left[\sum_{a\in\mathcal{A}}\frac{\pi(a|X)^2}{\pi_0(a|X)^{2\alpha-1}}\right]\,, \\
&= \frac{1}{n}\mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[\frac{\pi(A|X)}{\pi_0(A|X)^{2\alpha-1}}\right]\,.
\end{aligned}$$

□

E.2 Proofs for Off-Policy Learning

In this section, we provide the complete proofs for our OPL results in Section 8.3. We start with proving Theorem 4 in Section E.2.1. We then state the extension of Theorem 4

along with its proof in Section E.2.2. After that, in Section E.2.3, we provide the proof for Proposition 6. Finally, in Section E.2.4, we discuss in detail and prove our claims regarding the number of samples needed so that the performance of the learned policy is close to that of the optimal policy.

Notation: This section uses the risk formulation with $\mathcal{L}(\pi) = -V(\pi)$ and $\hat{\mathcal{L}}_n^\alpha(\pi) = -\hat{V}^\alpha(\pi)$. All bounds translate directly to the value formulation in the main paper. Recall that we assume the costs to be deterministic for simplicity: $C_i = c(X_i, A_i)$.

E.2.1 Proof of Theorem 4

In this section, we prove Theorem 4.

Proof. First, we decompose the difference $\mathcal{L}(\pi_\mathbb{Q}) - \hat{\mathcal{L}}_n^\alpha(\pi_\mathbb{Q})$ as

$$\mathcal{L}(\pi_\mathbb{Q}) - \hat{\mathcal{L}}_n^\alpha(\pi_\mathbb{Q}) = \underbrace{\mathcal{L}(\pi_\mathbb{Q}) - \frac{1}{n}\sum_{i=1}^n \mathcal{L}(\pi_\mathbb{Q}|X_i)}_{I_1} + \underbrace{\frac{1}{n}\sum_{i=1}^n \mathcal{L}(\pi_\mathbb{Q}|X_i) - \frac{1}{n}\sum_{i=1}^n \mathcal{L}^\alpha(\pi_\mathbb{Q}|X_i)}_{I_2} + \underbrace{\frac{1}{n}\sum_{i=1}^n \mathcal{L}^\alpha(\pi_\mathbb{Q}|X_i) - \hat{\mathcal{L}}_n^\alpha(\pi_\mathbb{Q})}_{I_3},$$

where

$$\begin{aligned}
\mathcal{L}(\pi_\mathbb{Q}) &= \mathbb{E}_{X\sim\nu, A\sim\pi_\mathbb{Q}(\cdot|X)}\left[c(X,A)\right], \\
\mathcal{L}(\pi_\mathbb{Q}|X_i) &= \mathbb{E}_{A\sim\pi_\mathbb{Q}(\cdot|X_i)}\left[c(X_i,A)\right], \\
\mathcal{L}^\alpha(\pi_\mathbb{Q}|X_i) &= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_\mathbb{Q}(A|X_i)}{\pi_0(A|X_i)^\alpha} c(X_i,A)\right], \\
\hat{\mathcal{L}}_n^\alpha(\pi) &= \frac{1}{n}\sum_{i=1}^n \frac{\pi(A_i|X_i)}{\pi_0(A_i|X_i)^\alpha} C_i .
\end{aligned}$$

Our goal is to bound $|\mathcal{L}(\pi_\mathbb{Q}) - \hat{\mathcal{L}}_n^\alpha(\pi_\mathbb{Q})|$ and thus we need to bound $|I_1| + |I_2| + |I_3|$. We start with $|I_1|$, Alquier (2021, Theorem 3.3) yields that following inequality holds with probability at least $1 - \delta/2$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$|I_1| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}}{2n}}. \tag{E.1}$$

Moreover, $|I_2|$ can be bounded by decomposing it as

$$\begin{aligned}
|I_2| &= \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{A \sim \pi_{\mathbb{Q}}(\cdot|X_i)} \left[c(X_i, A) \right] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{A \sim \pi_0(\cdot|X_i)} \left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0^{\alpha}(A|X_i)} c(X_i, A) \right] \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \pi_{\mathbb{Q}}(a|X_i) c(X_i, a) - \pi_0(a|X_i) \frac{\pi_{\mathbb{Q}}(a|X_i)}{\pi_0^{\alpha}(a|X_i)} c(X_i, a) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \left(\pi_{\mathbb{Q}}(a|X_i) - \frac{\pi_{\mathbb{Q}}(a|X_i)}{\pi_0^{\alpha-1}(a|X_i)} \right) c(X_i, a) \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \left(1 - \pi_0^{1-\alpha}(a|X_i) \right) \pi_{\mathbb{Q}}(a|X_i) c(X_i, a) \right| , \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \left| 1 - \pi_0^{1-\alpha}(a|X_i) \right| \pi_{\mathbb{Q}}(a|X_i) \left| c(X_i, a) \right| .
\end{aligned}$$

But $1 - \pi_0^{1-\alpha}(a|x) \geq 0$ and $|c(x, a)| \leq 1$ for any $a \in \mathcal{A}$ and $x \in \mathcal{X}$. Thus

$$|I_2| \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{A \sim \pi_{\mathbb{Q}}(\cdot|X_i)} \left[1 - \pi_0^{1-\alpha}(A|X_i) \right] . \tag{E.2}$$

Finally, we need to bound the main term $|I_3|$. To achieve this, we borrow the following technical lemma from Haddouche and Guedj (2022). It is slightly different from the one in Haddouche and Guedj (2022); their result holds for any $n \geq 1$ while we state a simpler version where n is fixed in advance.

Lemma 11. *Let $\mathcal{Z}$ be an instance space and let $S_n = (z_i)_{i \in [n]}$ be an n-sized dataset for some $n \geq 1$. Let $(\mathcal{F}_i)_{i \in \{0\} \cup [n]}$ be a filtration adapted to S_n. Also, let $\mathcal{H}$ be a hypothesis space and $(f_i(S_i, h))_{i \in [n]}$ be a martingale difference sequence for any $h \in \mathcal{H}$, that is for any $i \in [n]$, and $h \in \mathcal{H}$, we have that $\mathbb{E}\left[f_i(S_i, h) \,|\mathcal{F}_{i-1}\right] = 0$. Moreover, for any $h \in \mathcal{H}$, let $M_n(h) = \sum_{i=1}^{n} f_i(S_i, h)$. Then for any fixed prior, $\mathbb{P}$, on $\mathcal{H}$, any $\lambda > 0$, the following holds with probability $1 - \delta$ over the sample S_n, simultaneously for any $\mathbb{Q}$, on $\mathcal{H}$*

$$\left| \mathbb{E}_{h \sim \mathbb{Q}} \left[M_n(h) \right] \right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q} \| \mathbb{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2} \left(\mathbb{E}_{h \sim \mathbb{Q}} \left[\langle M \rangle_n(h) + [M]_n(h) \right] \right) ,$$

where $\langle M \rangle_n(h) = \sum_{i=1}^{n} \mathbb{E}\left[f_i(S_i, h)^2 \,|\mathcal{F}_{i-1}\right]$ and $[M]_n(h) = \sum_{i=1}^{n} f_i(S_i, h)^2$.

To apply Lemma 11, we need to construct an adequate martingale difference sequence $(f_i(S_i, h))_{i \in [n]}$ for $h \in \mathcal{H}$ that allows us to retrieve $|I_3|$. To achieve this, we define $S_n = (A_i)_{i \in [n]}$ as the set of n taken actions. Also, we let $(\mathcal{F}_i)_{i \in \{0\} \cup [n]}$ be a filtration adapted to S_n. For $h \in \mathcal{H}$, we define $f_i(S_i, h)$ as

$$f_i(S_i, h) = f_i(A_i, h) = \mathbb{E}_{A \sim \pi_0(\cdot|X_i)} \left[\frac{\mathbb{I}_{\{h(X_i) = A\}}}{\pi_0(A|X_i)^{\alpha}} c(X_i, A) \right] - \frac{\mathbb{I}_{\{h(X_i) = A_i\}}}{\pi_0(A_i|X_i)^{\alpha}} c(X_i, A_i) .$$

We stress that $f_i(S_i, h)$ only depends on the last action in S_i, A_i, and the predictor h. For this reason, we denote it by $f_i(A_i, h)$. The function f_i is indexed by i since it depends

on the fixed i-th context, X_i. The context X_i is fixed and thus randomness only comes from $A_i \sim \pi_0(\cdot|X_i)$. It follows that the expectations are under $A_i \sim \pi_0(\cdot|X_i)$. First, we have that $\mathbb{E}\left[f_i\left(A_i, h\right)|\mathcal{F}_{i-1}\right] = 0$ for any $i \in [n], h \in \mathcal{H}$. This follows from

$$\begin{aligned}
&\mathbb{E}\left[f_i\left(A_i, h\right)|\mathcal{F}_{i-1}\right] = \mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[f_i\left(A_i, h\right)\Big|A_1, \ldots, A_{i-1}\right], \\
&= \mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i, A)\right] - \frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i, A_i)\Big|A_1, \ldots, A_{i-1}\right], \\
&\overset{(i)}{=} \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i, A)\right] - \mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i, A_i)\Big|A_1, \ldots, A_{i-1}\right].
\end{aligned}$$

In (i) we use the fact that given X_i, $\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i, A)\right]$ is deterministic. Now A_i does not depend on $A_1, \ldots, A_{i-1}$ since logged data is i.i.d. Hence

$$\begin{aligned}
\mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i, A_i)\Big|A_1, \ldots, A_{i-1}\right] &= \mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i, A_i)\right], \\
&= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i, A)\right].
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{E}\left[f_i\left(A_i, h\right)|\mathcal{F}_{i-1}\right] \\
&= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i, A)\right] - \mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i, A_i)\Big|A_1, \ldots, A_{i-1}\right], \\
&= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i, A)\right] - \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i, A)\right], \\
&= 0\,.
\end{aligned}$$

Therefore, for any $h \in \mathcal{H}$, $(f_i(A_i, h))_{i\in[n]}$ is a martingale difference sequence. Hence we apply Lemma 11 and obtain that the following inequality holds with probability at least $1 - \delta/2$ for any $\mathbb{Q}$ on $\mathcal{H}$

$$\left|\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]\right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(4/\delta)}{\lambda} + \frac{\lambda}{2}\left(\mathbb{E}_{h\sim\mathbb{Q}}\left[\langle M\rangle_n(h) + [M]_n(h)\right]\right), \tag{E.3}$$

where

$$\begin{aligned}
M_n(h) &= \sum_{i=1}^n f_i\left(A_i, h\right), \\
\langle M\rangle_n(h) &= \sum_{i=1}^n \mathbb{E}\left[f_i\left(A_i, h\right)^2|\mathcal{F}_{i-1}\right], \\
[M]_n(h) &= \sum_{i=1}^n f_i\left(A_i, h\right)^2
\end{aligned}$$

Now these terms can be decomposed as

$$\begin{aligned}
\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right] &= \sum_{i=1}^{n}\mathbb{E}_{h\sim\mathbb{Q}}\left[f_i\left(A_i,h\right)\right], \\
&= \sum_{i=1}^{n}\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]-\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i)\right], \\
&\overset{(i)}{=} \sum_{i=1}^{n}\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]\right]-\mathbb{E}_{h\sim\mathbb{Q}}\left[\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i)\right], \\
&\overset{(ii)}{=} \sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(X_i)=A\}}\right]}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]-\frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(X_i)=A_i\}}\right]}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i), \\
&\overset{(iii)}{=} \sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]-\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i),
\end{aligned}$$

where we use the linearity of the expectation in both (i) and (ii). In (iii), we use our definition of policies in (8.10). Therefore, we have that

$$\begin{aligned}
\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right] &= \sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]-\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i), \\
&\overset{(i)}{=} \sum_{i=1}^{n}\mathcal{L}^\alpha(\pi_{\mathbb{Q}}|X_i)-n\hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}}), \\
&= nI_3,
\end{aligned} \tag{E.4}$$

where we used the fact that $C_i = c(X_i, A_i)$ for any $i \in [n]$ in (i).

Now we focus on the terms $\langle M\rangle_n(h)$ and $[M]_n(h)$. First, we have that

$$\begin{aligned}
f_i\left(A_i,h\right)^2 &= \left(\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]-\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i)\right)^2, \\
&= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]^2+\left(\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i)\right)^2 \\
&\qquad -2\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i), \\
&= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]^2+\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2 \\
&\qquad -2\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i).
\end{aligned} \tag{E.5}$$

Moreover, $f_i\left(A_i,h\right)^2$ does not depend on $A_1,\ldots,A_{i-1}$. Thus,

$$\begin{aligned}
\mathbb{E}\left[f_i\left(A_i,h\right)^2|\mathcal{F}_{i-1}\right] &= \mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[f_i\left(A_i,h\right)^2|\mathcal{F}_{i-1}\right], \\
&= \mathbb{E}_{A_i\sim\pi_0(\cdot|X_i)}\left[f_i\left(A_i,h\right)^2\right] = \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[f_i\left(A,h\right)^2\right].
\end{aligned}$$

Computing $\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[f_i\left(A,h\right)^2\right]$ using the decomposition in (E.5) yields

$$\begin{aligned}\mathbb{E}\left[f_i\left(A_i,h\right)^2|\mathcal{F}_{i-1}\right] &= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[f_i\left(A,h\right)^2\right],\\ &= -\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]^2+\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right] \qquad \text{(E.6)}\end{aligned}$$

Combining (E.5) and (E.6) leads to

$$\begin{aligned}\mathbb{E}\left[f_i\left(A_i,h\right)^2|\mathcal{F}_{i-1}\right]+f_i\left(A_i,h\right)^2 &= \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right]+\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2\\ &\quad-2\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^{\alpha}}c(X_i,A)\right]\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{\alpha}}c(X_i,A_i),\\ &\overset{(i)}{\leq}\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right]+\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2.\end{aligned} \qquad \text{(E.7)}$$

The inequality in (i) holds because $-2\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^\alpha}c(X_i,A)\right]\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha}c(X_i,A_i)\leq 0$. Therefore, we have that

$$\langle M\rangle_n(h)+[M]_n(h)\leq\sum_{i=1}^n\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{I}_{\{h(X_i)=A\}}}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right]+\frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2.$$

Finally, by using the linearity of the expectation and the definition of policies in (8.10), we get that

$$\begin{aligned}&\mathbb{E}_{h\sim\mathbb{Q}}\left[\langle M\rangle_n(h)+[M]_n(h)\right] \qquad \text{(E.8)}\\ &\leq\sum_{i=1}^n\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(X_i)=A\}}\right]}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right]+\frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{I}_{\{h(X_i)=A_i\}}\right]}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2,\\ &=\sum_{i=1}^n\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right]+\frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2. \qquad \text{(E.9)}\end{aligned}$$

Combining (E.3) and (E.8) yields

$$\begin{aligned}n|I_3| &= |\sum_{i=1}^n\mathcal{L}^\alpha(\pi_{\mathbb{Q}}|X_i)-n\hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}})|\\ &\leq\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log(4/\delta)}{\lambda}+\frac{\lambda}{2}\sum_{i=1}^n\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right]+\frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2.\end{aligned} \qquad \text{(E.10)}$$

This means that the following inequality holds with probability at least $1-\delta/2$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\begin{aligned}|I_3|\leq\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log(4/\delta)}{n\lambda}&+\frac{\lambda}{2n}\sum_{i=1}^n\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}c(X_i,A)^2\right]\\ &+\frac{\lambda}{2n}\sum_{i=1}^n\frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}c(X_i,A_i)^2.\end{aligned} \qquad \text{(E.11)}$$

However we know that $c(x,a)^2 \leq 1$ for any $x \in \mathcal{X}$ and $a \in \mathcal{A}$ and that $c(X_i, A_i) = C_i$ for any $i \in [n]$. Thus the following inequality holds with probability at least $1 - \delta/2$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$|I_3| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(4/\delta)}{n\lambda} + \frac{\lambda}{2n}\sum_{i=1}^{n} \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right] + \frac{\lambda}{2n}\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}C_i^2\,. \tag{E.12}$$

The union bound of (E.1) and (E.12) combined with the deterministic result in (E.2) yields that the following inequality holds with probability at least $1-\delta$ for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\begin{aligned}|\mathcal{L}(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}})| &\leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[1 - \pi_0^{1-\alpha}(A|X_i)\right] \\ &+ \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(4/\delta)}{n\lambda} + \frac{\lambda}{2n}\sum_{i=1}^{n} \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right] + \frac{\lambda}{2n}\sum_{i=1}^{n}\frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}C_i^2\,.\end{aligned} \tag{E.13}$$

□

E.2.2 Extensions of Theorem 4

Proposition 15 (Extension of Theorem 4 to hold simultaneously for any $\lambda \in (0,1)$)**.** *Let $n \geq 1$, $\delta \in [0,1]$, $\alpha \in [0,1]$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for any posterior $\mathbb{Q}$ on $\mathcal{H}$, and for any $\lambda \in (0,1)$ that*

$$|\mathcal{L}(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{KL'_1(\pi_{\mathbb{Q}},\lambda)}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{KL'_2(\pi_{\mathbb{Q}},\lambda)}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}})\,.$$

where

$$\begin{aligned}KL'_1(\pi_{\mathbb{Q}},\lambda) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\lambda}\,,\\ KL'_2(\pi_{\mathbb{Q}},\lambda) &= 2\big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\lambda}\big)\,,\\ B_n^{\alpha}(\pi_{\mathbb{Q}}) &= 1 - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[\pi_0^{1-\alpha}(A|X_i)\right]\,,\\ \operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}}C_i^2.\end{aligned}$$

Proof. Let $\delta \in (0,1)$. For any $i \geq 1$, we define $\lambda_i = 2^{-i}$ and let $\delta_i = \delta\lambda_i$. Then Theorem 4 yields that for any $i \geq 1$, the following inequality holds with probability at least $1-\delta_i$ for any $\mathbb{Q}$ on $\mathcal{H}$

$$|\mathcal{L}(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta_i}}{n\lambda_i} + \frac{\lambda_i}{2}\operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}})\,.$$

Now notice that $\sum_{i=1}^{\infty} \lambda_i = 1$, and hence $\sum_{i=1}^{\infty} \delta_i = \delta$. Therefore, the union bound of the above inequalities over $i \geq 1$ yields that with probability at least $1-\delta$, the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $i \geq 1$

$$|\mathcal{L}(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{4}{\delta_i}}{n\lambda_i} + \frac{\lambda_i}{2} \operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}}) \,. \tag{E.14}$$

Let $\lceil \cdot \rceil$ denote the ceiling function, then we have that for any $\lambda \in (0,1)$, there exists $j = \lceil \frac{-\log \lambda}{\log 2} \rceil \geq 1$ such that $\lambda/2 \leq \lambda_j \leq \lambda$. Since (E.14) holds for any $i \geq 1$, it holds in particular for j. In addition to this, we have that $\frac{1}{\lambda_j} \leq \frac{2}{\lambda}$, that $\lambda_j \leq \lambda$ and that $\frac{1}{\delta_j} = \frac{1}{\lambda_j \delta} \leq \frac{2}{\delta\lambda}$. This yields that the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $\lambda \in (0,1)$

$$|\mathcal{L}(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8\sqrt{n}}{\delta\lambda}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + 2\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8}{\delta\lambda}}{n\lambda} + \frac{\lambda}{2} \operatorname{Var}_n^{\alpha}(\pi_{\mathbb{Q}}) \,. \tag{E.15}$$

The additional 2 in $2\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8}{\delta\lambda}}{n\lambda}$ appears since we used that $\frac{1}{\lambda_j} \leq \frac{2}{\lambda}$. Similarly, the additional $\frac{2}{\lambda}$ in the logarithmic terms is due to the fact that $\frac{1}{\delta_j} \leq \frac{2}{\delta\lambda}$. Finally, setting

$$\begin{aligned} \mathrm{KL}'_1(\pi_{\mathbb{Q}}, \lambda) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8\sqrt{n}}{\delta\lambda} \,, \\ \mathrm{KL}'_2(\pi_{\mathbb{Q}}, \lambda) &= 2\big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8}{\delta\lambda}\big) \,, \end{aligned}$$

concludes the proof. □

Next, we provide a similar proof to extend Theorem 4 to any $\alpha \in (0,1]$. While we only provide a one-sided inequality, the same covering technique can be used to obtain the other side of the inequality.

Proposition 16 (One-sided extension of Theorem 4 to hold simultaneously for any $\alpha \in (0,1) \cup \{1\}$). *Let $n \geq 1$, $\delta \in [0,1]$, $\lambda > 0$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for any posterior $\mathbb{Q}$ on $\mathcal{H}$, and for any $\alpha \in (0,1]$ that*

$$\mathcal{L}(\pi_{\mathbb{Q}}) \leq \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}}) + \sqrt{\frac{KL''_1(\pi_{\mathbb{Q}}, \alpha)}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{KL''_2(\pi_{\mathbb{Q}}, \alpha)}{n\lambda} + \frac{\lambda}{2} \operatorname{Var}_n^{2\alpha}(\pi_{\mathbb{Q}}) \,.$$

where

$$
\begin{aligned}
\mathit{KL}''_1(\pi_{\mathbb{Q}}, \alpha) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8\sqrt{n}}{\delta\alpha}\,, \\
\mathit{KL}''_2(\pi_{\mathbb{Q}}, \alpha) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8}{\delta\alpha}\,, \\
B_n^{\alpha}(\pi_{\mathbb{Q}}) &= 1 - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[\pi_0^{1-\alpha}(A|X_i)\right]\,, \\
\mathrm{Var}_n^{\alpha}(\pi_{\mathbb{Q}}) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}} C_i^2.
\end{aligned}
$$

Proof. Let $\delta \in (0,1)$. For any $i \geq 0$, we define $\alpha_i = 2^{-i}$ and let $\delta_i = \delta\alpha_i/2$. Then Theorem 4 yields that for any $i \geq 0$, the following inequality holds with probability at least $1-\delta_i$ for any $\mathbb{Q}$ on $\mathcal{H}$

$$
|\mathcal{L}(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^{\alpha_i}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha_i}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta_i}}{n\lambda} + \frac{\lambda}{2}\mathrm{Var}_n^{\alpha_i}(\pi_{\mathbb{Q}})\,.
$$

Now notice that $\sum_{i=0}^{\infty}\alpha_i = 2$, and hence by definition of δ_i, we have $\sum_{i=0}^{\infty}\delta_i = \delta$. Therefore, the union bound of the above inequalities over $i \geq 0$ yields that with probability at least $1-\delta$, the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $i \geq 0$

$$
|\mathcal{L}(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^{\alpha_i}(\pi_{\mathbb{Q}})| \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta_i}}{2n}} + B_n^{\alpha_i}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta_i}}{n\lambda} + \frac{\lambda}{2}\mathrm{Var}_n^{\alpha_i}(\pi_{\mathbb{Q}})\,. \tag{E.16}
$$

Let $\lfloor\cdot\rfloor$ denote the floor function, then we have that for any $\alpha \in (0,1]$, there exists $j = \lfloor\frac{-\log\alpha}{\log 2}\rfloor \geq 0$ such that $\alpha \leq \alpha_j \leq 2\alpha$. Since (E.16) holds for any $i \geq 0$, it holds in particular for j. In addition, we have that $B_n^{\alpha}(\pi_{\mathbb{Q}})$ and $\hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}})$ are decreasing in α while $\mathrm{Var}_n^{\alpha}(\pi_{\mathbb{Q}})$ is increasing in α. Therefore, we have that $\hat{\mathcal{L}}_n^{\alpha_j}(\pi_{\mathbb{Q}}) \leq \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}})$, $B_n^{\alpha_j}(\pi_{\mathbb{Q}}) \leq B_n^{\alpha}(\pi_{\mathbb{Q}})$, and $\mathrm{Var}_n^{\alpha_j}(\pi_{\mathbb{Q}}) \leq \mathrm{Var}_n^{2\alpha}(\pi_{\mathbb{Q}})$. Moreover, we have that $\frac{1}{\delta_j} \leq \frac{2}{\delta\alpha}$. This yields that the following inequality holds with probability at least $1-\delta$ for any $\mathbb{Q}$ on $\mathcal{H}$ and for any $\alpha \in (0,1]$

$$
\mathcal{L}(\pi_{\mathbb{Q}}) \leq \hat{\mathcal{L}}_n^{\alpha}(\pi_{\mathbb{Q}}) + \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8\sqrt{n}}{\delta\alpha}}{2n}} + B_n^{\alpha}(\pi_{\mathbb{Q}}) + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{8}{\delta\alpha}}{n\lambda} + \frac{\lambda}{2}\mathrm{Var}_n^{2\alpha}(\pi_{\mathbb{Q}})\,. \tag{E.17}
$$

Finally, setting

$$
\begin{aligned}
\mathrm{KL}''_1(\pi_{\mathbb{Q}}, \alpha) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8\sqrt{n}}{\delta\alpha}\,, \\
\mathrm{KL}''_2(\pi_{\mathbb{Q}}, \alpha) &= D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log \frac{8}{\delta\alpha}\,,
\end{aligned}
$$

concludes the proof. □

E.2.3 Proof of Proposition 6

Haddouche and Guedj (2022, Theorem 7) provides an application of Lemma 11 to the general PAC-Bayes learning problems in Section 8.3.1. We cannot apply their theorem directly to get Proposition 6 for two reasons. They assume that the loss function is non-negative and they derive a one-sided generalization bound. In our case, the loss function is negative and we want to derive a two-sided generalization bound. Fortunately, we show with a slight modification of their proof that the result can be extended to two-sided inequalities with negative losses. In fact, the only requirement is that the sign of loss is fixed. We show next how this is achieved.

Proof. First, note that Lemma 11 does not make any assumption on the sign of the martingale difference sequence $(f_i(S_i,h))_{i\in[n]}$ nor on the sign of the terms that decompose it. Now similarly to the proof in Section E.2.1, we define $S_n = (X_i, A_i)_{i\in[n]}$ as the set of n observed contexts and taken actions. Also, we let $(\mathcal{F}_i)_{i\in\{0\}\cup[n]}$ be a filtration adapted to S_n. For $h \in \mathcal{H}$, we define $f_i\left(S_i, h\right)$ as

$$\begin{aligned} f_i\left(S_i, h\right) &= f_i\left(X_i, A_i, h\right) = f\left(X_i, A_i, h\right) , \\ &= \mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\mathbb{I}_{\{h(X)=A\}}}{\pi_0(A|X)^\alpha} c(X,A)\right] - \frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^\alpha} c(X_i, A_i) . \end{aligned}$$

Here $f_i(S_i,h)$ only depends on the last samples X_i, A_i and the predictor h. For this reason, we denote it by $f_i\left(X_i, A_i, h\right)$. Also, the function f_i does not depend on i and this is why we simplify the notation as $f_i\left(X_i, A_i, h\right) = f\left(X_i, A_i, h\right)$. Moreover, the randomness in $f\left(X_i, A_i, h\right)$ is only due $X_i \sim \nu$ and $A_i \sim \pi_0(\cdot|X_i)$; all other terms are deterministic. Thus the expectations are under $X_i \sim \nu, A_i \sim \pi_0(\cdot|X_i)$. Now similarly to the proof in Section E.2.1, we have that $\mathbb{E}\left[f\left(X_i, A_i, h\right)|\mathcal{F}_{i-1}\right] = 0$ for any $i \in [n], h \in \mathcal{H}$. Therefore, $(f(X_i, A_i, h))_{i\in[n]}$ is a martingale difference sequence for any $h \in \mathcal{H}$. Thus we apply Lemma 11 and get that that with probability at least $1-\delta$, the following holds simultaneously for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\left|\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]\right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\left(\mathbb{E}_{h\sim\mathbb{Q}}\left[\langle M\rangle_n(h) + [M]_n(h)\right]\right) , \tag{E.18}$$

where

$$\begin{aligned} M_n(h) &= \sum_{i=1}^n f\left(X_i, A_i, h\right) , \\ \langle M\rangle_n(h) &= \sum_{i=1}^n \mathbb{E}\left[f\left(X_i, A_i, h\right)^2 |\mathcal{F}_{i-1}\right] , \\ [M]_n(h) &= \sum_{i=1}^n f\left(X_i, A_i, h\right)^2 . \end{aligned}$$

Now we compute $\mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right]$ as

$$\begin{aligned} \mathbb{E}_{h\sim\mathbb{Q}}\left[M_n(h)\right] &= \sum_{i=1}^n \mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\pi_{\mathbb{Q}}(A|X)}{\pi_0(A|X)^\alpha} c(X,A)\right] - \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^\alpha} c(X_i, A_i) , \\ &= n\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\pi_{\mathbb{Q}}(A|X)}{\pi_0(A|X)^\alpha} c(X,A)\right] - \sum_{i=1}^n \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^\alpha} c(X_i, A_i) , \end{aligned} \tag{E.19}$$

where we used the linearity of the expectation $\mathbb{E}_{h\sim\mathbb{Q}}[\cdot]$ and the definition of policies in (8.10). Moreover, similarly to the proof in Section E.2.1, we have that

$$\begin{aligned}
\langle M\rangle_n(h) + [M]_n(h) &= \sum_{i=1}^{n} \mathbb{E}\left[f\left(X_i, A_i, h\right)^2 | \mathcal{F}_{i-1}\right] + f\left(X_i, A_i, h\right)^2 \\
&= \sum_{i=1}^{n} \mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\mathbb{I}_{\{h(X)=A\}}}{\pi_0(A|X)^{2\alpha}} c(X,A)^2\right] + \frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{2\alpha}} c(X_i,A_i)^2 \\
&\qquad - 2\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\mathbb{I}_{\{h(X)=A\}}}{\pi_0(A|X)^{\alpha}} c(X,A)\right] \frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{\alpha}} c(X_i,A_i)\,, \\
&\overset{(i)}{\leq} n\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\mathbb{I}_{\{h(X)=A\}}}{\pi_0(A|X)^{2\alpha}} c(X,A)^2\right] + \sum_{i=1}^{n} \frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{2\alpha}} c(X_i,A_i)^2\,,
\end{aligned} \tag{E.20}$$

where (i) holds since $-2\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\mathbb{I}_{\{h(X)=A\}}}{\pi_0(A|X)^{\alpha}} c(X,A)\right] \frac{\mathbb{I}_{\{h(X_i)=A_i\}}}{\pi_0(A_i|X_i)^{\alpha}} c(X_i,A_i) \leq 0$ for any $i \in [n]$. This is where the non-negative loss assumption is not needed. Our loss $L_\alpha(h,x,a,c) = \frac{\mathbb{I}_{\{h(X)=A\}}}{\pi_0(a|x)^\alpha} c$ is negative since $c \in [-1,0]$. However, we only need the product between the loss and its expectation to be non-negative. This holds in particular when the loss has a fixed sign. In that case, the expectation of the loss and the loss itself will have the same sign and thus their product will be non-negative. In our case, the loss has a fixed negative sign and this is all we needed. Now notice that

$$\begin{aligned}
n\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\pi_{\mathbb{Q}}(A|X)}{\pi_0(A|X)^{\alpha}} c(X,A)\right] &= nR^\alpha(\pi_{\mathbb{Q}})\,, \\
\sum_{i=1}^{n} \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{\alpha}} c(X_i,A_i) &= n\hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}})\,,
\end{aligned}$$

where we used that $c(X_i, A_i) = C_i$ for any $i \in [n]$ in the second equality. Using these two equalities and plugging (E.19) and (E.20) in (E.18) yields that with probability at least $1-\delta$, the following holds simultaneously for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\begin{aligned}
n\left|\mathcal{L}^\alpha(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}})\right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(2/\delta)}{\lambda} + \frac{\lambda}{2}\Bigg(&n\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\pi_{\mathbb{Q}}(A|X)}{\pi_0(A|X)^{2\alpha}} c(X,A)^2\right] \\
&+ \sum_{i=1}^{n} \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}} c(X_i,A_i)^2\Bigg)\,.
\end{aligned} \tag{E.21}$$

Again we used the linearity of the expectation $\mathbb{E}_{h\sim\mathbb{Q}}[\cdot]$ and the definition of policies in (8.10). Finally, we have that $c(X_i, A_i) = C_i$ for any $i \in [n]$. Thus with probability at least $1-\delta$ the following inequality holds for any distribution $\mathbb{Q}$ on $\mathcal{H}$

$$\begin{aligned}
\left|\mathcal{L}^\alpha(\pi_{\mathbb{Q}}) - \hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}})\right| \leq \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log(2/\delta)}{n\lambda} + \frac{\lambda}{2}&\mathbb{E}_{X\sim\nu, A\sim\pi_0(\cdot|X)}\left[\frac{\pi_{\mathbb{Q}}(A|X)}{\pi_0(A|X)^{2\alpha}} c(X,A)^2\right] \\
&+ \frac{\lambda}{2n}\sum_{i=1}^{n} \frac{\pi_{\mathbb{Q}}(A_i|X_i)}{\pi_0(A_i|X_i)^{2\alpha}} C_i^2\,.
\end{aligned} \tag{E.22}$$

This concludes the proof. □

E.2.4 Sample Complexity

Notation reminder: Minimizing risk $\mathcal{L}(\pi)$ is equivalent to maximizing value $V(\pi) = -\mathcal{L}(\pi)$. Achieving $\mathcal{L}(\hat{\pi}) \leq \mathcal{L}(\pi_*) + \epsilon$ is equivalent to $V(\hat{\pi}) \geq V(\pi_*) - \epsilon$.

Proposition 17. *Let $\mathcal{M}_1(\mathcal{H})$ be the set of probability distributions on the hypothesis space $\mathcal{H}$, and let $\lambda > 0$, $n \geq 1$, $\delta \in [0,1]$, $\alpha \in [0,1]$, and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, we have*

$$\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) \leq \mathcal{L}(\pi_{\mathbb{Q}_*}) + 2\sqrt{\frac{KL_1(\pi_{\mathbb{Q}_*})}{2n}} + 2B_n^\alpha(\pi_{\mathbb{Q}_*}) + 2\frac{KL_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \lambda \operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}_*})\,.$$

where $\pi_{\hat{\mathbb{Q}}_n}$ is the learned policy with $\hat{\mathbb{Q}}_n = \operatorname{argmin}_{\mathbb{Q}\in\mathcal{M}_1(\mathcal{H})} \hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}}) + \sqrt{\frac{KL_1(\pi_{\mathbb{Q}})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}}) + \frac{KL_2(\pi_{\mathbb{Q}})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}})$, $\mathbb{Q}_* = \operatorname{argmin}_{\mathbb{Q}\in\mathcal{M}_1(\mathcal{H})} \mathcal{L}(\pi_{\mathbb{Q}})$, and

$$KL_1(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4\sqrt{n}}{\delta}\,, \qquad KL_2(\pi_{\mathbb{Q}}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \log\frac{4}{\delta}\,,$$
$$B_n^\alpha(\pi_{\mathbb{Q}}) = 1 - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[\pi_0^{1-\alpha}(A|X_i)\right]\,,$$
$$\operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{A\sim\pi_0(\cdot|X_i)}\left[\frac{\pi_{\mathbb{Q}}(A|X_i)}{\pi_0(A|X_i)^{2\alpha}}\right] + \frac{\pi_{\mathbb{Q}}(A_i|X_i)C_i^2}{\pi_0(A_i|X_i)^{2\alpha}}\,.$$

Proof. First, Theorem 4 holds for any potentially data dependent distribution $\mathbb{Q}$ on $\mathcal{H}$. In particular, we have that with probability at least $1-\delta$ the following inequalities hold simultaneously for $\hat{\mathbb{Q}}_n$ and $\mathbb{Q}_*$

$$|\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) - \hat{\mathcal{L}}_n^\alpha(\pi_{\hat{\mathbb{Q}}_n})| \leq \sqrt{\frac{\mathrm{KL}_1(\pi_{\hat{\mathbb{Q}}_n})}{2n}} + B_n^\alpha(\pi_{\hat{\mathbb{Q}}_n}) + \frac{\mathrm{KL}_2(\pi_{\hat{\mathbb{Q}}_n})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\hat{\mathbb{Q}}_n})\,,$$
$$|\mathcal{L}(\pi_{\mathbb{Q}_*}) - \hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}_*})| \leq \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}_*})\,.$$

Taking only one side of these inequalities yields that with probability at least $1-\delta$ the following inequalities hold simultaneously for $\hat{\mathbb{Q}}_n$ and $\mathbb{Q}_*$

$$\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) \leq \underbrace{\hat{\mathcal{L}}_n^\alpha(\pi_{\hat{\mathbb{Q}}_n}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\hat{\mathbb{Q}}_n})}{2n}} + B_n^\alpha(\pi_{\hat{\mathbb{Q}}_n}) + \frac{\mathrm{KL}_2(\pi_{\hat{\mathbb{Q}}_n})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\hat{\mathbb{Q}}_n})}_{(I)}\,,$$
$$\hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}_*}) \leq \mathcal{L}(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}_*})\,.$$

Now using the definition of $\pi_{\hat{\mathbb{Q}}_n}$, we know that

$$I \leq \hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}_*})\,.$$

This yields that with probability at least $1-\delta$ the following inequalities hold simultaneously for $\hat{\mathbb{Q}}_n$ and $\mathbb{Q}_*$

$$\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) \leq \hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}_*})\,,$$
$$\hat{\mathcal{L}}_n^\alpha(\pi_{\mathbb{Q}_*}) \leq \mathcal{L}(\pi_{\mathbb{Q}_*}) + \sqrt{\frac{\mathrm{KL}_1(\pi_{\mathbb{Q}_*})}{2n}} + B_n^\alpha(\pi_{\mathbb{Q}_*}) + \frac{\mathrm{KL}_2(\pi_{\mathbb{Q}_*})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_{\mathbb{Q}_*})\,.$$

Computing the sum of these two inequalities concludes the proof. □

Corollary 2 (Special case of Proposition 17)**.** *Let $\mathcal{H} = \left\{h_\theta\,;\theta \in \mathbb{R}^{dK}\right\}$ of mappings $h_\theta(x) = \operatorname{argmax}_{a\in\mathcal{A}} \phi(x)^\top \theta_a$ for any $x \in \mathcal{X}$. Let $n \geq 1$, $\delta \in [0,1]$, $\alpha \in [0,1]$, and let $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, we have that*

$$\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) \leq \mathcal{L}(\pi_{\mathbb{Q}_*}) + \frac{\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}}}{\sqrt{n}} + 2(1-K^{\alpha-1}) + \frac{\|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta}}{\sqrt{n}} + \frac{K^{2\alpha-1}+K^{2\alpha}}{\sqrt{n}}\,.$$

where $\pi_{\hat{\mathbb{Q}}_n}$ is the learned policy with $\hat{\mathbb{Q}}_n = \operatorname{argmin}_{\mathbb{Q}=\mathcal{N}(\mu,I_{dK})} \hat{\mathcal{L}}_n^\alpha(\pi_\mathbb{Q}) + \sqrt{\frac{\mathrm{KL}_1(\pi_\mathbb{Q})}{2n}} + B_n^\alpha(\pi_\mathbb{Q}) + \frac{\mathrm{KL}_2(\pi_\mathbb{Q})}{n\lambda} + \frac{\lambda}{2}\operatorname{Var}_n^\alpha(\pi_\mathbb{Q})$, $\mathbb{Q}_* = \operatorname{argmin}_{\mathbb{Q}=\mathcal{N}(\mu,I_{dK})} \mathcal{L}(\pi_\mathbb{Q})$.

Proof. This result follows from the general Proposition 17 by simply setting $\mathbb{P} = \mathcal{N}(\mu_0, I_{dK})$ and $\mathbb{Q}_* = \mathcal{N}(\mu_*, I_{dK})$. First, since the covariance matrices of both distributions are I_{dK}, their KL divergence is $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) = \|\mu_* - \mu_0\|^2/2$. Moreover, since the logging policy is uniform then $B_n^\alpha(\pi_\mathbb{Q}) = (1-K^{\alpha-1})$ and $\operatorname{Var}_n^\alpha(\pi_\mathbb{Q}) \leq K^{2\alpha-1}+K^{2\alpha}$. Using these quantities, setting $\lambda = 1/\sqrt{n}$ and applying Proposition 17 yields that with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, we have that

$$\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) \leq \mathcal{L}(\pi_{\mathbb{Q}_*}) + \frac{\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}}}{\sqrt{n}} + 2(1-K^{\alpha-1}) + \frac{\|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta}}{\sqrt{n}} + \frac{K^{2\alpha-1}+K^{2\alpha}}{\sqrt{n}}\,.$$

This concludes the proof. □

The above corollary allows us to give insights into the sample complexity of our procedure. That is, the number of samples needed so that the performance of the learned policy $\pi_{\hat{\mathbb{Q}}_n}$ is close to that of the optimal one. Let $\epsilon > 2(1-K^{\alpha-1})$ for $\alpha \in [1-\log 2/\log K, 1]$. This condition on α ensures that $\epsilon \in [0,1]$ and it is mild as α is often close to 1. Let δ, then the following implication holds

$$\epsilon \geq \frac{\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}}}{\sqrt{n}} + 2(1-K^{\alpha-1}) + \frac{\|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta}}{\sqrt{n}} + \frac{K^{2\alpha-1}+K^{2\alpha}}{\sqrt{n}}$$
$$\implies \mathbb{P}(\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) \leq \mathcal{L}(\pi_{\mathbb{Q}_*}) + \epsilon) \geq 1-\delta\,. \tag{E.23}$$

First, we use that $\sqrt{\|\mu_* - \mu_0\|^2 + 2\log\frac{4\sqrt{n}}{\delta}} \le \|\mu_* - \mu_0\| + \sqrt{2\log\frac{4\sqrt{n}}{\delta}}$. Moreover we bound $K^{2\alpha-1} + K^{2\alpha} \le 2K^{2\alpha}$. Then the implication in (E.23) becomes

$$\sqrt{n} \ge \frac{\|\mu_* - \mu_0\| + \|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta} + \sqrt{2\log\frac{4\sqrt{n}}{\delta}} + 2K^{2\alpha}}{\epsilon - 2(1 - K^{\alpha-1})} \implies \mathbb{P}(\mathcal{L}(\pi_{\hat{\mathbb{Q}}_n}) \le \mathcal{L}(\pi_{\mathbb{Q}_*}) + \epsilon) \ge 1 - \delta\,. \tag{E.24}$$

We only provide intuition on the sample complexity and aim at having easy-to-interpret terms. Thus we omit the logarithmic terms in (E.24) and assume that $\|\mu_* - \mu_0\|^2 \ge \|\mu_* - \mu_0\|$. This leads to the claim made in Section 8.4.1. Of course, a more precise sample complexity analysis can be made by studying the function $h(x) = \sqrt{x} - \sqrt{2\log\frac{4\sqrt{x}}{\delta}}/(\epsilon - 2(1 - K^{\alpha-1}))$ and finding x such that $f(x) \ge \frac{\|\mu_* - \mu_0\| + \|\mu_* - \mu_0\|^2 + 2\log\frac{4}{\delta} + 2K^{2\alpha}}{\epsilon - 2(1-K^{\alpha-1})}$.

E.3 Experiments

E.3.1 Setup

We consider the standard supervised-to-bandit conversion (Agarwal et al., 2014). Precisely, let $\mathcal{S}_n^{\text{TR}}$ and $\mathcal{S}_{n_{\text{TS}}}^{\text{TS}}$ be the training and testing set of a classification dataset, respectively. First, we transform the training set $\mathcal{S}_n^{\text{TR}}$ to a logged bandit data $\mathcal{D}_n$ as described in Algorithm 3. The resulting logged data $\mathcal{D}_n$ is then used to train our policies. After that, the learned policies are tested on $\mathcal{S}_{n_{\text{TS}}}^{\text{TS}}$ as described in Algorithm 4. We consider that the resulting reward in Algorithm 4 is a good proxy for the unknown true reward of the learned policies. This will be our performance metric, the higher the better.

In our experiments, we use the following image classification datasets `MNIST` (LeCun et al., 1998), `FashionMNIST` (Xiao et al., 2017), `EMNIST` (Cohen et al., 2017) and `CIFAR100` (Krizhevsky et al., 2009). We provide a summary of the statistics of these datasets in Table E.1. Algorithm 3 takes as input a logging policy π_0 which we define as

$$\pi_0(a|x) = \frac{\exp(\eta_0 \cdot \phi(x)^\top \mu_{0,a})}{\sum_{a' \in \mathcal{A}} \exp(\eta_0 \cdot \phi(x)^\top \mu_{0,a'})}\,, \qquad \forall (x, a) \in \mathcal{X} \times \mathcal{A}\,. \tag{E.25}$$

Here $\phi(x) \in \mathbb{R}^d$ is the feature transformation function that outputs a d-dimensional vector, $\mu_0 = (\mu_{0,a})_{a \in \mathcal{A}} \in \mathbb{R}^{dK}$ are learnable parameters and η_0 is an inverse-temperature parameter for the `softmax` in (E.25). We explain next how these quantities are derived in detail.

The feature transformation function $\phi(x) \in \mathbb{R}^d$: for all the datasets, except `CIFAR100`, the feature transformation function $\phi(\cdot)$ is defined as $\phi(x) = \frac{x}{\|x\|}$ for any $x \in \mathcal{X}$. That is, we simply normalize the features $x \in \mathcal{X}$ by their L_2 norm $\|x\|$. In contrast, `CIFAR100` is a more challenging problem. Thus we use transfer learning to extract features $\phi(x)$ expressive enough so that a linear `softmax` model would enjoy a reasonable performance. Precisely, we retrieve the last hidden layer of a `ResNet-50` network, pre-trained on the ImageNet dataset, to output 2048-dimensional features. Finally, the obtained features

Table E.1: Statistics of the datasets used in our experiments.

Data set	Nbr. train samples n	Nbr. test samples n_{TS}	Nbr. actions K	Dimension d
MNIST	60000	10000	10	784
FashionMNIST	60000	10000	10	784
EMNIST	112800	18800	47	784
CIFAR100	50000	10000	100	2048

are normalized as $\frac{x}{\|x\|}$ and this whole process (ResNet-50 + normalization) corresponds to $\phi(\cdot)$ for CIFAR100.

The parameters $\mu_0 = (\mu_{0,a})_{a\in\mathcal{A}} \in \mathbb{R}^{dK}$: we learn the parameters μ_0 using 5% of the training set $\mathcal{S}_n^{\text{TR}}$. Precisely, we use the cross-entropy loss with an L_2 regularization of 10^{-6} to prevent the logging policy π_0 from being degenerate. This ensures that the learning policies are absolutely continuous with respect to the logging policy π_0, a condition under which standard IPS is unbiased. In optimization, we use Adam (Kingma and Ba, 2014) with a learning rate of 0.1 for 10 epochs. In all the experiments, we set the prior $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for the Gaussian policies in (8.13) and we set it as $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK}) \times \mathrm{G}(0,1)^K$ for the mixed-logit policies in (8.12). Our theory requires that the prior does not depend on data. Given that μ_0 is learned on the 5% portion of data, we only train our learning policies on the remaining 95% portion of the data to match our theoretical requirements.

The inverse-temperature parameter $\eta_0 \in \mathbb{R}$: this controls the performance of the logging policy. A high positive value of η_0 leads to a well-performing logging policy, while a negative one leads to a low-performing logging policy. When $\eta_0 = 0$, π_0 is identical to the uniform policy. In our experiments η_0 varies between 0 and 1.

Algorithm 3 Supervised-to-bandit: creating logged data

Input: training classification set $\mathcal{S}_n^{\text{TR}} = \{(X_i, y_i)\}_{i=1}^n$, logging policy π_0.
Output: logged bandit data $\mathcal{D}_n = (X_i, A_i, C_i)_{i\in[n]}$.
Initialize $\mathcal{D}_n = \{\}$
for $i = 1, \ldots, n$ **do**
  $A_i \sim \pi_0(\cdot|X_i)$
  $C_i = -\mathbb{I}_{\{A_i=y_i\}}$
  $\mathcal{D}_n \leftarrow \mathcal{D}_n \cup \{(X_i, A_i, C_i)\}$.

Algorithm 4 Supervised-to-bandit: testing policies

Input: image classification dataset $\mathcal{S}_{n_{\text{TS}}}^{\text{TS}} = \{(X_i, y_i)\}_{i=1}^{n_{\text{TS}}}$, learned policy $\hat{\pi}_n$.
Output: reward r.
for $i = 1, \ldots, n_{TS}$ **do**
  $A_i \sim \hat{\pi}_n(\cdot|X_i)$
  $R_i = \mathbb{I}_{\{A_i=y_i\}}$
$r = \frac{1}{n_{\text{TS}}} \sum_{i=1}^{n_{\text{TS}}} R_i$.

Now it remains to explain the learning policies $\pi_{\mathbb{Q}}$ and the corresponding closed-form bounds using either our results or those in existing works (London and Sandler, 2019;

Sakhi et al., 2022).

E.3.2 Policies

Here we present the two families of policies that we use in our experiments, Gaussian and mixed-logit policies.

Mixed-Logit

Let $\mathcal{H} = \left\{h_{\theta,\gamma}\,; \theta \in \mathbb{R}^{dK}, \gamma \in \mathbb{R}^K\right\}$ be a hypothesis space of mappings $h_{\theta,\gamma}(x) = \operatorname{argmax}_{a\in\mathcal{A}} \phi(x)^\top \theta_a + \gamma_a$ for any $x \in \mathcal{X}$. Here $\phi(x)$ outputs a d-dimensional representation of context $x \in \mathcal{X}$. Now assume that for any $a \in \mathcal{A}$, γ_a is a standard Gumbel perturbation, $\gamma_a \sim \mathrm{G}(0,1)$, then we have that

$$\begin{aligned}\pi_\theta^{\text{SOF}}(a|x) &= \frac{\exp(\phi(x)^\top \theta_a)}{\sum_{a'\in\mathcal{A}} \exp(\phi(x)^\top \theta_{a'})}\,, \\ &= \mathbb{E}_{\gamma\sim \mathrm{G}(0,1)^K}\left[\mathbb{I}_{\{h_{\theta,\gamma}(x)=a\}}\right]\,. \end{aligned} \tag{E.26}$$

In addition, we randomize θ such as $\theta \sim \mathcal{N}(\mu, \sigma^2 I_{dK})$ where $\mu \in \mathbb{R}^{dK}$ and $\sigma > 0$. It follows that the posterior $\mathbb{Q}$ is a multivariate Gaussian $\mathcal{N}(\mu, \sigma^2 I_{dK})$ over the parameters θ with standard Gumbel perturbations $\gamma \sim \mathrm{G}(0,1)^K$. We denote such policies by $\pi_{\mu,\sigma}^{\text{MIXL}}$ and they are defined as

$$\begin{aligned}\pi_{\mu,\sigma}^{\text{MIXL}}(a|x) &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}\left[\frac{\exp(\phi(x)^\top \theta_a)}{\sum_{a'\in\mathcal{A}} \exp(\phi(x)^\top \theta_{a'})}\right]\,, \\ &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}\left[\pi_\theta^{\text{SOF}}(a|x)\right]\,, \\ &= \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK}),\gamma\sim\mathrm{G}(0,1)^K}\left[\mathbb{I}_{\{h_{\theta,\gamma}(x)=a\}}\right]\,. \end{aligned} \tag{E.27}$$

To sample from the mixed-logit policies $\pi_{\mu,\sigma}^{\text{MIXL}}$, we first sample $\theta \sim \mathcal{N}(\mu, \sigma^2 I_{dK})$ and $\gamma \sim \mathrm{G}(0,1)^K$ and then set the sampled action as $a \leftarrow h_{\theta,\gamma}(x)$. Now we also need to compute the gradient of the expectation in (E.27). This needs additional care since the distribution under which we take the expectation depends on the parameters μ, σ. Fortunately, the reparameterization trick can be used in this case. Roughly speaking, it allows us to express a gradient of the expectation in (E.27) as an expectation of a gradient. In our case, we use the *local* reparameterizaton trick (Kingma et al., 2015) which is known for reducing the variance of stochastic gradients. Precisely, we rewrite (E.27) as

$$\begin{aligned}\pi_{\mu,\sigma}^{\text{MIXL}}(a|x) &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\|\phi(x)\|^2 I_K)}\left[\frac{\exp(\phi(x)^\top \mu_a + \sigma\epsilon_a)}{\sum_{a'\in\mathcal{A}} \exp(\phi(x)^\top \mu_{a'} + \sigma\epsilon_{a'})}\right]\,. \\ &= \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I_K)}\left[\frac{\exp(\phi(x)^\top \mu_a + \sigma\epsilon_a)}{\sum_{a'\in\mathcal{A}} \exp(\phi(x)^\top \mu_{a'} + \sigma\epsilon_{a'})}\right]\,, \end{aligned}$$

where we used that $\|\phi(x)\|^2 = 1$ since features are normalized. It follows that gradients read

$$\nabla_{\mu,\sigma}\pi_{\mu,\sigma}^{\text{MIXL}}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I_K)}\left[\nabla_{\mu,\sigma}\frac{\exp(\phi(x)^\top \mu_a + \sigma\epsilon_a)}{\sum_{a'\in\mathcal{A}} \exp(\phi(x)^\top \mu_{a'} + \sigma\epsilon_{a'})}\right]\,.$$

Moreover, the propensities are approximated as

$$\pi^{\text{MIXL}}_{\mu,\sigma}(a|x) \approx \frac{1}{S}\sum_{i\in[S]} \frac{\exp(\phi(x)^\top \mu_a + \sigma\epsilon_{i,a})}{\sum_{a'\in\mathcal{A}} \exp(\phi(x)^\top \mu_{a'} + \sigma\epsilon_{i,a'})}\,, \quad \epsilon_i \sim \mathcal{N}(0, I_K)\,, \forall i \in [S]\,. \tag{E.28}$$

In all our experiments, we set $S = 32$.

Gaussian

We define the hypothesis space $\mathcal{H} = \left\{h_\theta\,; \theta \in \mathbb{R}^{dK}\right\}$ of mappings $h_\theta(x) = \text{argmax}_{a\in\mathcal{A}}\, \phi(x)^\top \theta_a$ for any $x \in \mathcal{X}$. It follows that the learning policies $\pi_{\mathbb{Q}} = \pi^{\text{GAUS}}_{\mu,\sigma}$ read

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\theta\sim\mathcal{N}(\mu,\sigma^2 I_{dK})}\left[\mathbb{I}_{\{h_\theta(x)=a\}}\right]\,. \tag{E.29}$$

To see why this can be beneficial (Sakhi et al., 2022), let π_* be the optimal policy. Given $x \in \mathcal{X}$, $\pi_*(\cdot|x)$ should be deterministic; it chooses the best action for context x with probability 1. That is, there exists $\mu_* \in \mathbb{R}^{dK}$ such that $\pi_* = \mathbb{I}_{\{h_{\mu_*}(x)=a\}}$. When $\mu \to \mu_*$ and $\sigma \to 0$, the Gaussian policy in (E.29) approaches π_*. In contrast, the mixed-logit policy in (E.27) approaches $\pi^{\text{SOF}}_{\mu_*}$. However, $\pi^{\text{SOF}}_{\mu_*}$ is not deterministic due to the additional randomness in γ and is equal to π_* only if $\phi(x)^\top \mu_{*,a_*(x)} \to \infty$. This explains the choice of removing the Gumbel noise. First, Sakhi et al. (2022) showed that (E.29) can be written as

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\Big[\prod_{a'\neq a} \Phi\big(\epsilon + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma\|\phi(x)\|}\big)\Big]\,,$$

where Φ is the cumulative distribution function of a standard normal variable. But $\|\phi(x)\| = 1$ in all our experiments. Thus

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\Big[\prod_{a'\neq a} \Phi\big(\epsilon + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma}\big)\Big]\,.$$

Then similarly to mixed-logit policies, the gradient reads

$$\nabla_{\mu,\sigma}\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\Big[\nabla_{\mu,\sigma}\prod_{a'\neq a} \Phi\big(\epsilon + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma}\big)\Big]\,.$$

Moreover, the propensities are approximated as

$$\pi^{\text{GAUS}}_{\mu,\sigma}(a|x) \approx \frac{1}{S}\sum_{i\in[S]}\prod_{a'\neq a} \Phi\big(\epsilon_i + \frac{\phi(x)^\top(\mu_a - \mu_{a'})}{\sigma}\big)\,, \quad \epsilon_i \sim \mathcal{N}(0,1)\,, \forall i \in [S]\,. \tag{E.30}$$

In all our experiments, we set $S = 32$.

E.3.3 Baselines

Here we present all the methods that we use in our experiments. For each method, we state the result that holds for any learning policy π. After that, we derive the corresponding

closed-form bounds for Gaussian and mixed-logit policies that we presented previously. All the baselines require computing the KL divergence between the prior $\mathbb{P}$ and the posterior $\mathbb{Q}$. Thus before presenting them, we state the following lemma that allows bounding the KL divergence between the prior $\mathbb{P}$ and the posterior $\mathbb{Q}$ in the cases of mixed-logit or Gaussian policies.

Lemma 12 (KL divergence for Gaussian distributions with Gumbel noise)**.** *For distributions* $\mathbb{P} = \mathcal{N}(\mu_0, \sigma_0^2 I_{dK}) \times \mathrm{G}(0,1)^K$ *and* $\mathbb{Q} = \mathcal{N}(\mu, \sigma^2 I_{dK}) \times \mathrm{G}(0,1)^K$*, with* $\mu_0, \mu \in \mathbb{R}^{dK}$ *and* $0 < \sigma^2 \leq \sigma_0^2 < \infty$*,*

$$D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) \leq \frac{\|\mu - \mu_0\|^2}{2\sigma_0^2} + \frac{dK}{2} \log \frac{\sigma_0^2}{\sigma^2}\,.$$

Moreover, this result holds when the Gumbel noise is removed. That is when $\mathbb{P} = \mathcal{N}(\mu_0, \sigma_0^2 I_{dK})$ *and* $\mathbb{Q} = \mathcal{N}(\mu, \sigma^2 I_{dK})$*.*

We borrow this lemma from London and Sandler (2019). In particular, Lemma 12 shows that the KL terms for both policies can be bounded by the same quantity. As a result, the corresponding bounds will be the same; the only difference is the space of learning policies on which we optimize. For completeness, however, we write these bounds for both types of policies although they are similar. Since existing approaches are not named, we name them as **(Author, Policy)** where **Author** $\in$ **{Ours, London et al., Sakhi et al. 1, Sakhi et al. 2}** and **Policy** $\in$ **{Gaussian, Mixed-Logit}** . Here **Ours**, **London et al.**, **Sakhi et al. 1** and **Sakhi et al. 2** correspond to Theorem 4, London and Sandler (2019, Theorem 1), Sakhi et al. (2022, Proposition 1), Sakhi et al. (2022, Proposition 3), respectively. For example, London and Sandler (2019, Theorem 1) leads to two baselines **(London et al., Gaussian)** and **(London et al., Mixed-Logit)**. In all our experiments, the learning policies are trained using Adam (Kingma and Ba, 2014) with a learning rate of 0.1 for 20 epochs.

Ours, Theorem 4

(Ours, Gaussian) Here we use the Gaussian policies in (E.29). Thus we only replace the term, $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})$, with its closed-form bound in Lemma 12. This leads to the following objective.

$$\min_{\mu \in \mathbb{R}^{dK}, \sigma > 0} \Bigg(\hat{\mathcal{L}}_n^{\alpha}\left(\pi_{\mu,\sigma}^{\mathrm{GAUS}}\right) + \sqrt{\frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4\sqrt{n}}{\delta}}{2n}} + B_n^{\alpha}(\pi_{\mu,\sigma}^{\mathrm{GAUS}}) + \frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4}{\delta}}{n\lambda} + \frac{\lambda}{2}\mathrm{Var}_n^{\alpha}(\pi_{\mu,\sigma}^{\mathrm{GAUS}}) \Bigg)\,,$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for Gaussian policies. Moreover, we set $\lambda = \sqrt{2\frac{\frac{\|\mu-\mu_0\|^2}{2} - \frac{dK}{2}\log\sigma^2 + \log\frac{4}{\delta}}{n\,\mathrm{Var}_n^{\alpha}(\pi_{\mu,\sigma}^{\mathrm{GAUS}})}}$.

(Ours, Mixed-Logit) Here we use the mixed-logit policies in (E.27). Thus we only replace the terms, $D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})$, with their closed-form bound in Lemma 12. This leads to

the following objective.

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0}\left(\hat{\mathcal{L}}_n^{\alpha}\left(\pi_{\mu,\sigma}^{\text{MIXL}}\right)+\sqrt{\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{4\sqrt{n}}{\delta}}{2n}}+B_n^{\alpha}(\pi_{\mu,\sigma}^{\text{MIXL}})\right.$$
$$\left.+\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{4}{\delta}}{n\lambda}+\frac{\lambda}{2}\operatorname{Var}_n^{\alpha}(\pi_{\mu,\sigma}^{\text{MIXL}})\right),$$

where we used that $\sigma_0=1$ since our prior is $\mathbb{P}=\mathcal{N}(\eta_0\mu_0,I_{dK})\times\mathrm{G}(0,1)^K$ for mixed-logit policies. Moreover, we set $\lambda=\sqrt{2\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{4}{\delta}}{n\operatorname{Var}_n^{\alpha}(\pi_{\mu,\sigma}^{\text{MIXL}})}}$.

London and Sandler (2019, Theorem 1)

Proposition 18. *Let $\tau\in(0,1)$, $n\geq 1$, $\delta\in(0,1)$ and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n\sim\mu_{\pi_0}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$ that*

$$R\left(\pi_{\mathbb{Q}}\right)\leq\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mathbb{Q}}\right)+\sqrt{\frac{2\left(\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mathbb{Q}}\right)+\frac{1}{\tau}\right)\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})+\log\frac{n}{\delta}\right)}{\tau(n-1)}}+\frac{2\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})+\log\frac{n}{\delta}\right)}{\tau(n-1)}.\tag{E.31}$$

Baseline 1: (London et al., Gaussian) Here we use the Gaussian policies in (E.29). Thus we only replace the terms, $D_{\text{KL}}(\mathbb{Q}\|\mathbb{P})$, with their closed-form bound in Lemma 12. This leads to the following objective.

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0}\left(\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{GAUS}}\right)+\sqrt{\frac{2\left(\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{GAUS}}\right)+\frac{1}{\tau}\right)\left(\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{n}{\delta}\right)}{\tau(n-1)}}\right.$$
$$\left.+\frac{2\left(\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{n}{\delta}\right)}{\tau(n-1)}\right),$$

where we used that $\sigma_0=1$ since our prior is $\mathbb{P}=\mathcal{N}(\eta_0\mu_0,I_{dK})$ for Gaussian policies.

Baseline 2: (London et al., Mixed-Logit) Here we consider the mixed-logit policies in (E.27). Since the additional Gumbel noise does not affect the KL divergence (Lemma 12), we have the same objective as in the Gaussian case. That is

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0}\left(\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{MIXL}}\right)+\sqrt{\frac{2\left(\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mu,\sigma}^{\text{MIXL}}\right)+\frac{1}{\tau}\right)\left(\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{n}{\delta}\right)}{\tau(n-1)}}\right.$$
$$\left.+\frac{2\left(\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{n}{\delta}\right)}{\tau(n-1)}\right),$$

where we used that $\sigma_0=1$ since our prior is $\mathbb{P}=\mathcal{N}(\eta_0\mu_0,I_{dK})\times\mathrm{G}(0,1)^K$ for mixed-logit policies.

Sakhi et al. (2022, Proposition 1)

Proposition 19. *Let $\tau \in (0,1)$, $n \geq 1$, $\delta \in (0,1)$ and let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$ that*

$$R\left(\pi_{\mathbb{Q}}\right) \leq \min_{\lambda>0} \frac{1}{\tau\left(e^{\lambda}-1\right)}\left(1-e^{-\tau\lambda\hat{\mathcal{L}}_n^{\tau}(\pi_{\mathbb{Q}})+\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log\frac{2\sqrt{n}}{\delta}}{n}}\right). \tag{E.32}$$

Baseline 3: (Sakhi et al. 1, Gaussian) Here we use the Gaussian policies in (E.29).

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0,\lambda>0}\left(\frac{1}{\tau\left(e^{\lambda}-1\right)}\left(1-e^{-\tau\lambda\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mu,\sigma}^{\mathrm{GAUS}}\right)+\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{2\sqrt{n}}{\delta}}{n}}\right)\right), \tag{E.33}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK})$ for Gaussian policies.

Baseline 4: (Sakhi et al. 1, Mixed-Logit) Here we consider the mixed-logit policies in (E.27).

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0,\lambda>0}\left(\frac{1}{\tau\left(e^{\lambda}-1\right)}\left(1-e^{-\tau\lambda\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mu,\sigma}^{\mathrm{MIXL}}\right)+\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{2\sqrt{n}}{\delta}}{n}}\right)\right). \tag{E.34}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0\mu_0, I_{dK}) \times \mathrm{G}(0,1)^K$ for mixed-logit policies.

Sakhi et al. (2022, Proposition 3)

Proposition 20. *Let $\tau \in (0,1)$, $n \geq 1$, $\delta \in (0,1)$, let $\mathbb{P}$ be a fixed prior on $\mathcal{H}$, and let $\Lambda = \{\lambda_i\}_{i\in[n_\lambda]}$ a set of n_λ positive scalars. Then with probability at least $1-\delta$ over draws $\mathcal{D}_n \sim \mu_{\pi_0}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$ and any $\lambda_i \in \Lambda$,*

$$\mathcal{L}\left(\pi_{\mathbb{Q}}\right) \leq \hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mathbb{Q}}\right) + \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P})+\log\frac{2n_\lambda}{\delta}}{\lambda} + \frac{\lambda}{n}g\left(\frac{\lambda}{\tau n}\right)\mathcal{V}_n^{\tau}\left(\pi_{\mathbb{Q}}\right), \tag{E.35}$$

where $g : u \to \frac{\exp(u)-1-u}{u^2}$ and $\mathcal{V}_n^{\tau}(\pi_{\mathbb{Q}}) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{A\sim\pi_{\mathbb{Q}}(\cdot|X_i)}\left[\frac{\pi_0(A|X_i)}{\max(\tau,\pi_0(A|X_i))^2}\right]$.

Baseline 5: (Sakhi et al. 2, Gaussian) Here we consider the Gaussian policies in (E.29).

$$\min_{\mu\in\mathbb{R}^{dK},\sigma>0,\lambda\in\Lambda}\left(\hat{\mathcal{L}}_n^{\tau}\left(\pi_{\mu,\sigma}^{\mathrm{GAUS}}\right) + \sqrt{\frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{4\sqrt{n}}{\delta}}{2n}} + \frac{\frac{\|\mu-\mu_0\|^2}{2}-\frac{dK}{2}\log\sigma^2+\log\frac{2n_\lambda}{\delta}}{\lambda} + \frac{\lambda}{n}g\left(\frac{\lambda}{\tau n}\right)\mathcal{V}_n^{\tau}\left(\pi_{\mu,\sigma}^{\mathrm{GAUS}}\right)\right), \tag{E.36}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0 \mu_0, I_{dK})$ for Gaussian policies.

Baseline 6: (Sakhi et al. 2, Mixed-Logit) Here we consider the mixed-logit policies in (E.27).

$$\min_{\mu \in \mathbb{R}^{dK}, \sigma > 0, \lambda \in \Lambda} \left(\hat{\mathcal{L}}_n^{\tau} \left(\pi_{\mu,\sigma}^{\text{MIXL}} \right) + \sqrt{\frac{\frac{\|\mu - \mu_0\|^2}{2} - \frac{dK}{2} \log \sigma^2 + \log \frac{4\sqrt{n}}{\delta}}{2n}} + \frac{\frac{\|\mu - \mu_0\|^2}{2} - \frac{dK}{2} \log \sigma^2 + \log \frac{2n_\lambda}{\delta}}{\lambda} + \frac{\lambda}{n} g\left(\frac{\lambda}{\tau n} \right) \mathcal{V}_n^{\tau} \left(\pi_{\mu,\sigma}^{\text{MIXL}} \right) \right), \tag{E.37}$$

where we used that $\sigma_0 = 1$ since our prior is $\mathbb{P} = \mathcal{N}(\eta_0 \mu_0, I_{dK}) \times \mathrm{G}(0,1)^K$ for mixed-logit policies.

E.3.4 Additional Results and Discussion

In Figure E.1, we report the reward of the learned policy using one of the considered methods. We make the following observations:

- **Choice of τ and α:** in Figure E.1, we set $\tau = 1/\sqrt[4]{n} \approx 0.06$ and $\alpha = 1 - 1/\sqrt[4]{n} \approx 0.94$ so that when n is large enough, both $\hat{\mathcal{L}}_n^{\tau}(\pi)$ and $\hat{\mathcal{L}}_n^{\alpha}(\pi)$ approach $\hat{\mathcal{L}}_n^{\text{IPS}}(\pi)$ (Ionides, 2008). This is because standard IPS should be preferred when $n \to \infty$. For completeness, we also show in Figure E.2 that the choice of α and τ does not affect the conclusions that we make here. We also include in Figure E.2 the results with an adaptive and data-dependent α obtained using (8.15) in Section 8.3.4. The results in Figure E.2 will be discussed in detail after we finish analyzing the results in Figure E.1.
- **Overall performance:** our method outperforms the baselines for any class of learning policies (Gaussian or mixed-logit) and any choice of logging policies. The only exception is when the logging policy is uniform.
- **Effect of the class of learning policies:** the class of policies, Gaussian or mixed-logit, affects the performance of all the baselines. In general, Gaussian policies behave better than mixed-logit policies. However, this is less significant for our method; the performance of both Gaussian and mixed-logit policies are comparable, and in both cases, our method outperforms the baselines with Gaussian policies. Therefore, in general, Gaussian policies should be preferred over mixed-logit policies. But in case engineering constraints impose the choice of mixed-logit or softmax policies, then the performance of our method is robust to this choice.
- **Effect of the logging policy:** our method reaches the maximum reward even when the logging policy is not performing well. In contrast, the baselines only reach their best reward when the logging policy is already well-performing ($\eta_0 \approx 1$), in which case minor to no improvements are made. Note that the baselines have a better reward than ours when the logging policy is uniform. But our method has better reward when the logging policy is not uniform, that is when $\eta_0 > 0$. This is more common in practice since the logging policy is deployed in production and thus it is expected to perform better than the uniform policy.

In Figure E.2, we compare our method to **(Sakhi et al. 2)** with Gaussian policies since this was the best-performing baseline in our experiments in Figure E.1. Note that we did not include `CIFAR100` in Figure E.1 as it was computationally heavy to run these experiments with varying η_0, α and τ for a very high-dimensional dataset such as `CIFAR100`. We consider 20 varying values of τ and α evenly spaced in $(0, 1)$. We also include the results using the adaptive tuning procedure of α described in Section 8.3.4 (green curve). We make the following observations:

- **Adaptive and data-dependent α:** This procedure is reliable since the performance with an adaptive α (green curve) is comparable with the best possible choice of α. This is consistent for the three datasets.
- **Effect of the choice α:** as we observed before, the only case where the choice of α may lead to bad-performing policies is when the logging policy is uniform. When the logging policy is not uniform, our method outperforms the best baseline with the best τ for a wide range of values of α. Also, note that there is no very bad choice of α, in contrast with $\tau \approx 0$ that led to a very bad performing policy that slightly improved upon the logging policy. This attests to the robustness of our method to the choice of α. Moreover, our bound regularizes better α; it contains a bias-variance trade-off term for α. Also, the bound of **(Sakhi et al. 2)** has a $1/\tau$ making it vacuous for small values of τ.
- **Best choice of α:** To see the effect of α for varying problems, we consider the following experiment. We split the logging policies into two groups. The first is *modest logging* which corresponds to logging policies whose η_0 is between 0 and 0.5. This includes uniform logging policies and other average-performing logging policies. The second is *good logging* which corresponds to logging policies whose η_0 is between 0.5 and 1. After that, for each α, we compute the average reward of the learned policy across either the group of modest or good logging policies. For each dataset, this leads to the two red and green curves in the second row of Figure E.2. Overall, we observe that $\alpha \approx 0.7$ leads to the best performance for the *modest logging* group. Thus when the performance of the logging policy is average, regularizing the importance weights can be critical. In contrast, when the performance of the logging policy is already good, regularization is less needed and we can set $\alpha \approx 1$. Fortunately, one of the main strengths of this work is that our bound also holds for standard IPS recovered for $\alpha = 1$. The bounds in all prior works cannot provide good performance for standard IPS due to their dependency on $1/\tau$.

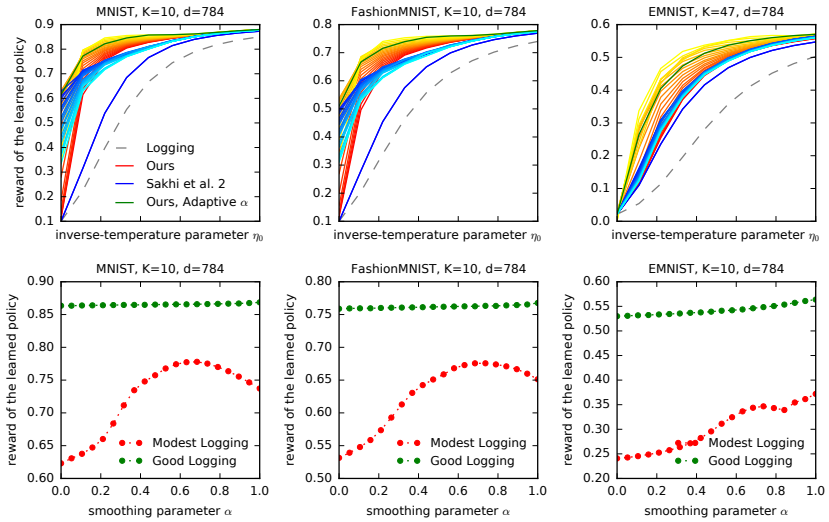


Figure E.2: In the first row, we report the reward of the learned policy with 20 evenly space values of $\tau \in (0, 1)$ and $\alpha \in (0, 1)$ and varying $\eta_0 \in [0, 1]$, and for an adaptive and data-dependent α obtained using (8.15) in Section 8.3.4. The blue-to-cyan colors correspond to different values of τ. The lighter the color, the higher the value of τ. For instance, the cyan lines correspond to high values of τ while the blue ones correspond to very small values of τ. Similarly, the red-to-yellow colors correspond to different values α. The lighter the color, the higher the value of α. For instance, the yellow lines correspond to high values of α while the red ones correspond to very small values of α. Finally, the green curve corresponds to the reward of the learned policy using an adaptive and data-dependent α described in (8.15) (Section 8.3.4). In the second row, we report the *average* reward of the learned policies using our method across the modest logging group ($\eta_0 \in [0, 0.5]$ in red) and the good logging group ($\eta_0 \in [0.5, 1]$ in green).

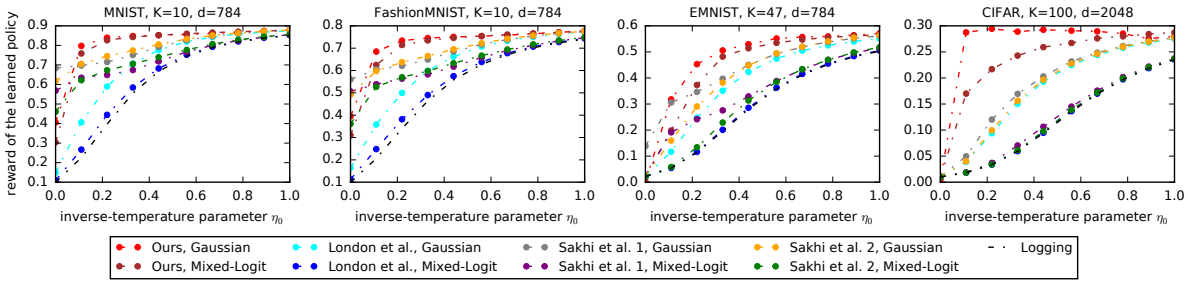


Figure E.1: The reward of the learned policy for four datasets with varying quality of the logging policy $\eta_0 \in [0, 1]$.

E.3.5 Learning Principles

Here we compare our bound in Theorem 4 and our learning principle in (8.17) to the one in London and Sandler (2019). We do not include the learning principle in Swaminathan and Joachims (2015a) since the one in London and Sandler (2019) enjoys similar performance and is far more scalable. The learning principle of London and Sandler (2019) is defined

as

$$\min_{\mu} \hat{\mathcal{L}}_n^{\tau}(\pi_\mu) + \lambda \|\mu - \mu_0\|^2 \,. \tag{E.38}$$

where λ is a tunable hyper-parameters, π_μ is the softmax policy defined in (E.26) and $\mu \in \mathbb{R}^{dK}$ is its parameter vector. This learning principle is referred to as **(London et al., LP)**. In contrast, our learning principle is defined as

$$\hat{\mathcal{L}}_n^{\alpha}(\pi_\mu) + \lambda_1 \|\mu - \mu_0\|^2 + \lambda_2 \operatorname{Var}_n^{\alpha}(\pi_\mu) + \lambda_3 B_n^{\alpha}(\pi_\mu) \,, \tag{E.39}$$

where λ_1, λ_2 and λ_3 are tunable hyper-parameters and π_μ is the Gaussian policy in (8.13) with a fixed $\sigma = 1$. Our learning principle is referred to as **(Ours, LP)**. Finally, our bound in Theorem 4 with Gaussian policies is referred to as **(Ours, Bound)**. Similarly to the previous experiments, we set $\tau = 1/\sqrt[4]{n} \approx 0.06$ and $\alpha = 1 - 1/\sqrt[4]{n} \approx 0.94$ so that when n is large enough, both $\hat{\mathcal{L}}_n^{\tau}(\pi)$ and $\hat{\mathcal{L}}_n^{\alpha}(\pi)$ approach $\hat{\mathcal{L}}_n^{\text{IPS}}(\pi)$ (Ionides, 2008). For the learning principles, we tried multiple values of hyper-parameters $\lambda, \lambda_1, \lambda_2$ and λ_3, all between 10^{-5} and 10^{-1}. For instance, we found that the best hyper-parameter for London and Sandler (2019) is $\lambda = 10^{-5}$ which matches the value they found in their `FashionMNIST` experiments. For our learning principle, the best hyper-parameters were $\lambda_1 = 10^{-5}, \lambda_2 = 10^{-5}$ and $\lambda_3 = 10^{-5}$. In contrast, our bound does not require hyper-parameter tuning. We report in Figure E.3 the reward of the learned policy on the `FashionMNIST` for all these methods with varying values of hyper-parameters. To reduce clutter, we only report the reward for good choices of hyper-parameters $\lambda, \lambda_1, \lambda_2$ and λ_3. We observe that for a wide range of hyper-parameters, our learning principle outperforms the one in London and Sandler (2019). However, both learning principles are sensitive to the choice of hyper-parameters. In contrast, our bound does not require the tuning of any additional hyper-parameter and it achieves the best performance except for the uniform logging policy. In addition to being more theoretically grounded, this approach also enjoys favorable empirical performance without additional hyper-parameter tuning, an important practical consideration.

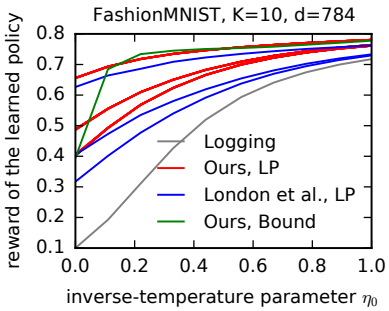


Figure E.3: The reward of the learned policy using either our bound in Theorem 4 (referred to as **(Ours, Bound)** in green), our learning principle in (8.17) (referred to as **(Ours, LP)** in red for multiple values of hyper-parameters) or the learning principle in London and Sandler (2019) (referred to as **(London et al., LP)** in blue) for multiple values of hyper-parameters).

E.3.6 Other Importance Weight Corrections

Su et al. (2020); Metelli et al. (2021) also proposed corrections that are different from hard clipping (a detailed comparison is given in Section 8.2). However, they were not included in our main experiments since they do not provide generalization guarantees; they focus on OPE and only propose a heuristic for OPL in their Appendix B.2 and Section 6.1.2, respectively. Those heuristics are not based on theory, in contrast with ours which is directly derived from our generalization bound. However, for completeness, we also compare our regularization of importance weights to theirs. To make such a comparison, we use the hyper-parameters and tuning procedures provided in Section 6 and Appendix B.2 for Metelli et al. (2021) and Sections 5 and 6.1.2 for Su et al. (2020). Overall, we observe in Figure E.4 that our method outperforms these baselines in OPL and the gap is more significant when the logging policy is not performing well.

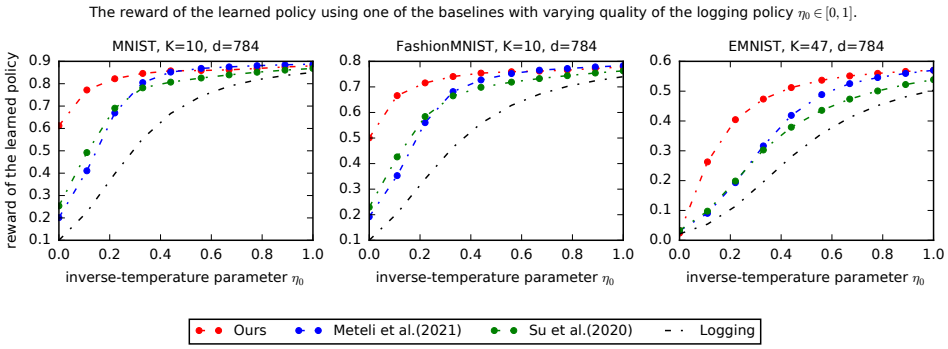


Figure E.4: The reward of the learned policy with varying quality of the logging policy $\eta_0 \in [0, 1]$ using either our regularization (α-IPS) or the ones in Su et al. (2020); Metelli et al. (2021).

Bibliography

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

M. Abeille and A. Lazaric. Linear Thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013a.

S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013b.

S. Agrawal and N. Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

P. Alquier. User-friendly introduction to pac-bayes bounds. *arXiv preprint arXiv:2110.11216*, 2021.

I. Aouali. Linear diffusion models meet contextual bandits with large action spaces. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

I. Aouali. Diffusion models meet contextual bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

I. Aouali and O. Sakhi. Off-policy learning in large action spaces: Optimization matters more than estimation. *arXiv preprint arXiv:2509.03456*, 2025.

I. Aouali, S. Ivanov, M. Gartrell, D. Rohde, F. Vasile, V. Zaytsev, and D. Legrand. Combining reward and rank signals for slate recommendation. *arXiv preprint arXiv:2107.12455*, 2021.

I. Aouali, A. Benhalloum, M. Bompaire, A. Ait Sidi Hammou, S. Ivanov, B. Heymann, D. Rohde, O. Sakhi, F. Vasile, and M. Vono. Reward optimizing recommendation using deep learning and fast maximum inner product search. In *proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4772–4773, 2022a.

I. Aouali, A. Benhalloum, M. Bompaire, B. Heymann, O. Jeunen, D. Rohde, O. Sakhi, and F. Vasile. Offline evaluation of reward-optimizing recommender systems: The case of simulation. *arXiv preprint arXiv:2209.08642*, 2022b.

I. Aouali, A. A. S. Hammou, O. Sakhi, D. Rohde, and F. Vasile. Probabilistic rank and reward: A scalable model for slate recommendation. *arXiv preprint arXiv:2208.06263*, 2022c.

I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Exponential smoothing for off-policy learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 984–1017. PMLR, 2023a.

I. Aouali, B. Kveton, and S. Katariya. Mixed-effect thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2087–2115. PMLR, 2023b.

I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Unified pac-bayesian study of pessimism for offline policy learning with regularized importance sampling. In *Uncertainty in Artificial Intelligence*, pages 88–109. PMLR, 2024.

I. Aouali, V.-E. Brunel, D. Rohde, and A. Korba. Bayesian off-policy evaluation and learning for large action spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR, 2025.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

M. G. Azar, A. Lazaric, and E. Brunskill. Sequential transfer in multi-armed bandit with finite set of models. In *Advances in Neural Information Processing Systems 26*, pages 2220–2228, 2013.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

H. Bastani, D. Simchi-Levi, and R. Zhu. Meta dynamic pricing: Transfer learning across experiments. *CoRR*, abs/1902.10918, 2019. URL https://arxiv.org/abs/1902.10918.

S. Basu, B. Kveton, M. Zaheer, and C. Szepesvari. No regrets for learning the prior in bandits. In *Advances in Neural Information Processing Systems 34*, 2021.

B. Bercu and A. Touati. Exponential inequalities for self-normalized martingales with applications. 2008.

C. M. Bishop. *Pattern Recognition and Machine Learning*, volume 4 of *Information science and statistics*. Springer, 2006.

L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

O. Catoni. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

L. Cella, A. Lazaric, and M. Pontil. Meta-learning with stochastic linear bandits. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

L. Cella, K. Lounici, and M. Pontil. Multi-task representation learning with stochastic linear bandits. *arXiv preprint arXiv:2202.10066*, 2022.

O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *Advances in Neural Information Processing Systems 24*, pages 2249–2257, 2012.

M. Chen, R. Gummadi, C. Harris, and D. Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

M. Cief, J. Golebiowski, P. Schmidt, Z. Abedjan, and A. Bekasov. Learning action embeddings for off-policy evaluation. In *European Conference on Information Retrieval*, pages 108–122. Springer, 2024.

P. Clavier, T. Huix, and A. Durmus. Vits: Variational inference thomson sampling for contextual bandits. *arXiv preprint arXiv:2307.10167*, 2023.

G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, volume 2, page 3, 2008.

A. A. Deshmukh, U. Dogan, and C. Scott. Multi-task learning for contextual bandits. In *Advances in Neural Information Processing Systems 30*, pages 4848–4856, 2017.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

M. Dimakopoulou, N. Vlassis, and T. Jebara. Marginal posterior sampling for slate bandits. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2223–2229. International Joint Conferences on Artificial Intelligence Organization, 2019.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. *International Conference on Machine Learning*, 2011.

M. Dudík, D. Erhan, J. Langford, and L. Li. Sample-efficient nonstationary policy evaluation for contextual bandits. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 247–254, Arlington, Virginia, USA, 2012. AUAI Press.

M. Dudik, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85, pages 67–82, 2018.

M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

A. Farid and A. Majumdar. Generalization bounds for meta-learning via pac-bayes and uniform stability. *Advances in Neural Information Processing Systems*, 34:2173–2186, 2021.

S. Filippi, O. Cappe, A. Garivier, and C. Szepesvari. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594, 2010.

H. Flynn, D. Reeb, M. Kandemir, and J. Peters. Pac-bayes bounds for bandit problems: A survey and experimental comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15308–15327, 2023.

D. J. Foster, C. Gentile, M. Mohri, and J. Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33:11478–11489, 2020.

G. Gabbianelli, G. Neu, and M. Papini. Importance-weighted offline learning done right. In *International Conference on Algorithmic Learning Theory*, pages 614–634. PMLR, 2024.

G. Garrigos and R. M. Gower. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pages 757–765, 2014.

A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 198–206, 2018.

A. Gilotte, O. Sakhi, I. Aouali, and B. Heymann. Offline contextual bandit with counterfactual sample identification. *arXiv preprint arXiv:2509.10520*, 2025.

A. Gopalan, S. Mannor, and Y. Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on Machine Learning*, pages 100–108, 2014.

B. Guedj. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

S. Gupta, S. Chaudhari, S. Mukherjee, G. Joshi, and O. Yagan. A unified approach to translate classical bandit algorithms to the structured bandit setting. *CoRR*, abs/1810.08164, 2018. URL https://arxiv.org/abs/1810.08164.

M. Haddouche and B. Guedj. Pac-bayes with unbounded losses through supermartingales. *arXiv preprint arXiv:2210.00928*, 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

J. Hong, B. Kveton, M. Zaheer, Y. Chow, A. Ahmed, and C. Boutilier. Latent bandits revisited. In *Advances in Neural Information Processing Systems 33*, 2020.

J. Hong, B. Kveton, S. Katariya, M. Zaheer, and M. Ghavamzadeh. Deep hierarchy in bandits. In *Proceedings of the 39th International Conference on Machine Learning*, 2022a.

J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh. Hierarchical Bayesian bandits. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022b.

J. Hong, B. Kveton, M. Zaheer, S. Katariya, and M. Ghavamzadeh. Multi-task off-policy learning from bandit feedback. In *International Conference on Machine Learning*, pages 13157–13173. PMLR, 2023.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

J. Hu, X. Chen, C. Jin, L. Li, and L. Wang. Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pages 4349–4358. PMLR, 2021.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

O. Jeunen and B. Goethals. Pessimistic reward models for off-policy learning in recommendation. In *Fifteenth ACM Conference on Recommender Systems*, pages 63–74, 2021.

Y. Jin, Z. Yang, and Z. Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

A. Korba and F. Portier. Adaptive importance sampling meets mirror descent: a bias-variance tradeoff. In *International Conference on Artificial Intelligence and Statistics*, pages 11503–11527. PMLR, 2022.

N. Korda, E. Kaufmann, and R. Munos. Thompson sampling for 1-dimensional exponential family bandits. *Advances in neural information processing systems*, 26, 2013.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

I. Kuzborskij and C. Szepesvári. Efron-stein pac-bayesian inequalities. *arXiv preprint arXiv:1909.01931*, 2019.

I. Kuzborskij, C. Vernade, A. Gyorgy, and C. Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648. PMLR, 2021.

B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

B. Kveton, M. Konobeev, M. Zaheer, C.-W. Hsu, M. Mladenov, C. Boutilier, and C. Szepesvari. Meta-Thompson sampling. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

S. Lam and J. Herlocker. MovieLens Dataset. http://grouplens.org/datasets/movielens/, 2016.

T. Lattimore and R. Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems 27*, pages 550–558, 2014.

T. Lattimore and C. Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

L. Li, W. Chu, J. Langford, and R. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

L. Li, W. Chu, J. Langford, and X. Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

L. Li, Y. Lu, and D. Zhou. Provably optimal algorithms for generalized linear contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2071–2080, 2017.

D. Liang and N. Vlassis. Local policy improvement for recommender systems. *arXiv preprint arXiv:2212.11431*, 2022.

D. Lindley and A. Smith. Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(1):1–18, 1972.

B. London and T. Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR, 2019.

X. Lu and B. Van Roy. Information-theoretic confidence bounds for reinforcement learning. In *Advances in Neural Information Processing Systems 32*, 2019.

R. D. Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

C. J. Maddison, D. Tarlow, and T. Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014.

O.-A. Maillard and S. Mannor. Latent bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pages 136–144, 2014.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

D. A. McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 230–234, 1998.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

J. Mei, C. Xiao, B. Dai, L. Li, C. Szepesvari, and D. Schuurmans. Escaping the gravitational pull of softmax. In *Advances in Neural Information Processing Systems*, volume 33, pages 21130–21140. Curran Associates, Inc., 2020a.

J. Mei, C. Xiao, C. Szepesvári, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020b.

A. M. Metelli, A. Russo, and M. Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34:8119–8132, 2021.

A. Nair, A. Gupta, M. Dalal, and S. Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

N. Nguyen, I. Aouali, A. György, and C. Vernade. Prior-dependent allocations for bayesian fixed-budget best-arm identification in structured bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 379–387. PMLR, 2025.

M. Papini, A. M. Metelli, L. Lupo, and M. Restelli. Optimistic policy optimization via multiple importance sampling. In *International Conference on Machine Learning*, pages 4989–4999. PMLR, 2019.

A. Peleg, N. Pearl, and R. Meirr. Metalearning linear bandits by prior update. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 2022.

J. Peng, H. Zou, J. Liu, S. Li, Y. Jiang, J. Pei, and P. Cui. Offline policy evaluation in large action spaces via outcome-oriented action grouping. In *Proceedings of the ACM Web Conference 2023*, pages 1220–1230, 2023.

X. B. Peng, A. Kumar, G. Zhang, and S. Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

J. Peters. Reinforcement learning by reward-weighted regression. In *NIPS 2006 Workshop: Towards a New Reinforcement Learning?*, 2006.

J. Rappaz, J. McAuley, and K. Aberer. *Recommendation on Live-Streaming Platforms: Dynamic Availability and Repeat Consumption*, page 390–399. Association for Computing Machinery, 2021.

I. Rejwan and Y. Mansour. Top-k combinatorial bandits with full-bandit feedback. In *ALT*, pages 752–776, 2020.

S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.

D. A. Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

C. Riquelme, G. Tucker, and J. Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

N. Sachdeva, Y. Su, and T. Joachims. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 965–975, 2020.

N. Sachdeva, L. Wang, D. Liang, N. Kallus, and J. McAuley. Off-policy evaluation for large action spaces via policy convolution. In *Proceedings of the ACM Web Conference 2024*, pages 3576–3585, 2024.

Y. Saito and T. Joachims. Off-policy evaluation for large action spaces via embeddings. *arXiv preprint arXiv:2202.06317*, 2022.

Y. Saito, Q. Ren, and T. Joachims. Off-policy evaluation for large action spaces via conjunct effect modeling. In *international conference on Machine learning*, pages 29734–29759. PMLR, 2023.

Y. Saito, J. Yao, and T. Joachims. POTEC: Off-policy contextual bandits for large action spaces via policy decomposition. In *The Thirteenth International Conference on Learning Representations*, 2025.

O. Sakhi, S. Bonner, D. Rohde, and F. Vasile. Blob: A probabilistic model for recommendation that combines organic and bandit signals. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 783–793, 2020.

O. Sakhi, N. Chopin, and P. Alquier. Pac-bayesian offline contextual bandits with guarantees. *arXiv preprint arXiv:2210.13132*, 2022.

O. Sakhi, D. Rohde, and N. Chopin. Fast slate policy optimization: Going beyond plackett-luce. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=f7a8XCRtUu.

O. Sakhi, I. Aouali, P. Alquier, and N. Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. *Advances in Neural Information Processing Systems*, 37:80706–80755, 2024.

S. Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639 – 658, 2010.

A. Shrivastava and P. Li. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

M. Simchowitz, C. Tosh, A. Krishnamurthy, D. Hsu, T. Lykouris, M. Dudik, and R. Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. In *Advances in Neural Information Processing Systems 34*, 2021.

A. Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Y. Su, M. Dimakopoulou, A. Krishnamurthy, and M. Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1): 1731–1755, 2015a.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015b.

A. Swaminathan, A. Krishnamurthy, A. Agarwal, M. Dudik, J. Langford, D. Jose, and I. Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

M. F. Taufiq, A. Doucet, R. Cornish, and J.-F. Ton. Marginal density ratio for off-policy evaluation in contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

S. Tomkins, P. Liao, P. Klasnja, and S. Murphy. Intelligentpooling: Practical thompson sampling for mhealth. *Machine learning*, 110(9):2685–2727, 2021.

N. Tripuraneni, C. Jin, and M. Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

I. Urteaga and C. Wiggins. Variational inference for the multi-armed contextual bandit. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 698–706, 2018.

T. Van Erven and P. Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

M. Wan, R. Misra, N. Nakashole, and J. J. McAuley. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics, 2019.

R. Wan, L. Ge, and R. Song. Metadata-based multi-task bandits with Bayesian hierarchical models. In *Advances in Neural Information Processing Systems 34*, 2021.

R. Wan, L. Ge, and R. Song. Towards scalable and robust structured bandits: A meta-learning framework. *CoRR*, abs/2202.13227, 2022. URL https://arxiv.org/abs/2202.13227.

L. Wang, A. Krishnamurthy, and A. Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. *arXiv preprint arXiv:2306.07923*, 2023.

Y.-X. Wang, A. Agarwal, and M. Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Z. Wang, A. Novikov, K. Zolna, J. S. Merel, J. T. Springenberg, S. E. Reed, B. Shahriari, N. Siegel, C. Gulcehre, N. Heess, et al. Critic regularized regression. *Advances in Neural Information Processing Systems*, 33:7768–7778, 2020.

N. Weiss. *A Course in Probability*. Addison-Wesley, 2005.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Y. Xu and A. Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.

J. Yang, W. Hu, J. D. Lee, and S. S. Du. Impact of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*, 2020.

T. Yu, B. Kveton, Z. Wen, R. Zhang, and O. Mengshoel. Graphical models meet bandits: A variational Thompson sampling approach. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

W. Zhu. Classification of mnist handwritten digit database using neural network. *Proceedings of the research school of computer science. Australian National University, Acton, ACT*, 2601, 2018.

Y. Zhu, D. J. Foster, J. Langford, and P. Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pages 27428–27453. PMLR, 2022.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Apprentissage en ligne et hors politique pour les grands espaces d'actions

Mots clés : apprentissage en ligne, apprentissage hors politique, apprentissage contrefactuel, bandits contextuels, grands espaces d'actions

Résumé : Cette thèse étudie l'apprentissage de politiques dans les systèmes interactifs où un agent observe un contexte, choisit une action parmi un très grand ensemble, puis reçoit un retour partiel. Le cadre principal est celui des bandits contextuels, avec deux paradigmes : l'apprentissage en ligne, où l'agent interagit séquentiellement avec l'environnement et minimise le regret, et l'apprentissage hors politique, où il apprend à partir de données journalisées par une politique de logging. Dans les grands espaces d'actions, ces deux cadres soulèvent des difficultés majeures : exploration coûteuse, faible couverture des données, forte variance des poids d'importance, biais d'extrapolation et objectifs difficiles à optimiser. La première partie propose des méthodes bayésiennes structurées pour l'apprentissage en ligne. Nous introduisons `meTS`, une extension de Thompson sampling fondée sur des effets mixtes, puis `dTS`, qui exploite des priors inspirés des modèles de diffusion. Ces méthodes partagent l'information entre actions et obtiennent des garanties de regret dépendant d'un nombre effectif d'actions. La seconde partie traite l'apprentissage hors politique. Nous proposons `sDM`, une méthode directe structurée fondée sur des variables latentes, montrons que l'erreur d'optimisation peut dominer l'erreur d'estimation dans les grands espaces d'actions, et introduisons des objectifs de vraisemblance pondérée par la politique, concaves et efficaces à optimiser. Enfin, nous développons des méthodes pessimistes différentiables fondées sur le lissage exponentiel et des bornes PAC-bayésiennes pour contrôler le compromis biais-variance des estimateurs par importance sampling.

Title : On-Policy and Off-Policy Learning for Large Action Spaces

Keywords : on-policy learning, off-policy learning, counterfactual learning, contextual bandits, large action spaces

Abstract : This thesis studies policy learning in interactive systems where an agent observes a context, selects an action from a very large set, and receives partial feedback. The main framework is contextual bandits, with two paradigms: on-policy learning, where the agent interacts sequentially with the environment and minimizes regret, and off-policy learning, where it learns from logged data collected by a logging policy. In large action spaces, both settings face major challenges: inefficient exploration, sparse data coverage, high-variance importance weights, extrapolation bias, and difficult optimization landscapes. The first part develops structured Bayesian methods for on-policy learning. We introduce `meTS`, a mixed-effect extension of Thompson sampling, and `dTS`, which leverages diffusion-inspired priors to model dependencies between actions. These methods share information across actions and yield regret guarantees depending on an effective number of actions. The second part addresses off-policy learning. We propose `sDM`, a structured direct method based on latent variables, show that optimization error can dominate estimation error in large action spaces, and introduce policy-weighted log-likelihood objectives that are concave and efficiently optimizable. Finally, we develop differentiable pessimistic methods based on exponential smoothing and PAC-Bayesian bounds to control the bias-variance trade-off of regularized importance-sampling estimators.

Institut Polytechnique de Paris
91120 Palaiseau, France