Bayesian Off-Policy Evaluation and Learning for Large Action Spaces

Imad Aouali CREST, ENSAE and Criteo
Victor-Emmanuel Brunel CREST, ENSAE
David Rohde Criteo
Anna Korba CREST, ENSAE

Abstract

In interactive systems, actions are often correlated, presenting an opportunity for more sample-efficient off-policy evaluation (OPE) and learning (OPL) in large action spaces. We introduce a unified Bayesian framework to capture these correlations through structured and informative priors. In this framework, we propose `sDM`, a generic Bayesian approach for OPE and OPL, grounded in both algorithmic and theoretical foundations. Notably, `sDM` leverages action correlations without compromising computational efficiency. Moreover, inspired by online Bayesian bandits, we introduce Bayesian metrics that assess the average performance of algorithms across multiple problem instances, deviating from the conventional worst-case assessments. We analyze `sDM` in OPE and OPL, highlighting the benefits of leveraging action correlations. Empirical evidence showcases the strong performance of `sDM`.

1 INTRODUCTION

An off-policy contextual bandit [Dudík et al., 2011] provides a practical framework for optimizing decision-making using logged data from an agent's interactions with an online environment [Bottou et al., 2013]. The data, comprising context-action-reward tuples, is generated when the agent observes a context, takes an action, and receives a reward based on that action and context. This data allows two key tasks: off-policy evaluation (OPE) [Dudík et al., 2011], which estimates the expected reward of a new policy, and off-policy learning (OPL) [Swaminathan and Joachims, 2015a], which learns an improved policy. Many interactive systems, like online advertising, can be modeled this way, where the context is user features, the action is product selection, and the reward is the click-through rate.

For small action spaces, existing methods based on the inverse propensity scoring (IPS) estimator [Horvitz and Thompson, 1952] are effective and come with theoretical guarantees. However, in large action spaces, these methods suffer from high bias and variance. The bias is due to the logging policy's[1] deficient support [Sachdeva et al., 2020], while the variance is due to high importance weight values [Swaminathan et al., 2017]. To address this, Saito and Joachims [2022] introduced MIPS, an IPS variant for OPE in large action spaces. MIPS relies on informative embeddings (and their distribution) that fully capture the causal effects of actions on rewards. Numerous recent variants of MIPS have been proposed [Peng et al., 2023, Sachdeva et al., 2023, Cief et al., 2024, Taufiq et al., 2024, Saito et al., 2023] to relax the original assumptions made by MIPS. All these studies are variants of IPS, designed to capture some underlying structure of the problem under different/milder assumptions than MIPS.

While much of the existing work focuses on IPS, an alternative is the direct method (DM) [Jeunen and Goethals, 2021], which directly learns a reward model. DMs typically introduce less variance than IPS but can be more prone to bias if the model is poorly suited to the data. DMs are widely used in practice, particularly in recommender systems [Sakhi et al., 2020, Jeunen and Goethals, 2021, Aouali et al., 2022], but their theory is less explored. This work seeks to analyze DMs from a Bayesian perspective and improve their efficiency in OPE and OPL by incorporating informative priors, especially in large action spaces. Similar to how MIPS [Saito and Joachims, 2022] extended IPS with auxiliary information, we aim to enhance DMs by integrating such information. However, unlike MIPS's focus on OPE, we offer theoretical and empirical insights for both OPE and OPL, relying on different as-

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

[1] The logging policy is the policy of the agent used to collect logged data.

sumptions than MIPS and its variants.

Contributions. We introduce a Bayesian method, called structured Direct Method (`sDM`), which leverages informative priors to share reward information across actions. When one action is observed, `sDM` updates its knowledge about similar actions, improving statistical efficiency without compromising computational scalability, making it suitable for large action spaces: a key limitation of traditional DMs. To evaluate `sDM`, we introduce Bayesian metrics for OPE and OPL that measure average performance across different instances, diverging from the frequentist worst-case focus. These metrics capture the benefits of informative priors, enhancing our understanding of DMs, especially given the relatively underdeveloped statistical analysis of DMs in OPE and OPL. Finally, we evaluate `sDM` using both synthetic and real data.

Related work. Bayesian models for OPL have been studied by Lazaric and Ghavamzadeh [2010], Hong et al. [2023], but these works focused on multi-task learning. Moreover, the models they use differ from ours. Hong et al. [2023] also introduced a Bayesian metric for OPL that evaluates performance in a single environment drawn from the prior given a fixed logged dataset. In contrast, our metric assesses the average performance across all environments sampled from the prior and all logged datasets, which is more typical in Bayesian analyses (e.g., Bayesian regret in Russo and Van Roy [2014]). Furthermore, we cover OPE, expanding beyond the OPL-focused scope of Lazaric and Ghavamzadeh [2010], Hong et al. [2023]. Finally, our theoretical results are more general and rely on much weaker assumptions. For more related works, refer to Appendix B.

2 BACKGROUND

Let $[n] = \{1, \dots, n\}$ for any positive integer n. Random variables are denoted with capital letters, and their realizations with the respective lowercase letters, except for Greek letters. Let X, Y be two random variables, with a slight abuse of notation, the distribution (or density) of $X \mid Y = y$ evaluated at x is denoted by $p(x \mid y)$. Let $a_1, \dots, a_n$ be n vectors of $\mathbb{R}^d$, $a = (a_i)_{i\in[n]} \in \mathbb{R}^{nd}$ is their nd-dimensional concatenation, $\otimes$ denotes the Kronecker product, and $\mathcal{O}$ the big-O notation. For any vector $a \in \mathbb{R}^d$ and any positive-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, we define $\|a\|_\Sigma = \sqrt{a^\top \Sigma a}$. Finally, $\lambda_1(A)$ and $\lambda_d(A)$ denote the maximum and minimum eigenvalue of matrix A, respectively.

Agents are represented by stochastic policies $\pi \in \Pi$, where Π is the set of policies. Precisely, given a context $x \in \mathbb{R}^d$, $\pi(\cdot \mid x)$ is a probability distribution over a finite *action set* $\mathcal{A} = [K]$, where K is the number of actions. Agents interact with a *contextual bandit* environment over n rounds. In round $i \in [n]$, the agent observes a *context* $X_i \sim \nu$, where ν is a distribution whose support $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Then the agent takes an *action* $A_i \sim \pi_0(\cdot \mid X_i)$ from the action set $\mathcal{A}$, where π_0 is the *logging policy*. Finally, assuming a parametric model, the agent receives a *stochastic reward* $R_i \sim p(\cdot \mid X_i; \theta_{*,A_i})$. Here $p(\cdot \mid x; \theta_{*,a})$ is the *reward distribution* for action a in context x, where $\theta_{*,a} \in \mathbb{R}^d$ is an *unknown d-dimensional parameter* of action a. Let $\theta_* = (\theta_{*,a})_{a\in\mathcal{A}} \in \mathbb{R}^{dK}$ be the concatenation of action parameters. Then, $r(x, a; \theta_*) = \mathbb{E}_{R\sim p(\cdot|x;\theta_{*,a})}[R]$ is the *reward function* that outputs the expected reward of action a in context x. Finally, the goal is to find a policy $\pi \in \Pi$ that maximizes the value function $V(\pi; \theta_*) = \mathbb{E}_{X\sim\nu}\mathbb{E}_{A\sim\pi(\cdot|x)}[r(X, A; \theta_*)]$.

Let $S = (X_i, A_i, R_i)_{i\in[n]}$ be a set of random variables drawn i.i.d. as $X_i \sim \nu$, $A_i \sim \pi_0(\cdot \mid X_i)$ and $R_i \sim p(\cdot \mid X_i; \theta_{*,A_i})$; that we will refer to as the sample set. The goal of off-policy evaluation (OPE) and learning (OPL) is to build an estimator $\hat{V}(\pi, S)$ of $V(\pi; \theta_*)$ and then use $\hat{V}(\pi, S)$ to find a policy $\hat{\pi} \in \Pi$ that maximizes $V(\cdot; \theta_*)$. In OPE, it is common to use *inverse propensity scoring (IPS)* [Horvitz and Thompson, 1952, Dudík et al., 2012], which leverages importance sampling to estimate the value $V(\pi; \theta_*)$ as $\hat{V}_{\text{IPS}}(\pi, S) = \frac{1}{n}\sum_{i\in[n]} \frac{\pi(A_i|X_i)}{\pi_0(A_i|X_i)} R_i$ where for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, $\frac{\pi(a|x)}{\pi_0(a|x)}$ are the *importance weights*. In practice, IPS can suffer high variance, especially when the action space is large [Swaminathan et al., 2017, Saito and Joachims, 2022]. Moreover, while IPS is unbiased under the assumption that the logging policy has full support, it can induce a high bias when such an assumption is violated [Sachdeva et al., 2020], which is again likely when the action space is large. IPS, also, assumes access to the logging policy π_0. An alternative approach to IPS is to use a *direct method (DM)* [Jeunen and Goethals, 2021], that relies on a reward model $\hat{r}$ to estimate the value $V(\pi; \theta_*)$ as

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

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., 2022, 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.

3 STRUCTURED DM

3.1 Structured Priors

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

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

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. Under this prior, each action a has an associated parameter θ_a. Given the prior in (2), the posterior distribution of an action parameter follows a multivariate Gaussian: $\theta_a \mid S \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$, 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 $B_a = \sigma^{-2}\sum_{i\in[n]} \mathbb{I}_{\{A_i=a\}} R_i \phi(X_i)$. Note that G_a and B_a only use the subset of samples S 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 sample set S doesn't cover all actions. In particular, the posterior for an unseen action a, $\theta_a \mid S$, 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 $(\theta_a)_{a\in\mathcal{A}} \mid S$, 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 Fig. 1.

$$\begin{aligned} \psi &\sim q(\cdot)\,, \\ \theta_a \mid \psi &\sim p_a(\cdot; f_a(\psi))\,, \qquad \forall a \in \mathcal{A}\,, \\ R \mid \psi, \theta, X, A &\sim p(\cdot \mid X; \theta_A)\,. \end{aligned} \tag{3}$$

This prior is flexible as assuming the existence of a latent parameter ψ is a mild assumption (Appendix E.6). 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.

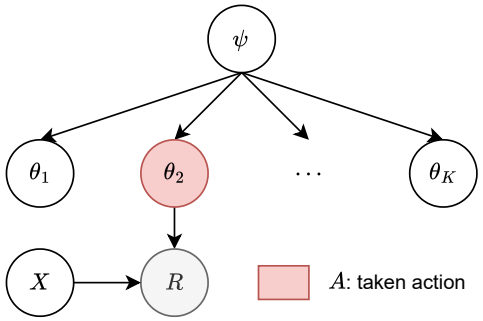

Figure 1: Graph representation of the structured prior.

Now, we discuss how to perform OPE and OPL under this general structured prior in (3), before applying it to linear-Gaussian distributions in Section 4.

3.2 Off-Policy Evaluation and Learning

Off-policy evaluation. OPE aims at estimating the value function $V(\pi; \theta_*)$ using the sample set S. In DMs, the estimator $\hat{V}_{\text{DM}}$ in (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 $\theta_a \mid S$ under model (3) since the reward is then estimated as $\hat{r}(x, a) = \mathbb{E}\left[r(x, a; \theta) \mid S\right]$ for any $(x, a) \in \mathcal{X} \times \mathcal{A}$, and this estimate is plugged into $\hat{V}_{\text{DM}}$ in (1) to estimate $V(\pi; \theta_*)$. Thus, we need to derive the posterior density of the action parameter θ_a, $p(\theta_a \mid S)$, under the structured prior in (3), which reads

$$p(\theta_a \mid S) = \int_\psi p(\theta_a \mid \psi, S) p(\psi \mid S)\, \mathrm{d}\psi\,, \tag{4}$$

where $\psi \mid S$ is the latent posterior and $\theta_a \mid \psi, S$ is the *conditional* action posterior. To compute $p(\theta_a \mid S)$, we first compute $p(\theta_a \mid \psi, S)$ and $p(\psi \mid S)$ and then integrate out ψ following (4). First,

$$p(\theta_a \mid \psi, S) \propto \mathcal{L}_a(\theta_a) p_a(\theta_a; f_a(\psi))\,, \tag{5}$$

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 S where $A_i = a$). Similarly,

$$p(\psi \mid S) \propto \prod_{b\in\mathcal{A}} \int_{\theta_b} \mathcal{L}_b(\theta_b) p_b\left(\theta_b; f_b(\psi)\right) \mathrm{d}\theta_b \, q(\psi)\,, \quad (6)$$

This allows us to further develop (4) as

$$p(\theta_a \mid S) \propto \int_{\psi} \mathcal{L}_a(\theta_a) p_a(\theta_a; f_a(\psi)) \prod_{b\in\mathcal{A}} \int_{\theta_b} \mathcal{L}_b(\theta_b) p_b\left(\theta_b; f_b(\psi)\right) \mathrm{d}\theta_b \, q(\psi) \, \mathrm{d}\psi\,. \quad (7)$$

All the quantities inside the integrals in (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 4 in the fully linear case. Otherwise, the posterior should be approximated, e.g. through approximate sampling methods, see Appendix E.1 for more details.

Greedy off-policy learning. OPL aims at finding a policy π that maximizes the value $V(\pi;\theta_*)$. We act greedy with respect to our estimator $\hat{V}_{\mathrm{DM}}(\cdot, S)$ and define the learned policy as the one maximizing it: $\hat{\pi}_{\mathrm{G}} = \operatorname{argmax}_{\pi\in\Pi} \hat{V}_{\mathrm{DM}}(\pi, S)$. If the set of policies Π contains deterministic policies, then

$$\hat{\pi}_{\mathrm{G}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b\in\mathcal{A}} \hat{r}(x, b)\}\,. \quad (8)$$

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 $\hat{\pi}_{\mathrm{P}}(a \mid x) = \mathbb{1}\{a = \operatorname{argmax}_{b\in\mathcal{A}} \hat{r}(x,b) - u(x,b)\}$. If $u(x,a)$ does not depend on x and a, then pessimistic and greedy policies are the same. The advantage of one over another depends on the evaluation metric. Our metric is the Bayesian suboptimality (BSO), defined in Section 5. It assesses the average performance of algorithms across multiple problems rather than focusing on the worst-case. The Greedy policy is more suitable for BSO optimization than pessimism (demontrated theoretically in Section 5 and empirically in Appendix F.6).

4 LINEAR-GAUSSIAN CASE

In this section, we use linear functions f_a combined with Gaussian distributions for the structured prior (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. The whole prior is

$$\begin{aligned} \psi &\sim \mathcal{N}(\mu, \Sigma)\,, \\ \theta_a \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_a\psi,\, \Sigma_a\Big)\,, \qquad \forall a \in \mathcal{A}\,, \\ R \mid \psi, \theta, X, A &\sim \mathcal{N}(\phi(X)^\top\theta_A, \sigma^2)\,. \end{aligned} \quad (9)$$

This is an important model, in both practice and theory since linear models often lead to closed-form solutions that are both computationally tractable and allow theoretical analysis. To highlight the generality of (9), we provide some problems where it can be used.

4.1 Applications

Mixed-effect modeling. (9) allows modeling that action parameters depend on a linear mixture of effect parameters. 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. In movie recommendation, θ_a is the parameter of movie a, ψ_j is the parameter of theme j (adventure, romance, etc.), and $w_{a,j}$ quantifies the relevance of theme j to movie a. Similarly, in clinical trials, a drug is a combination of multiple ingredients, each with a specific dosage. Then θ_a is the parameter of drug a, ψ_j is the parameter of ingredient j, and $w_{a,j}$ is the dosage of ingredient j in drug a.

Low-rank modeling. (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.

4.2 Closed-Form Solutions for sDM

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

$$\tilde{\Sigma}_a^{-1} = \Sigma_a^{-1} + G_a\,, \quad \tilde{\Sigma}_a^{-1}\tilde{\mu}_a = \Sigma_a^{-1}\mathrm{W}_a\psi + B_a\,, \quad (10)$$

where $G_a = \sigma^{-2}\sum_{i\in[n]} \mathbb{1}\{A_i = a\}\phi(X_i)\phi(X_i)^\top$ and $B_a = \sigma^{-2}\sum_{i\in[n]} \mathbb{1}\{A_i = a\}R_i\phi(X_i)$. 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 $\psi \mid S \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$, where

$$\bar{\Sigma}^{-1} = \Sigma^{-1} + \sum_{a\in\mathcal{A}} \mathrm{W}_a^\top(\Sigma_a^{-1} - \Sigma_a^{-1}\tilde{\Sigma}_a\Sigma_a^{-1})\mathrm{W}_a\,,$$
$$\bar{\Sigma}^{-1}\bar{\mu} = \Sigma^{-1}\mu + \sum_{a\in\mathcal{A}} \mathrm{W}_a^\top\Sigma_a^{-1}\tilde{\Sigma}_a B_a. \quad (11)$$

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 (7), the action posterior is $\theta_a \mid S \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$, where

$$\hat{\Sigma}_a = \tilde{\Sigma}_a + \tilde{\Sigma}_a\Sigma_a^{-1}\mathrm{W}_a\bar{\Sigma}\mathrm{W}_a^\top\Sigma_a^{-1}\tilde{\Sigma}_a\,,$$
$$\hat{\mu}_a = \tilde{\Sigma}_a\big(\Sigma_a^{-1}\mathrm{W}_a\bar{\mu} + B_a\big)\,. \quad (12)$$

Finally, from (9), the reward function is $r(x, a; \theta) = \phi(x)^\top\theta$. Thus, the estimated reward is $\hat{r}(x, a) = \mathbb{E}[r(x, a; \theta) \mid S] = \phi(x)^\top\hat{\mu}_a$ for any $(x, a) \in \mathcal{X}\times\mathcal{A}$, which can then be plugged in (8) for decision-making, leading to $\hat{\pi}_{\mathrm{G}}(a \mid x) = \mathbb{1}\{a = \mathrm{argmax}_{b\in\mathcal{A}}\,\phi(x)^\top\hat{\mu}_b\}$.

To see why this is more beneficial than the standard prior (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 S. Thus $\hat{\mu}_a$ and $\hat{\Sigma}_a$ are also learned using the interactions with all the actions in S, in contrast with the standard prior (2) where they were learned using only the interaction with action a. The additional computational cost of considering the structured prior in (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 (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 $(\theta_a)_{a\in\mathcal{A}} \mid S$ requires converting and storing a $dK \times dK$ covariance matrix.

Remark 4.1. sDM with this linear Gaussian hierarchy 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. An extension of sDM to binary rewards is provided in Appendix E.1.

5 ANALYSIS

While we address both OPL and OPE for our theoretical results, we place more emphasis on OPL and present it first for clarity. Although it might seem more natural to start with OPE, we found that this structure makes our contributions easier to explain. All proofs are provided in Appendix E.

5.1 Bayesian OPE and OPL Metrics

We introduce a Bayesian OPL metric, called Bayesian suboptimality. A similar OPE metric, Bayesian Mean Squared Error (BMSE), is presented afterwards.

Bayesian Suboptimality. In OPL, 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 $\pi_* = \mathrm{argmax}_{\pi\in\Pi} V(\pi; \theta_*)$ 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_*, 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:

$$\mathrm{BSO}(\hat{\pi}) = \mathbb{E}[V(\pi_*; \theta_*) - V(\hat{\pi}; \theta_*)]\,, \quad (13)$$

where the expectation is taken over all random variables: the sample set S 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 $S \mid \theta_*$. The other method involves taking an expectation under the data S, followed by taking an expectation under the posterior $\theta_* \mid S$. 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 [Aouali et al., 2023b], and this is similarly achieved by the BSO (Appendix E.2).

Hong et al. [2023] also examined a concept of Bayesian suboptimality defined as $V(\pi_*; \theta_*) - V(\hat{\pi}; \theta_*) \mid S$, where the sample set S is fixed and randomness only

comes from θ_* sampled from the posterior. In contrast, our definition aligns with traditional online bandit settings (Bayesian regret [Russo and Van Roy, 2014]) by averaging the suboptimality across both the sample set S and parameter θ_* sampled from the prior. This leads to a different metric that captures the average performance under various data realizations and parameter draws. One major difference is that our metric favors the greedy policy (Section 5.2), while they use pessimism [Hong et al., 2023, Sections 4 and 5].

Bayesian mean squared error (BMSE). In OPE, we assess the quality of estimator $\hat{r}$ using the mean squarred error (MSE). Then, similarly to the BSO, we can define the BMSE of the estimator $\hat{r}$ for a fixed action a and context x as

$$\textsc{Bmse}(\hat{r}(x,a)) = \mathbb{E}\left[\left(\hat{r}(x,a) - r(x,a;\theta_*)\right)^2\right]. \quad (14)$$

The BMSE diverges from the MSE in that the expectation in BMSE is calculated over the sample set S and across the true parameter θ_* sampled from the prior.

5.2 Theoretical Results

Our theory relies on a well-specified prior assumption, where action parameters $\theta_{*,a}$ and true rewards are assumed to be drawn from the structured prior (9). This yields our bound on the BSO of `sDM`.

Theorem 5.1 (Covariance-Dependent Bound)**.** *Let $\pi_*(x)$ be the optimal action for context x. Then the BSO of* `sDM` *under the structured prior* (9) *satisfies*

$$\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) \leq 2\sqrt{d}\,\mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right]. \quad (15)$$

Scaling of the bound in Theorem 5.1 aligns with existing frequentist results [Jin et al., 2021, Theorem 4.4]. The main differences lie in the constants, which in our case reflect the benefits of using informative priors (Appendix E.2), and the fact that this rate is achieved using greedy policies (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 using BSO as a performance metric. Specifically, $\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) \leq \textsc{Bso}(\pi)$ for any policy π, including pessimistic ones [Jin et al., 2021]. Therefore, in the Bayesian setting and when using BSO as a performance metric, greedy policies should always be preferred to pessimistic ones. This fundamental difference is proven in Appendix E.5 and it is of independent interest beyond this work. Theorem 5.1 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, problem instances θ_*, and sample sets S, then the BSO bound is correspondingly small. In particular, the tightness of the bound depends on the degree to which the sample set covers the optimal actions on average.

Even without such an assumption, Theorem 5.1 still highlights the advantages of using `sDM` over the non-structured prior (2). To see this, notice that the parameters of the non-structured prior in (2), μ_a and Σ_a, are obtained by marginalizing out ψ in (9). In this case, $\mu_a = \mathrm{W}_a\mu$ and $\Sigma_a = \Sigma_a + \mathrm{W}_a\Sigma\mathrm{W}_a^\top$. The corresponding posterior covariance is $\hat{\Sigma}_a^{\mathrm{NS}} = ((\Sigma_a + \mathrm{W}_a\Sigma\mathrm{W}_a^\top)^{-1} + G_a)^{-1}$, and is generally larger than the covariance of `sDM`, $\hat{\Sigma}_a$ in (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 sample set S, `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 (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 a 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 5.2 and defer the complete general statement to Appendix E.7.

Theorem 5.2 (Scaling with n)**.** *For n large enough and under some mild conditions stated in Appendix E.7, the BSO of* `sDM` *under the structured prior* (9) *satisfies*

$$\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) = \mathcal{O}\Bigg(\sqrt{d\mathbb{E}_{X\sim\nu}\left[e^{-\frac{n\pi_0^2(\pi_*(X)|X)}{2}}\right] + \frac{d}{n}} + \sqrt{\frac{d}{n}\mathbb{E}_{X\sim\nu}\left[\frac{1}{\pi_0(\pi_*(X)|X) + \frac{1}{n}}\right]}\Bigg).$$

The above bound demonstrates `sDM`'s scaling with n without relying on the standard *well-explored dataset* assumption commonly used in the literature [Jin et al., 2021, Hong et al., 2023]. For instance, Hong et al. [2023] assumed that the dataset satisfies the condition $G_a \succ \gamma n G$, where $G = \mathbb{E}_{X\sim\nu}[\phi(X)\phi(X)^\top]$. Instead of such assumptions, we applied techniques from Oliveira [2016] and focused on high-probability events where these conditions naturally occur. This approach enables us to derive a scaling with n that directly links the constants to π_0, specifically the probability mass π_0 assigns to the optimal policy $(\pi_0(\pi_*(X) \mid X))$ averaged over contexts $X \sim \nu$. As a result, the bound becomes smaller or larger depending on how well the logging policy π_0 covers the optimal actions for each

context x. For instance, one could replace the logging policy π_0 with a uniform distribution or the optimal policy, leading both to $\mathcal{O}(1/\sqrt{n})$ where the latter has a tighter bound independent of K.

Technical Challenges. Our proof is novel and presents several technical challenges for two main reasons. First, we bound a new metric, BSO, which prevents the use of standard frequentist techniques. Second, we avoid the common well-explored dataset assumption and instead provide a bound that directly depends on the logging policy π_0, making the dependence explicit rather than implicit in an assumption. Deriving this bound without such assumptions is challenging because we needed to control how often each action appears in the dataset. To address this, we modeled the number of appearances as a Binomial random variable with carefully chosen parameters and applied results from Oliveira [2016], which provide high-probability bounds for our posterior covariance matrix. A sketch of the proof and an outline of these challenges are provided in Appendix E.7.

A similar result can be derive in OPE, where we bound the BMSE as follows.

Theorem 5.3 (OPE Result)**.** *Let $x \in \mathcal{X}$ and $a \in \mathcal{A}$, the BMSE of* `sDM` *under the structured prior* (9) *satisfies*

$$\textsc{Bmse}(\hat{r}(x,a)) \leq \mathbb{E}\left[\|\phi(x)\|^2_{\hat{\Sigma}_a}\right] .$$

For any context x and action a, the BMSE of `sDM` is bounded by $\mathbb{E}[\|\phi(x)\|^2_{\hat{\Sigma}_a}]$. This is expected, given that we assume access to well-specified prior and likelihood, eliminating bias and directly linking BMSE to the variance. Essentially, estimation accuracy increases as the posterior covariance of action a, $\hat{\Sigma}_a$, diminishes in the direction of $\phi(x)$. For the standard non-structured prior in (2), this would only happen if the context x and action a appear frequently in the sample set S. However, `sDM`'s use of structured priors ensures lower covariance in the direction of $\phi(x)$ even without observing context x and action a together. This is because `sDM` calculates posterior covariance for action a using all observed contexts and actions, which reduces its variance. We will now present our main OPL result.

6 EXPERIMENTS

We evaluate `sDM` using both synthetic and real datasets. For OPL, we use the average reward relative to the optimal policy as the evaluation metric, and for OPE, we use the mean squared error (MSE).

6.1 Synthetic Problems

Setting. We simulate synthetic data using the linear-Gaussian model in (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 $d' = 10$. 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 $K = 10^3$, and we use a uniform logging policy to collect data. Additional experiments with different values of d', K, d, and logging policies are presented in Appendix F.

Baselines. First, we use `sDM` under prior (9). Second, we examine `DM (Bayes)`, which uses the standard non-structured prior (2), where parameters μ_a and Σ_a are obtained by marginalizing out the latent parameters ψ in (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., 2023] are also included. Implementation details of baselines is provided in Appendix F.1.

Results. In Fig. 2 (first two columns), we plot the results for OPE and OPL. Overall, `sDM` consistently outperforms the baselines in both OPE and OPL. 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. `IPS` variants for large action spaces such as `MIPS` and `PC` outperform all other `IPS` variants and standard `DM` in OPE, but this increase in performance does not translate to OPL.

Scaling to large action spaces. `sDM` achieves improved scalability compared to standard DM as 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 sample set S, instead of only using interactions with the specific action a. This is crucial, especially given that many actions may not even be observed in S. To show `sDM`'s improved scalability, we compare it to the most competitive baseline in OPL simulations, `DM (Bayes)`, for varying $K \in [10, 10^5]$ with $n = 10^3$. The results in Fig. 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 recom-

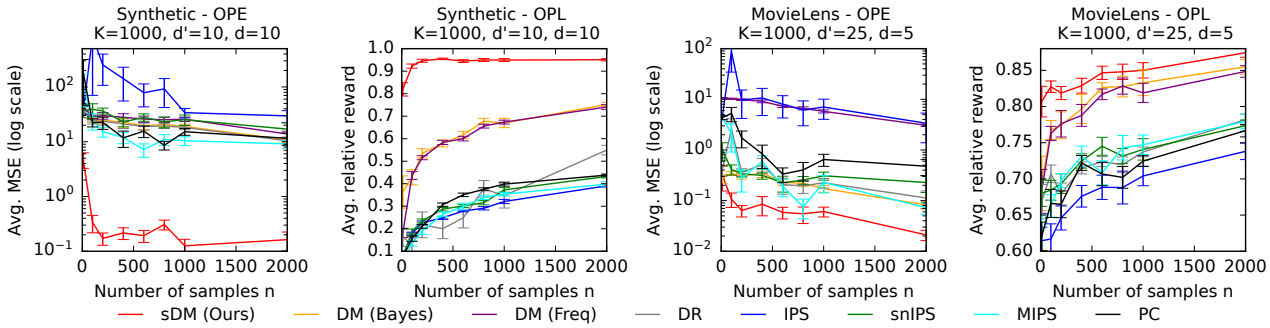

Figure 2: Performance of sDM and baselines on synthetic and MovieLens problems.

mender systems where unique embeddings are learned for each product, it still enjoys good scalability.

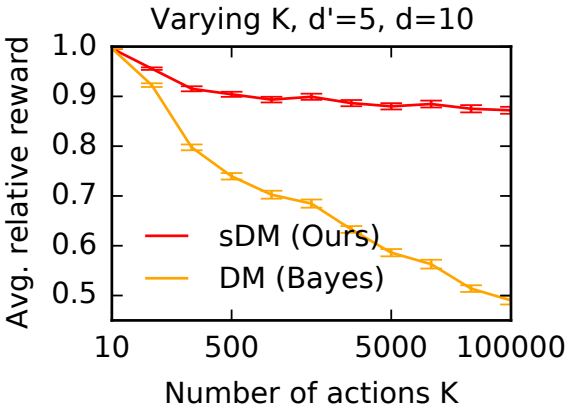

Figure 3: sDM *vs.* DM (Bayes) for varying K.

Likelihood and Prior Misspecification. We consider the case where sDM's assumed likelihood is misspecified. Specifically, we simulate binary rewards using a Bernoulli-logistic model, while sDM assumes a linear-Gaussian likelihood. Other DMs (DM (Bayes) and DM (Freq)) also use misspecified likelihoods. Results are shown in Fig. 4. Despite the misspecification, sDM significantly outperforms all methods in OPL. However, sDM experiences a drop in OPE performance. This suggests that while the assumed Gaussian reward model does not predict the reward accurately, it still captures the ranking of actions. IPS variants like MIPS, PC, and snIPS perform well in OPE, but they fail to match (or even come close to) sDM's performance in OPL, even with misspecified likelihood. Additional experiments with misspecified priors, as presented in Appendix F.4, yield similar findings.

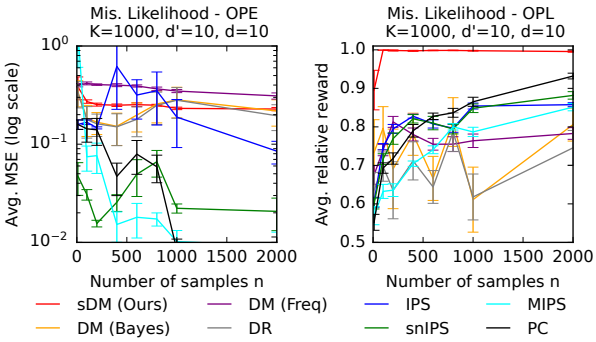

Figure 4: Effect of Misspecification.

6.2 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 (2) is inferred from data, where we set μ_a to be the mean of movie vectors across all dimensions, and $\Sigma_a = \mathrm{diag}(v)$, 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 4, 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., 2023, 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 $K = 10^3$ randomly selected movies.

Results. Results are in Fig. 2 (last two columns). Even though the latent structure assumed by sDM is not inherently present in MovieLens, sDM still outperforms the baselines by learning it offline.

6.3 KuaiRec Problems

We also compare sDM to baselines using the KuaiRec dataset [Gao et al., 2022], which is similar to MovieLens but with a key difference. Unlike MovieLens, which contains sparse user-movie interactions, KuaiRec offers a fully observed user-item interaction matrix. This allows us to evaluate algorithms without relying on semi-synthetic simulations based on low-rank factorization models to complete the matrix. This setup enables a comparison of sDM with baselines in a setting where neither the prior nor the likelihood is well-specified for sDM (and other DMs). The true reward model is unknown and not necessarily Gaussian, and no simulated environment is sampled from a prior. Due to space constraints, details on data preprocessing and filtering of KuaiRec, along with additional OPE results, are provided in Appendix F.7.

The results are shown in Fig. 5. First, sDM outperforms all DM variants, including DM (Bayes) and DM (Freq), making it the most competitive DM baseline in these experiments. When compared to IPS variants, sDM outperforms all of them except for MIPS, where the performance comparison depends on the sample size n. In low-data regimes (small n), sDM outperforms MIPS, whereas MIPS outperforms sDM when the number of samples is sufficiently large. sDM slightly outperforms MIPS when $n = 2000$.

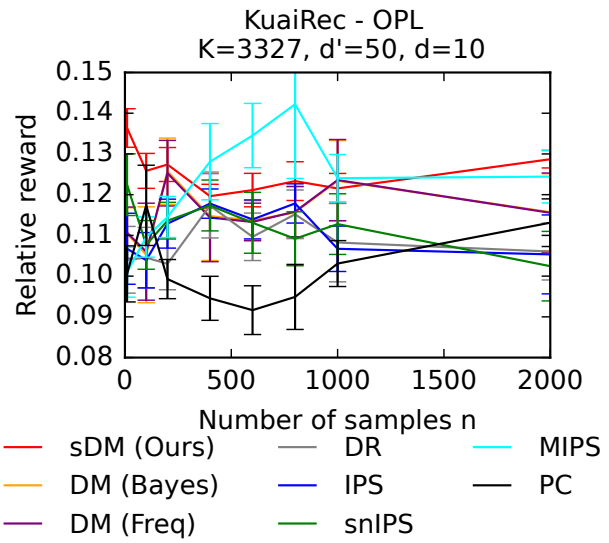

Figure 5: OPL Results on KuaiRec.

7 CONCLUSION

We introduced sDM, a novel approach for OPE and OPL that leverages latent structures among actions to enhance statistical efficiency while maintaining computational feasibility, particularly in scenarios with large action spaces and limited data. We developed a Bayesian framework to assess performance, showing the superiority of greedy policies over pessimistic ones and providing theoretical justifications for sDM's advantages. However, a key limitation of our theory is the assumption of a well-specified prior. While we have empirically studied the effects of misspecification, a theoretical investigation remains for future work. Additionally, extending sDM to handle non-linear hierarchies is a promising direction for further exploration.

Acknowledgments

We sincerely thank the anonymous reviewers for their valuable feedback and constructive suggestions, which have helped improve the clarity and quality of this work. We are also grateful to Hugo Chardon for pointing us to key references that enabled us to bound the covariance matrix under less restrictive conditions.

References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pages 127–135, 2013.

Imad Aouali. Linear diffusion models meet contextual bandits with large action spaces. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

Imad Aouali, Achraf Ait Sidi Hammou, Sergey Ivanov, Otmane Sakhi, David Rohde, and Flavian Vasile. Probabilistic rank and reward: A scalable model for slate recommendation, 2022.

Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential smoothing for off-policy learning. In *International Conference on Machine Learning*, pages 984–1017. PMLR, 2023a.

Imad Aouali, Branislav Kveton, and Sumeet Katariya. Mixed-effect thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2087–2115. PMLR, 2023b.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

Matej Cief, Jacek Golebiowski, Philipp Schmidt, Ziawasch Abedjan, and Artur Bekasov. Learning action embeddings for off-policy evaluation. In *European Conference on Information Retrieval*, pages 108–122. Springer, 2024.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *International Conference on Machine Learning*, 2011.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong 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.

Miroslav Dudik, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. Kuairec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 540–550, 2022.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon 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.

Joey Hong, Branislav Kveton, Manzil Zaheer, Sumeet Katariya, and Mohammad Ghavamzadeh. Multi-task off-policy learning from bandit feedback. In *International Conference on Machine Learning*, pages 13157–13173. PMLR, 2023.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Olivier Jeunen and Bart Goethals. Pessimistic reward models for off-policy learning in recommendation. In *Fifteenth ACM Conference on Recommender Systems*, pages 63–74, 2021.

Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba 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.

Shyong Lam and Jon Herlocker. MovieLens Dataset. http://grouplens.org/datasets/movielens/, 2016.

Tor Lattimore and Csaba Szepesvari. *Bandit Algorithms*. Cambridge University Press, 2019.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Alessandro Lazaric and Mohammad Ghavamzadeh. Bayesian multi-task reinforcement learning. In *ICML-27th international conference on machine learning*, pages 599–606. Omnipress, 2010.

Lihong Li, Wei Chu, John Langford, and Robert Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings*

of the 19th International Conference on World Wide Web, 2010.

Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR, 2019.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34:8119–8132, 2021.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166: 1175–1194, 2016.

Jie Peng, Hao Zou, Jiashuo Liu, Shaoming Li, Yibao Jiang, Jian Pei, and Peng 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.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Noveen Sachdeva, Yi Su, and Thorsten 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.

Noveen Sachdeva, Lequn Wang, Dawen Liang, Nathan Kallus, and Julian McAuley. Off-policy evaluation for large action spaces via policy convolution. *arXiv preprint arXiv:2310.15433*, 2023.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. *arXiv preprint arXiv:2202.06317*, 2022.

Yuta Saito, Qingyang Ren, and Thorsten Joachims. Off-policy evaluation for large action spaces via conjunct effect modeling. In *international conference on Machine learning*, pages 29734–29759. PMLR, 2023.

Otmane Sakhi, Stephen Bonner, David Rohde, and Flavian 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.

Otmane Sakhi, Nicolas Chopin, and Pierre Alquier. Pac-bayesian offline contextual bandits with guarantees. *arXiv preprint arXiv:2210.13132*, 2022.

Claudia Schillings, Björn Sprungk, and Philipp Wacker. On the convergence of the laplace approximation and noise-level-robustness of laplace-based monte carlo methods for bayesian inverse problems. *Numerische Mathematik*, 145:915–971, 2020.

David J Spiegelhalter and Steffen L Lauritzen. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605, 1990.

Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, pages 6005–6014. PMLR, 2019.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015a.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. *advances in neural information processing systems*, 28, 2015b.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miro Dudik, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. *Advances in Neural Information Processing Systems*, 30, 2017.

Muhammad Faaiz Taufiq, Arnaud Doucet, Rob Cornish, and Jean-Francois Ton. Marginal density ratio for off-policy evaluation in contextual bandits. *Advances in Neural Information Processing Systems*, 36, 2024.

William 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.

Lequn Wang, Akshay Krishnamurthy, and Aleksandrs Slivkins. Oracle-efficient pessimism: Offline policy optimization in contextual bandits. *arXiv preprint arXiv:2306.07923*, 2023.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.

Neil Weiss. *A Course in Probability*. Addison-Wesley, 2005.

Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *International Conference on Machine Learning*, pages 27428–27453. PMLR, 2022.

Shi Zong, Hao Ni, Kenny Sung, Nan Rosemary Ke, Zheng Wen, and Branislav Kveton. Cascading bandits for large-scale recommendation problems. *arXiv preprint arXiv:1603.05359*, 2016.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

ORGANIZATION OF THE SUPPLEMENTARY MATERIAL

The supplementary material is organized as follows.

- In **Appendix A**, we provide a detailed notation.
- In **Appendix B**, we provide an extended related work discussion.
- In **Appendix C**, we outline the posterior derivations for the standard prior in (2).
- In **Appendix D**, we outline the posterior derivations under the structured prior in (9).
- In **Appendix E**, we prove the claims made in Section 5.
- In **Appendix F**, we present supplementary experiments.
- In **Appendices G and H**, we discuss the broader impact of this work and the moderate amount of computing needed.

A DETAILED NOTATION

For any positive integer n, we define $[n]$ as the set $1, 2, \dots, n$. We represent the identity matrix of dimension $d \times d$ as I_d. In our notation, unless explicitly stated otherwise, the i-th coordinate of a vector v is denoted as v_i. However, if the vector is already indexed, such as x_i, we use the notation $x_{i,j}$ to represent the j-th entry of the vector x_i. When dealing with matrices, we refer to the (i, j)-th entry of a matrix A as $\mathrm{A}_{i,j}$. Also, $\lambda_1(\mathrm{A})$ and $\lambda_d(\mathrm{A})$ refer to the maximum and minimum eigenvalues of matrix A, respectively. Moreover, for any positive-definite matrix A and vector x, we let $\|x\|_A = \sqrt{x^\top A x}$.

Now, consider a collection of n vectors, denoted as $a_1 \in \mathbb{R}^d, a_2 \in \mathbb{R}^d, \dots, a_n \in \mathbb{R}^d$. Then we use $a = (a_i)_{i\in[n]} \in \mathbb{R}^{nd}$ to represent a nd-dimensional vector formed by concatenating the vectors $a_1, a_2, \dots, a_n$. The operator $\mathrm{Vec}(\cdot)$ is used to vectorize a matrix or a set of vectors. Let $\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_n$ be a collection of n matrices, each of dimension $d \times d$. The notation $\mathrm{diag}((\mathrm{A}_i)_{i\in[n]}) \in \mathbb{R}^{nd\times nd}$ represents a block diagonal matrix where $\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_n$ are the main-diagonal blocks. Similarly, $(\mathrm{A}_i)_{i\in[n]} \in \mathbb{R}^{nd\times d}$ denotes the $nd \times d$ matrix formed by concatenating matrices $\mathrm{A}_1, \mathrm{A}_2, \dots, \mathrm{A}_n$. To provide a more visual representation, consider a collection of vectors $a_1 \in \mathbb{R}^d, a_2 \in \mathbb{R}^d, \dots, a_n \in \mathbb{R}^d$ and matrices $\mathrm{A}_1 \in \mathbb{R}^{d\times d}, \mathrm{A}_2 \in \mathbb{R}^{d\times d}, \dots, \mathrm{A}_n \in \mathbb{R}^{d\times d}$. We can represent them as follows

$$(a_i)_{i\in[n]} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \in \mathbb{R}^{nd}, \quad \mathrm{diag}((\mathrm{A}_i)_{i\in[n]}) = \begin{pmatrix} \mathrm{A}_1 & 0 & \cdots & 0 \\ 0 & \mathrm{A}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathrm{A}_n \end{pmatrix} \in \mathbb{R}^{nd\times nd},$$

$$(\mathrm{A}_i)_{i\in[n]} = \begin{pmatrix} \mathrm{A}_1 \\ \mathrm{A}_2 \\ \vdots \\ \mathrm{A}_n \end{pmatrix} \in \mathbb{R}^{nd\times d}.$$

B EXTENDED RELATED WORK

Online learning under uncertainty is often modeled within the framework of contextual bandits [Lattimore and Szepesvari, 2019, Li et al., 2010, Chu et al., 2011]. Naturally, learning within this framework follows an online paradigm. However, practical applications often entail a large action space and a significant emphasis on short-term gains. This presents a challenge where the agent needs to exhibit a *risk-averse* behavior, contradicting the foundational principle of online algorithms designed to explore actions for long-term benefit [Auer et al., 2002, Thompson, 1933, Russo et al., 2018, Agrawal and Goyal, 2013, Abbasi-Yadkori et al., 2011]. While several practical online algorithms have emerged to efficiently explore the action space in contextual bandits [Zong et al., 2016, Zhu et al., 2022, Aouali et al., 2023b, Aouali, 2023], a notable gap exists in the quest for an offline procedure

capable of optimizing decision-making using historical data. Fortunately, we are often equipped with a sample set collected from past interactions. Leveraging this data, agents can enhance their policies offline [Swaminathan and Joachims, 2015a, London and Sandler, 2019, Sakhi et al., 2022], consequently improving the overall system performance. This study is primarily concerned with this offline or *off-policy* formulation of contextual bandits [Dudík et al., 2011, Dudík et al., 2012, Dudik et al., 2014, Wang et al., 2017, Farajtabar et al., 2018]. There are two main tasks in off-policy contextual bandits, first, off-policy evaluation (OPE) [Dudík et al., 2011] which involves estimating policy performance using historical data, simulating as if these evaluations were performed while the policy is interacting with the environment in real-time. Subsequently, the derived estimator is refined to closely approximate the optimal policy, a process known as off-policy learning (OPL) [Swaminathan and Joachims, 2015a]. Next, we review both OPE and OPL.

B.1 Off-Policy Evaluation

Recent years have witnessed a surge in interest in OPE, with several key works contributing to this field [Dudík et al., 2011, Dudík et al., 2012, Dudik et al., 2014, Wang et al., 2017, Farajtabar et al., 2018, Su et al., 2019, 2020, Metelli et al., 2021, Kuzborskij et al., 2021, Saito and Joachims, 2022, Sakhi et al., 2020, Jeunen and Goethals, 2021]. The literature on OPE can be broadly categorized into three main approaches. The first approach referred to as the direct method (DM) [Jeunen and Goethals, 2021, Hong et al., 2023], involves learning a model that approximates the expected reward. This model is then used to estimate the performance of evaluated policies. The second approach, inverse propensity scoring (IPS) [Horvitz and Thompson, 1952, Dudík et al., 2012], aims to estimate the reward of evaluated policies by correcting for the preference bias of the logging policy in the sample set. IPS is unbiased when there's an assumption that the evaluation policy is absolutely continuous with respect to the logging policy. However, it can exhibit high variance and substantial bias when this assumption is violated [Sachdeva et al., 2020]. Various techniques have been introduced to address the variance issue, such as clipping importance weights [Ionides, 2008, Swaminathan and Joachims, 2015a], smoothing them [Aouali et al., 2023a], self-normalization of weights [Swaminathan and Joachims, 2015b], etc. [Gilotte et al., 2018]. The third approach, known as doubly robust (DR) [Robins and Rotnitzky, 1995, Bang and Robins, 2005, Dudík et al., 2011, Dudik et al., 2014, Farajtabar et al., 2018], combines elements of DM and IPS, helping to reduce variance. Assessing the accuracy of an OPE estimator, denoted as $\hat{R}_n(\pi)$, is typically done using the mean squared error (MSE). It may be relevant to note that Metelli et al. [2021] argued that high-probability concentration rates should be preferred over the MSE to evaluate OPE estimators as they provide non-asymptotic guarantees. In this work, we presented a direct method for OPE, for which we could derive both an MSE and high-probability concentration bound under our assumptions. However, we focused more on OPL.

B.2 Off-Policy Learning

OPL under IPS. Prior research on OPL predominantly revolved around developing learning principles influenced by generalization bounds for IPS. Swaminathan and Joachims [2015a] proposed to penalize the IPS estimator with a variance term leading to a learning principle that promotes policies that achieve high estimated reward and exhibit minimal empirical variance. In a recent development, London et al. [London and Sandler, 2019] made a connection between PAC-Bayes theory and OPL that inspired several works in the same vein [Sakhi et al., 2022, Aouali et al., 2023a]. London and Sandler [2019] proposed a learning principle that favors policies with high estimated reward while keeping the parameter close to that of the logging policy in terms of L_2 distance. In particular, this learning principle is scalable. Then Sakhi et al. [2022], Aouali et al. [2023a] took a different direction of deriving tractable generalization bounds and optimizing them directly as they are. Finally, Wang et al. [2023] recently introduced an efficient learning principle with guarantees for a specific choice of their hyperparameter. All of these learning principles are related to the concept of pessimism which we discuss next.

OPL under DM. The majority of OPL approaches in contextual bandits rely on the principle of pessimism [Jeunen and Goethals, 2021, Jin et al., 2021, Hong et al., 2023]. In essence, these methods construct confidence intervals for reward estimates $\hat{r}(x,a)$, that hold simultaneously for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ and satisfy the following condition

$$|r(x,a;\theta) - \hat{r}(x,a)| \leq u(x,a)\,, \qquad \forall (x,a) \in \mathcal{X} \times \mathcal{A}$$

Subsequently, the policy learned with this pessimistic approach is defined as

$$\hat{\pi}_{\mathrm{P}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b \in \mathcal{A}} \hat{r}(x,b) - u(x,b)\}\,. \tag{16}$$

Note that when the function $u(x,a)$ is independent of both x and a, the pessimistic and greedy policies become equivalent. However, in this work, we highlight the critical role of the selected metric and the underlying assumptions about $\hat{r}$ in choosing between pessimistic and greedy policies. In particular, we introduce the concept of Bayesian suboptimality (BSO), as defined in Section 5.1. BSO evaluates the average performance of algorithms across multiple problem instances, and in this context, the Greedy policy emerges as a more suitable choice for minimizing the BSO. A detailed comparison of pessimistic and greedy policies is presented in Appendix E.5.

In addition to not using pessimism, our work also considers structured contextual bandit problems, upon which we build informative priors. Notably, while structured problems have been previously investigated in the context of multi-task learning by Lazaric and Ghavamzadeh [2010], Hong et al. [2023], these works primarily focused on cases where there are multiple contextual bandit instances that bear similarity to one another. Interaction with one bandit instance contributes to the agent's understanding of other instances. In contrast, our work tackles a single contextual bandit instance with a large action space and considers an inherent structure among actions within this bandit instance. Furthermore, the structure we address in the single-instance problem is different and more general compared to the ones explored in the aforementioned multi-task learning works.

C POSTERIOR DERIVATIONS UNDER STANDARD PRIORS

Here we derive the posterior under the standard prior in (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 C.1 (Independence)**.** (X, A) is independent of θ, and the θ_a, for $a \in \mathcal{A}$ are independent.

Derivation of $p(\theta_a \mid S)$ for the standard prior in (2). We start by recalling the standard prior in (2)

$$\begin{aligned} \theta_a &\sim \mathcal{N}(\mu_a, \Sigma_a)\,, && \forall a \in \mathcal{A}\,, \\ R \mid \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2)\,, \end{aligned} \tag{17}$$

where $\mathcal{N}(\mu_a, \Sigma_a)$ is the prior on the action parameter θ_a. Let $\theta = (\theta_a)_{a \in \mathcal{A}} \in \mathbb{R}^{dK}$, $\Sigma_{\mathcal{A}} = \mathrm{diag}(\Sigma_a)_{a \in \mathcal{A}} \in \mathbb{R}^{dK \times dK}$ 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 (17) as

$$\begin{aligned} \theta &\sim \mathcal{N}(\mu_{\mathcal{A}}, \Sigma_{\mathcal{A}})\,, \\ R \mid \theta, X, A &\sim \mathcal{N}((u_A \otimes \phi(X))^\top \theta, \sigma^2)\,. \end{aligned} \tag{18}$$

Then the *joint* action posterior $p(\theta \mid S)$ decomposes as

$$
\begin{aligned}
p(\theta \mid S) &= p(\theta \mid (X_i, A_i, R_i)_{i\in[n]}) \overset{(i)}{\propto} p((R_i)_{i\in[n]} \mid \theta, (X_i, A_i)_{i\in[n]}) p(\theta \mid (X_i, A_i)_{i\in[n]}) \,, \\
&\overset{(ii)}{=} p((R_i)_{i\in[n]} \mid \theta, (X_i, A_i)_{i\in[n]}) p(\theta) \overset{(iii)}{=} \prod_{i\in[n]} p(R_i \mid \theta, X_i, A_i) p(\theta) \,, \\
&\overset{(iv)}{=} \prod_{i\in[n]} \mathcal{N}(R_i; (u_{A_i} \otimes \phi(X_i))^\top \theta, \sigma^2) \mathcal{N}(\theta; \mu_{\mathcal{A}}, \Sigma_{\mathcal{A}}) \,, \\
&= \exp\left[-\frac{1}{2} \left(v \sum_{i=1}^n (R_i^2 - 2R_i (u_{A_i} \otimes \phi(X_i))^\top \theta + ((u_{A_i} \otimes \phi(X_i))^\top \theta)^2) + \theta^\top \Lambda_{\mathcal{A}} \theta - 2\theta^\top \Lambda_{\mathcal{A}} \mu_{\mathcal{A}} + \mu_{\mathcal{A}}^\top \Lambda_{\mathcal{A}} \mu_{\mathcal{A}} \right) \right] \,, \\
&\propto \exp\left[-\frac{1}{2} \left(\theta^\top (v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))(u_{A_i} \otimes \phi(X_i))^\top + \Lambda_{\mathcal{A}}) \theta - 2\theta^\top \left(v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))^\top R_i + \Lambda_{\mathcal{A}} \mu_{\mathcal{A}} \right) \right) \right] \,, \\
&= \exp\left[-\frac{1}{2} \left(\theta^\top (v \sum_{i=1}^n (u_{A_i} u_{A_i}^\top \otimes \phi(X_i)\phi(X_i)^\top) + \Lambda_{\mathcal{A}}) \theta - 2\theta^\top \left(v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))^\top R_i + \Lambda_{\mathcal{A}} \mu_{\mathcal{A}} \right) \right) \right] \,, \\
&\overset{(v)}{\propto} \mathcal{N}\left(\theta; \hat{\mu}_{\mathcal{A}}, \left(\hat{\Lambda}_{\mathcal{A}} \right)^{-1} \right) \,.
\end{aligned}
$$

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 i.i.d. 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 $\hat{\mu}_a = \hat{\Lambda}_{\mathcal{A}}^{-1} \big(v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))^\top R_i + \Lambda_{\mathcal{A}} \mu_{\mathcal{A}} \big)$. Now notice that $\hat{\Lambda}_{\mathcal{A}} = \mathrm{diag}(\Sigma_a^{-1} + G_a)_{a\in\mathcal{A}}$. Thus, $p(\theta \mid S) = \mathcal{N}(\theta, \hat{\mu}_{\mathcal{A}}, \Lambda_{\mathcal{A}}^{-1})$ where $\hat{\mu}_{\mathcal{A}} = (\hat{\mu}_a)_{a\in\mathcal{A}}$ and $\Lambda_{\mathcal{A}} = \mathrm{diag}(\hat{\Lambda}_a)_{a\in\mathcal{A}}$, with

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

Since the covariance matrix of $p(\theta \mid S)$ is diagonal by block, we know that the marginals $\theta_a \mid S$ also have a Gaussian density $p(\theta_a \mid S) = \mathcal{N}(\theta_a; \hat{\mu}_a, \hat{\Sigma}_a)$ where $\hat{\Sigma}_a = \hat{\Lambda}_a^{-1}$.

□

D POSTERIOR DERIVATIONS UNDER STRUCTURED PRIORS

Here we derive the posteriors under the structured prior in (9). Precisely, we derive the latent posterior density of $\psi \mid S$, the conditional posterior density of $\theta \mid S, \psi$. Then, we derive the marginal posterior $\theta \mid S$. Posterior derivations rely on the following assumption.

Assumption D.1 (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.

D.1 Latent Posterior

Derivation of $p(\psi \mid S)$. First, recall that our model in (9) reads

$$
\begin{aligned}
\psi &\sim \mathcal{N}(\mu, \Sigma) \,, \\
\theta_a \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_a \psi, \Sigma_a \Big) \,, && \forall a \in \mathcal{A} \,, \\
R \mid \psi, \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2) \,.
\end{aligned} \tag{19}
$$

Then we first rewrite it as

$$
\begin{aligned}
\psi &\sim \mathcal{N}(\mu, \Sigma) \,, \\
\theta \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_{\mathcal{A}} \psi, \Sigma_{\mathcal{A}} \Big) \,, \\
R \mid \psi, \theta, X, A &\sim \mathcal{N}((u_A \otimes \phi(X))^\top \theta, \sigma^2) \,.
\end{aligned} \tag{20}
$$

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 D.1. 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 $I_a = |S_a| = \sum_{i\in[n]} \mathbb{1}\{A_i = a\}$ as the number of times action a appears in the sample set S. Now let $h_a(\psi) = \int_{\theta_a} \left(\prod_{i\in I_a} \mathcal{N}(R_i; \phi(X_i)^\top \theta_a, \sigma^2)\right) \mathcal{N}\left(\theta_a; \mathrm{W}_a\psi, \Sigma_a\right) \mathrm{d}\theta$. Then we have that

$$p(\psi \mid S) \propto \prod_{a\in\mathcal{A}} h_a(\psi) \mathcal{N}(\psi; \mu, \Sigma) \,. \tag{21}$$

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

$$\begin{aligned}
h_a(\psi) &= \int_{\theta_a} \left(\prod_{i\in I_a} \mathcal{N}(R_i; \phi(X_i)^\top \theta_a, \sigma^2) \right) \mathcal{N}\left(\theta_a; \mathrm{W}_a\psi, \Sigma_a\right) \mathrm{d}\theta_a \,, \\
&\propto \int_{\theta_a} \exp\left[-\frac{1}{2} v \sum_{i\in I_a} (R_i - \phi(X_i)^\top \theta_a)^2 - \frac{1}{2} (\theta_a - \mathrm{W}_a\psi)^\top \Lambda_a (\theta_a - \mathrm{W}_a\psi) \right] \mathrm{d}\theta_a \,, \\
&= \int_{\theta_a} \exp\Big[-\frac{1}{2} \Big(v \sum_{i\in I_a} (R_i^2 - 2R_i \theta_a^\top \phi(X_i) + (\theta_a^\top \phi(X_i))^2) + \theta_a^\top \Lambda_a \theta_a - 2\theta_a^\top \Lambda_a \mathrm{W}_a\psi \\
&\qquad\qquad + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi) \Big) \Big] \mathrm{d}\theta_a \,, \\
&\propto \int_{\theta_a} \exp\Big[-\frac{1}{2} \Big(\theta_a^\top \left(v \sum_{i\in I_a} \phi(X_i)\phi(X_i)^\top + \Lambda_a \right) \theta_a - 2\theta_a^\top \left(v \sum_{i\in I_a} R_i \phi(X_i) + \Lambda_a \mathrm{W}_a\psi \right) \\
&\qquad\qquad + (\mathrm{W}_a\psi)^\top \Lambda_a (\mathrm{W}_a\psi) \Big) \Big] \mathrm{d}\theta_a \,.
\end{aligned}$$

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

$$
\begin{aligned}
f(\psi) &\propto \int_{\theta_a} \exp\left[-\frac{1}{2}\left(\theta_a^\top U_a\theta_a - 2\theta_a^\top U_aV_a\left(B_a+\Lambda_a\mathrm{W}_a\psi\right) + (\mathrm{W}_a\psi)^\top\Lambda_a(\mathrm{W}_a\psi)\right)\right]\mathrm{d}\theta_a\,,\\
&= \int_{\theta_a} \exp\left[-\frac{1}{2}\left(\theta_a^\top U_a\theta_a - 2\theta_a^\top U_a\beta_a + (\mathrm{W}_a\psi)^\top\Lambda_a(\mathrm{W}_a\psi)\right)\right]\mathrm{d}\theta_a\,,\\
&= \int_{\theta_a} \exp\left[-\frac{1}{2}\left((\theta_a-\beta_a)^\top U_a(\theta_a-\beta_a) - \beta_a^\top U_a\beta_a + (\mathrm{W}_a\psi)^\top\Lambda_a(\mathrm{W}_a\psi)\right)\right]\mathrm{d}\theta_a\,,\\
&\propto \exp\left[-\frac{1}{2}\left(-\beta_a^\top U_a\beta_a + (\mathrm{W}_a\psi)^\top\Lambda_a(\mathrm{W}_a\psi)\right)\right]\,,\\
&= \exp\left[-\frac{1}{2}\left(-\left(B_a+\Lambda_a\mathrm{W}_a\psi\right)^\top V_a\left(B_a+\Lambda_a\mathrm{W}_a\psi\right) + (\mathrm{W}_a\psi)^\top\Lambda_a(\mathrm{W}_a\psi)\right)\right]\,,\\
&\propto \exp\left[-\frac{1}{2}\left(\psi^\top\mathrm{W}_a^\top\left(\Lambda_a-\Lambda_aV_a\Lambda_a\right)\mathrm{W}_a\psi - 2\psi^\top\left(\mathrm{W}_a^\top\Lambda_aV_aB_a\right)\right)\right]\,,\\
&\propto \mathcal{N}(\psi;\bar\mu_a,\bar\Sigma_a)\,,
\end{aligned}
$$

where

$$
\begin{aligned}
\bar\Sigma_a^{-1} &= \mathrm{W}_a^\top\left(\Lambda_a-\Lambda_aV_a\Lambda_a\right)\mathrm{W}_a = \mathrm{W}_a^\top\left(\Sigma_a^{-1}-\Sigma_a^{-1}(G_a+\Sigma_a^{-1})^{-1}\Sigma_a^{-1}\right)\mathrm{W}_a\,,\\
\bar\Sigma_a^{-1}\bar\mu_a &= \mathrm{W}_a^\top\Lambda_aV_aB_a = \mathrm{W}_a^\top\Sigma_a^{-1}(G_a+\Sigma_a^{-1})^{-1}B_a\,.
\end{aligned}
\tag{22}
$$

However, we know from (21) that $p(\psi\mid S)\propto\prod_{a\in\mathcal{A}}h_a(\psi)\mathcal{N}(\psi;\mu,\Sigma)$. But $h_a(\psi)$ is proportional to $\mathcal{N}(\psi;\bar\mu_a,\bar\Sigma_a)$ for any a. Thus $p(\psi\mid S)$ can be seen as the product of $K+1$ multivariate Gaussian distributions $\mathcal{N}(\mu,\Sigma)$ and $\mathcal{N}(\bar\mu_a,\bar\Sigma_a)$ for $a\in\mathcal{A}$ (the terms h_a). Thus, $p(\psi\mid S)$ is also a multivariate Gaussian distribution $\mathcal{N}(\bar\mu,\bar\Sigma^{-1})$, with

$$
\bar\Sigma^{-1} = \Sigma^{-1}+\sum_{a\in\mathcal{A}}\bar\Sigma_a^{-1} = \Sigma^{-1}+\sum_{a\in\mathcal{A}}\mathrm{W}_a^\top\left(\Sigma_a^{-1}-\Sigma_a^{-1}(G_a+\Sigma_a^{-1})^{-1}\Sigma_a^{-1}\right)\mathrm{W}_a\,, \tag{23}
$$

$$
\bar\Sigma^{-1}\bar\mu = \Sigma^{-1}\mu+\sum_{a\in\mathcal{A}}\bar\Sigma_a^{-1}\bar\mu_a = \Sigma^{-1}\mu+\sum_{a\in\mathcal{A}}\mathrm{W}_a^\top\Sigma_a^{-1}(G_a+\Sigma_a^{-1})^{-1}B_a\,. \tag{24}
$$

□

D.2 Conditional Posterior

Derivation of $p(\theta_a \mid \psi, S)$. Let $v = \sigma^{-2}$, $\Lambda_a = \Sigma_a^{-1}$. We consider the model rewritten in (20), then the *joint* action posterior $p(\theta \mid S)$ decomposes as

$$
\begin{aligned}
p(\theta \mid \psi, S) &= p(\theta \mid \psi, (X_i, A_i, R_i)_{i\in[n]}) \overset{(i)}{\propto} p((R_i)_{i\in[n]} \mid \theta, \psi, (X_i, A_i)_{i\in[n]}) p(\theta \mid \psi, (X_i, A_i)_{i\in[n]}) \,, \\
&\overset{(ii)}{=} p((R_i)_{i\in[n]} \mid \theta, (X_i, A_i)_{i\in[n]}) p(\theta \mid \psi) \overset{(iii)}{=} \prod_{i\in[n]} p(R_i \mid \theta, X_i, A_i) p(\theta \mid \psi) \,, \\
&\overset{(iv)}{=} \prod_{i\in[n]} \mathcal{N}(R_i; (u_{A_i} \otimes \phi(X_i))^\top \theta, \sigma^2) \mathcal{N}(\theta; \mathrm{W}_{\mathcal{A}} \psi, \Sigma_{\mathcal{A}}) \,, \\
&= \exp\Big[-\frac{1}{2} \Big(v \sum_{i=1}^n (R_i^2 - 2R_i (u_{A_i} \otimes \phi(X_i))^\top \theta + ((u_{A_i} \otimes \phi(X_i))^\top \theta)^2) + \theta^\top \Lambda_{\mathcal{A}} \theta - 2\theta^\top \Lambda_{\mathcal{A}} \mathrm{W}_{\mathcal{A}} \psi \\
&\qquad\qquad + \psi^\top \mathrm{W}_{\mathcal{A}}^\top \Lambda_{\mathcal{A}} \mathrm{W}_{\mathcal{A}} \psi \Big) \Big] \,, \\
&\propto \exp\left[-\frac{1}{2} \left(\theta^\top (v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))(u_{A_i} \otimes \phi(X_i))^\top + \Lambda_{\mathcal{A}}) \theta - 2\theta^\top \left(v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))^\top R_i + \Lambda_{\mathcal{A}} \mathrm{W}_{\mathcal{A}} \psi \right) \right) \right] \,, \\
&= \exp\Big[-\frac{1}{2} \Big(\theta^\top (v \sum_{i=1}^n (u_{A_i} u_{A_i}^\top \otimes \phi(X_i)\phi(X_i)^\top) + \Lambda_{\mathcal{A}}) \theta - 2\theta^\top \left(v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))^\top R_i + \Lambda_{\mathcal{A}} \mathrm{W}_{\mathcal{A}} \psi \right) \Big) \Big] \,, \\
&\overset{(v)}{\propto} \mathcal{N}\left(\theta; \tilde{\mu}_{\mathcal{A}}, \left(\tilde{\Lambda}_{\mathcal{A}} \right)^{-1} \right) \,,
\end{aligned}
$$

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 i.i.d. 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 $\tilde{\mu}_a = \tilde{\Lambda}_{\mathcal{A}}^{-1} \big(v \sum_{i=1}^n (u_{A_i} \otimes \phi(X_i))^\top R_i + \Lambda_{\mathcal{A}} \mathrm{W}_{\mathcal{A}} \psi \big)$. Now notice that $\tilde{\Lambda}_{\mathcal{A}} = \mathrm{diag}(\Sigma_a^{-1} + G_a)_{a\in\mathcal{A}}$. Thus, $p(\theta \mid S) = \mathcal{N}(\theta, \tilde{\mu}_{\mathcal{A}}, \Lambda_{\mathcal{A}}^{-1})$ where $\tilde{\mu}_{\mathcal{A}} = (\tilde{\mu}_a)_{a\in\mathcal{A}}$ and $\Lambda_{\mathcal{A}} = \mathrm{diag}(\tilde{\Lambda}_a)_{a\in\mathcal{A}}$, 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 $p(\theta \mid \psi, S)$ is diagonal by block. Thus $\theta_a \mid \psi, S$ for $a \in \mathcal{A}$ are independent and have a Gaussian density $p(\theta_a \mid S) = \mathcal{N}(\theta_a; \tilde{\mu}_a, \tilde{\Sigma}_a)$ where $\tilde{\Sigma}_a = \tilde{\Lambda}_a^{-1}$. □

D.3 Action Posterior

Derivation of $p(\theta_a \mid S)$. We know that $\theta_a \mid S, \psi \sim \mathcal{N}(\tilde{\mu}_a, \tilde{\Sigma}_a)$ and $\psi \mid S \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$. Thus the posterior density of $\theta_a \mid S$ is also Gaussian since Gaussianity is preserved after marginalization [Koller and Friedman, 2009]. We let $\theta_a \mid S \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$. 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

$$
\begin{aligned}
\tilde{\Sigma}_a &= (G_a + \Lambda_a)^{-1} \\
\mathbb{E}\left[\theta_a \,|\, \psi, S\right] &= \tilde{\Sigma}_a \left(B_a + \Lambda_a \mathrm{W}_a \psi\right)
\end{aligned}
$$

First, given S, $\tilde{\Sigma}_a = (G_a + \Lambda_a)^{-1}$ and B_a are constant (do not depend on ψ). Thus

$$
\begin{aligned}
\hat{\mu}_a = \mathbb{E}\left[\theta_a \,|\, S\right] = \mathbb{E}\left[\mathbb{E}\left[\theta_a \,|\, \psi, S\right] \,|\, S\right] = \mathbb{E}_{\psi\sim\mathcal{N}(\bar{\mu},\bar{\Sigma})} \left[\tilde{\Sigma}_a \left(B_a + \Lambda_a \mathrm{W}_a \psi\right) \right] &= \tilde{\Sigma}_a \left(B_a + \Lambda_a \mathrm{W}_a \mathbb{E}_{\psi\sim\mathcal{N}(\bar{\mu},\bar{\Sigma})} \left[\psi\right]\right) \,, \\
&= \tilde{\Sigma}_a \left(B_a + \Lambda_a \mathrm{W}_a \bar{\mu}\right) \,.
\end{aligned}
$$

This concludes the computation of $\hat{\mu}_a$. Similarly, given S, $\tilde{\Sigma}_a = (G_a + \Lambda_a)^{-1}$ and B_a are constant (do not depend on ψ), yields two things. First,

$$
\mathbb{E}\left[\mathrm{cov}\left[\theta_a \,|\, \psi, S\right] \,|\, S\right] = \mathbb{E}\left[\tilde{\Sigma}_a \,\middle|\, S\right] = \tilde{\Sigma}_a \,.
$$

Second,

$$\begin{aligned}\operatorname{cov}\left[\mathbb{E}\left[\theta_a \mid \psi, S\right] \mid S\right] &= \operatorname{cov}\left[\tilde{\Sigma}_a \Lambda_a \mathrm{W}_a \psi \,\middle|\, S\right] \\ &= \tilde{\Sigma}_a \Lambda_a \mathrm{W}_a \operatorname{cov}\left[\psi \mid S\right] \mathrm{W}_a^\top \Lambda_a \tilde{\Sigma}_a \\ &= \tilde{\Sigma}_a \Lambda_a \mathrm{W}_a \bar{\Sigma} \mathrm{W}_a^\top \Lambda_a \tilde{\Sigma}_a \,.\end{aligned}$$

Finally, the total covariance decomposition [Weiss, 2005] yields that

$$\hat{\Sigma}_a = \operatorname{cov}\left[\theta_a \mid S\right] = \mathbb{E}\left[\operatorname{cov}\left[\theta_a \mid \psi, S\right] \mid S\right] + \operatorname{cov}\left[\mathbb{E}\left[\theta_a \mid \psi, S\right] \mid S\right] = \tilde{\Sigma}_a + \tilde{\Sigma}_a \Lambda_a \mathrm{W}_a \bar{\Sigma} \mathrm{W}_a^\top \Lambda_a \tilde{\Sigma}_a \,.$$

This concludes the proof.

□

E MISSING DISCUSSIONS AND PROOFS

E.1 Linear Hierarchy With Non-Linear Rewards

In this subsection, the action and latent parameters are generated as in (9). But the reward is no longer Gaussian. Precisely, let $\mathrm{Ber}(p)$ be a Bernoulli distribution with mean p, and $g(u) = 1/(1+\exp(-u))$ be the logistic function. We define $p(\cdot \mid x; \theta_a) = \mathrm{Ber}(g(\phi(x)^\top \theta_a))$ which yields

$$\begin{aligned}\psi &\sim \mathcal{N}(\mu, \Sigma)\,, && \\ \theta_a \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_a \psi,\, \Sigma_a\Big)\,, && \forall a \in \mathcal{A}\,, \\ R \mid \psi, \theta, X, A &\sim \mathrm{Ber}(g(\phi(X)^\top \theta_A))\,. &&\end{aligned} \tag{25}$$

Step (A). In this case, we cannot derive closed-form posteriors. But we can approximate the likelihoods $\mathcal{L}_a(\cdot)$ by multivariate Gaussian distributions such as

$$\mathcal{L}_a(\cdot) \approx \mathcal{N}(\cdot; \breve{\mu}_a, (\breve{G}_a)^{-1})\,. \tag{26}$$

where $\breve{\mu}_a$ and $\breve{G}_a$ are the maximum likelihood estimator (MLE) and the Hessian of the negative log-likelihood $-\log \mathcal{L}_a(\cdot)$, respectively:

$$\begin{aligned}\breve{\mu}_a &= \operatorname*{argmax}_{\theta \in \mathbb{R}^d} \log \mathcal{L}_a(\theta)\,, \\ \breve{G}_a &= \sum_{i \in [n]} \mathbb{1}\{A_i = a\} \dot{g}\left(X_i^\top \breve{\mu}_a\right) \phi(X_i) \phi(X_i)^\top\,.\end{aligned}$$

The latter approximation is slightly different from the popular Laplace approximation[2] and is sometimes referred to as Bernstein-von-Mises approximation, see [Schillings et al., 2020, Remark 3] for a detailed discussion. It is simpler and more computationally friendly than the Laplace approximation in our setting. Indeed, interestingly, (26) allows us to use the Gaussian posteriors (10), (11) and (12) in **Step (A)** in Section 4, except that G_a and B_a are replaced by $\breve{G}_a$ and $\breve{G}_a \breve{\mu}_a$, respectively (notice that the inverse $\breve{G}_a^{-1}$ may not exist but it is not used in our formulas). There is a clear intuition behind this. First, $G_a \leftarrow \breve{G}_a$ captures the change of curvature due to the nonlinearity of the mean function g. Moreover, $B_a \leftarrow \breve{G}_a \breve{\mu}_a$ follows from the fact that the MLE of the action parameter θ_a with linear rewards in (9) is $G_a^{-1} B_a$, and it corresponds to $\breve{\mu}_a$ in this case.

Step (B). The reward function under (25) is $r(x, a; \theta) = g(\phi(x)^\top \theta)$. Thus our reward estimate is

$$\hat{r}(x, a) = g\Big(\frac{\phi(x)^\top \hat{\mu}_a}{\sqrt{1 + \pi/8 \|\phi(x)\|_{\hat{\Sigma}_a}}}\Big)\,,$$

[2] Laplace approximation's mean is given by the maximum a-posteriori estimate (MAP) of the posterior and its covariance is the inverse Hessian of the negative log posterior density.

where we use an approximation of the expectation of the sigmoid function under a Gaussian distribution [Spiegelhalter and Lauritzen, 1990, Appendix A].

Step (C). Since g is increasing, the learned policy is

$$\hat{\pi}_{\mathrm{G}}(a \mid x) = \mathbb{1}\Big\{a = \operatorname*{argmax}_{b \in \mathcal{A}} \frac{\phi(x)^\top \hat{\mu}_b}{\sqrt{1 + \pi/8 \|\phi(x)\|_{\hat{\Sigma}_b}}}\Big\}.$$

E.2 Frequentist vs. Bayesian Suboptimality

Before diving into our main result, a Bayesian suboptimality bound for the structured prior in (3), we motivate its use by analyzing both Bayesian and frequentist suboptimality for the simpler non-structured prior in (2). While this analysis uses the non-structured prior, our main result applies directly to the structured one. The key distinction between Bayesian and frequentist suboptimality lies in their confidence intervals, as discussed in [Hong et al., 2023, Section 4.2]. However, we still provide a self-contained explanation for completeness, considering our slightly different results and proofs. Also, note that Hong et al. [2023] also explored a notion of Bayesian suboptimality that they defined as $V(\pi_*; \theta_*) - V(\hat{\pi}; \theta_*) \mid S$. Precisely, they fix the sample set S and consider only randomness in θ_* sampled from the prior. In contrast, our definition aligns with traditional online bandits (Bayesian regret) by averaging suboptimality across both the sample set and θ_* sampled from the prior: $\mathbb{E}\left[V(\pi_*; \theta_*) - V(\hat{\pi}; \theta_*)\right]$. These metrics are different and have distinct implications. Interestingly, ours favors the greedy policy (Appendix E.5), while theirs incentivizes pessimistic policies.

Frequentist confidence intervals. Assume that for any $a \in \mathcal{A}$, there exists a parameter $\theta_{*,a} \in \mathbb{R}^d$ generating the reward as $R \mid \theta_*, X, A \sim \mathcal{N}(\phi(X)^\top \theta_{*,A}, \sigma^2)$. Then the posterior under the non-structured prior in (2) satisfies

$$\mathbb{P}\left(\forall x \in \mathcal{X} \,:\, |r(x, a; \theta_*) - \hat{r}(x, a)| \leq \left(\alpha(d, \delta) + \frac{c_a + \|\mu_a\|_2}{\sqrt{\lambda_d(\Sigma_a)}}\right) \|\phi(X)\|_{\hat{\Sigma}_a} \mid \tilde{S}\right) \geq 1 - \delta\,, \tag{27}$$

where $\tilde{S} = (X_i, A_i)_{i \in [n]}$, the randomness in $\mathbb{P}$ is over the true reward noise $\mathcal{N}(0, \sigma^2)$, and $\alpha(d, \delta) = \sqrt{d + 2\sqrt{d \log \frac{1}{\delta}} + 2 \log \frac{1}{\delta}}$. Now let us compare this confidence interval to the Bayesian one below.

Bayesian confidence intervals. Here we further assume that for any $a \in \mathcal{A}$, the action parameter $\theta_{*,a}$ is generated from a know Gaussian distribution as $\theta_{*,a} \sim \mathcal{N}(\mu_a, \Sigma_a)$ whose parameters μ_a, Σ_a match the ones used in the non-structured prior (2). Then the posterior under the non-structured prior in (2) satisfies

$$\mathbb{P}\left(\forall x \in \mathcal{X} \,:\, |r(x, a; \theta_*) - \hat{r}(x, a)| \leq \alpha(d, \delta) \|\phi(X)\|_{\hat{\Sigma}_a} \,\middle|\, S\right) \geq 1 - \delta \tag{28}$$

Here the randomness in $\mathbb{P}$ is over the action parameter uncertainty $\mathcal{N}(0, \Sigma_a)$ as we condition on $S = (X_i, A_i, R_i)_{i \in [n]}$. The proof is given after Lemma E.1. In Bayesian bounds, $\theta_{*,a}$ is treated as a random variable with distribution $\mathcal{N}(\mu_a, \Sigma_a)$ and it is included in the randomness in $\mathbb{P}$. The randomization of the data $(X_i, A_i, R_i)_{i \in [n]}$ is also included in $\mathbb{P}$. However, we condition on S, and hence they are treated as fixed. In contrast with frequentist bounds where $\theta_{*,a}$ is fixed while the reward is random and included in the randomness in $\mathbb{P}$ without conditioning on it. The Bayesian confidence intervals capture that the agent has access to the prior on the true action parameters and hence the additional misspecification term $\frac{c_a + \|\mu_a\|_2}{\sqrt{\lambda_d(\Sigma_a)}}$ is not present, as opposed to its frequentist counterpart. This is the main reason why BSO captures the benefits of knowing and leveraging informative priors.

Lemma E.1 (Bayesian bound). *Let $a \in \mathcal{A}$, $\delta \in (0, 1)$ and $\alpha(d, \delta) = \sqrt{d + 2\sqrt{d \log \frac{1}{\delta}} + 2 \log \frac{1}{\delta}}$. Then*

$$\mathbb{P}\left(\forall x \in \mathcal{X} \,:\, |r(x, a; \theta_*) - r(x, a; \hat{\mu})| \leq \alpha(d, \delta) \|\phi(X)\|_{\hat{\Sigma}_a} \,\middle|\, S\right) \geq 1 - \delta \tag{29}$$

Proof. First, we have that

$$\begin{aligned} |r(x, a; \theta_*) - r(x, a; \hat{\mu})| &= |\phi(X)^\top \theta_{*,a} - \phi(X)^\top \hat{\mu}_a| = |\phi(X)^\top (\theta_{*,a} - \hat{\mu}_a)|\,, \\ &= |\phi(X)^\top \hat{\Sigma}_a^{\frac{1}{2}} \hat{\Sigma}_a^{-\frac{1}{2}} (\theta_a - \hat{\mu}_a)| \leq \|\phi(X)\|_{\hat{\Sigma}_a} \|\theta_a - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}}\,, \end{aligned}$$

where we use the Cauchy-Schwarz inequality in the last step. Now we assumed that the true action parameters $\theta_{*,a}$ are random and their prior distribution matches our model in (9). Therefore, $\theta_{*,a} \mid S$ have the same density as our posterior $\theta_a \mid S$, and hence $\theta_{*,a} \mid S \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$. This means that $\theta_{*,a} - \hat{\mu}_a \mid S \sim \mathcal{N}(0, \hat{\Sigma}_a)$. 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

$$\mathbb{P}\left(\|\theta_{*,a} - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}} \leq \alpha(d,\delta) \,\middle|\, S\right) \geq 1 - \delta\,.$$

Finally using that for any $x \in \mathcal{A}$, $|r(x,a;\theta_*) - r(x,a;\hat{\mu})| \leq \|\phi(X)\|_{\hat{\Sigma}_a} \|\theta_a - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}}$ concludes the proof. □

Lemma E.2 (Frequentist bound)**.** *Let $a \in \mathcal{A}$, $\delta \in (0,1)$ and $\alpha(d,\delta) = \sqrt{d + 2\sqrt{d \log \frac{1}{\delta}} + 2\log \frac{1}{\delta}}$. Then*

$$\mathbb{P}\left(\forall x \in \mathcal{X} \,:\, |r(x,a;\theta_*) - r(x,a;\hat{\mu})| \leq \left(\alpha(d,\delta) + \frac{c_a + \|\hat{\mu}_a\|_2}{\sqrt{\lambda_d(\Sigma_a)}}\right) \|\phi(X)\|_{\hat{\Sigma}_a} \,\middle|\, \tilde{S}\right) \geq 1 - \delta\,, \tag{30}$$

where $\tilde{S} = (X_i, A_i)_{i \in [n]}$.

Proof. Similarly to Lemma E.1, we have that $|r(x,a;\theta_*) - r(x,a;\hat{\mu})| \leq \|\phi(X)\|_{\hat{\Sigma}_a} \|\theta_a - \hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}}$. Note that (X_i, A_i) and $\theta_{*,a}$ are fixed; the randomness only comes from R_i for $i \in [n]$. Thus $\hat{\Sigma}_a$ is fixed while $\hat{\mu}_a$ is random. Keeping this in mind, we have that $R_i \sim \mathcal{N}(\phi(X_i)^\top \theta_{*,a}, \sigma^2)$ and thus

$$\sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} \phi(X_i) R_i \sim \mathcal{N}\Big(\sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} \phi(X_i)\phi(X_i)^\top \theta_{*,a}, \sigma^2 \sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} \phi(X_i)\phi(X_i)^\top\Big)\,. \tag{31}$$

But $\sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} \phi(X_i) R_i = \sigma^2 B_a$ and $\sum_{i \in [n]} \mathbb{I}_{\{A_i = a\}} \phi(X_i)\phi(X_i)^\top = \sigma^2 G_a$. Thus, we get that

$$\sigma^2 B_a \sim \mathcal{N}(\sigma^2 G_a \theta_{*,a}, \sigma^4 G_a)\,, \tag{32}$$

and hence

$$B_a \sim \mathcal{N}(G_a \theta_{*,a}, G_a)\,. \tag{33}$$

Now since Σ_a, μ_a and $\hat{\Sigma}_a$ are fixed, we have that

$$\hat{\mu}_a = \hat{\Sigma}_a(\Sigma_a^{-1}\mu_a + B_a) \sim \mathcal{N}(\hat{\Sigma}_a(\Sigma_a^{-1}\mu_a + G_a\theta_{*,a}), \hat{\Sigma}_a G_a \hat{\Sigma}_a)\,. \tag{34}$$

But notice that $G_a\theta_{*,a} = (G_a + \Sigma_a^{-1})\theta_{*,a} - \Sigma_a^{-1}\theta_{*,a} = \hat{\Sigma}_a^{-1}\theta_{*,a} - \Sigma_a^{-1}\theta_{*,a}$ and hence (34) becomes

$$\hat{\mu}_a \sim \mathcal{N}(\hat{\Sigma}_a\Sigma_a^{-1}(\mu_a - \theta_{*,a}) + \theta_{*,a}, \hat{\Sigma}_a G_a \hat{\Sigma}_a)\,. \tag{35}$$

Thus

$$\hat{\mu}_a - \theta_{*,a} \sim \mathcal{N}(\hat{\Sigma}_a\Sigma_a^{-1}(\mu_a - \theta_{*,a}), \hat{\Sigma}_a G_a \hat{\Sigma}_a)\,. \tag{36}$$

Multiplying by $\hat{\Sigma}_a^{-\frac{1}{2}}$ leads to

$$\hat{\Sigma}_a^{-\frac{1}{2}}(\hat{\mu}_a - \theta_{*,a}) \sim \mathcal{N}(\hat{\Sigma}_a^{\frac{1}{2}}\Sigma_a^{-1}(\mu_a - \theta_{*,a}), \hat{\Sigma}_a^{\frac{1}{2}} G_a \hat{\Sigma}_a^{\frac{1}{2}})\,. \tag{37}$$

Now we rewrite (37) as

$$\hat{\Sigma}_a^{-\frac{1}{2}}(\hat{\mu}_a - \theta_{*,a}) = \hat{\Sigma}_a^{\frac{1}{2}}\Sigma_a^{-1}(\mu_a - \theta_{*,a}) + \hat{\Sigma}_a^{\frac{1}{2}} G_a^{\frac{1}{2}} Z\,, \qquad Z \sim \mathcal{N}(0, I_d)\,. \tag{38}$$

Thus we have that

$$\begin{aligned}
\|\hat{\Sigma}_a^{-\frac{1}{2}}(\hat{\mu}_a - \theta_{*,a})\|_2 = \|\hat{\mu}_a - \theta_{*,a}\|_{\hat{\Sigma}_a^{-1}} &= \|\hat{\Sigma}_a^{\frac{1}{2}}\Sigma_a^{-1}(\mu_a - \theta_{*,a}) + \hat{\Sigma}_a^{\frac{1}{2}} G_a^{\frac{1}{2}} Z\|_2\,, \\
&\leq \|\hat{\Sigma}_a^{\frac{1}{2}}\Sigma_a^{-1}(\mu_a - \theta_{*,a})\|_2 + \|\hat{\Sigma}_a^{\frac{1}{2}} G_a^{\frac{1}{2}} Z\|_2\,, \\
&\leq \frac{\|\mu_a - \theta_{*,a}\|_2}{\sqrt{\lambda_d(\Sigma_a)}} + \|Z\|_2\,.
\end{aligned} \tag{39}$$

Finally, since $Z \sim \mathcal{N}(0, I_d)$, we apply Laurent and Massart [2000, Lemma 1] and get that

$$\mathbb{P}(\|Z\|_2 \leq \alpha(d,\delta)) \geq 1-\delta\,.$$

Combining this with (39) leads to

$$\mathbb{P}\left(\|\theta_{*,a}-\hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}} \leq \frac{\|\mu_a-\theta_{*,a}\|_2}{\sqrt{\lambda_d(\Sigma_a)}}+\alpha(d,\delta)\,\middle|\,(X_i,A_i)_{i\in[n]}\right) \geq 1-\delta\,.$$

Finally using that for any $x \in \mathcal{A}$, $|r(x,a;\theta_*)-r(x,a;\hat{\mu})| \leq \|\phi(X)\|_{\hat{\Sigma}_a}\|\theta_a-\hat{\mu}_a\|_{\hat{\Sigma}_a^{-1}}$ and using that $\|\theta_{*,a}-\hat{\mu}_a\|_2 \leq c_a+\|\hat{\mu}_a\|_2$ concludes the proof. □

E.3 Main Result

In this section, we prove Theorem 5.1. Recall that we make the following well-specified prior assumption.

Assumption E.3 (Well-specified priors)**.** Action parameters $\theta_{*,a}$ and rewards are drawn from (9).

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 $\hat{a}_x=\operatorname{argmax}_{b\in\mathcal{A}}\hat{r}(x,b)$. 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.

$$\begin{aligned}
\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) &= \mathbb{E}\left[V(\pi_*;\theta_*)-V(\hat{\pi};\theta_*)\right]\,,\\
&= \mathbb{E}\left[r(X,\pi_*(X);\theta_*)-r(X,\hat{\pi}_{\mathrm{G}}(X);\theta_*)\right]\,,\\
&= \mathbb{E}\left[r(X,\pi_*(X);\theta_*)-r(X,\hat{\pi}_{\mathrm{G}}(X);\hat{\mu})+r(X,\hat{\pi}_{\mathrm{G}}(X);\hat{\mu})-r(X,\hat{\pi}_{\mathrm{G}}(X);\theta_*)\right]\,,\\
&\leq \mathbb{E}\left[r(X,\pi_*(X);\theta_*)-r(X,\pi_*(X);\hat{\mu})+r(X,\hat{\pi}_{\mathrm{G}}(X);\hat{\mu})-r(X,\hat{\pi}_{\mathrm{G}}(X);\theta_*)\right]\,,\\
&\leq \mathbb{E}\left[r(X,\pi_*(X);\theta_*)-r(X,\pi_*(X);\hat{\mu})\right]+\mathbb{E}\left[r(X,\hat{\pi}_{\mathrm{G}}(X);\hat{\mu})-r(X,\hat{\pi}_{\mathrm{G}}(X);\theta_*)\right]\,.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}\left[r(X,\hat{\pi}_{\mathrm{G}}(X);\hat{\mu})-r(X,\hat{\pi}_{\mathrm{G}}(X);\theta_*)\right] &= \mathbb{E}\left[\mathbb{E}\left[r(X,\hat{\pi}_{\mathrm{G}}(X);\hat{\mu})-r(X,\hat{\pi}_{\mathrm{G}}(X);\theta_* \mid X,S\right]\right]\,,\\
&= \mathbb{E}\left[\mathbb{E}\left[\phi(X)^\top\hat{\mu}_{\hat{\pi}_{\mathrm{G}}(X)}-\phi(X)^\top\theta_{*,\hat{\pi}_{\mathrm{G}}(X)} \mid X,S\right]\right]\,,\\
&\overset{(i)}{=} \mathbb{E}\left[\mathbb{E}\left[\Phi(X,\hat{\pi}_{\mathrm{G}}(X))^\top\hat{\mu}-\Phi(X,\hat{\pi}_{\mathrm{G}}(X))^\top\theta_* \mid X,S\right]\right]\,,\\
&\overset{(ii)}{=} \mathbb{E}\left[\Phi(X,\hat{\pi}_{\mathrm{G}}(X))^\top\mathbb{E}\left[\hat{\mu}-\theta_* \mid X,S\right]\right]\,,\\
&\overset{(iii)}{=} \mathbb{E}\left[\Phi(X,\hat{\pi}_{\mathrm{G}}(X))^\top(\hat{\mu}-\mathbb{E}\left[\theta_* \mid X,S\right])\right]\,,\\
&\overset{(iv)}{=} 0\,.
\end{aligned}$$

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 S. In (iii), we used that $\hat{\mu}$ is deterministic given X and S. Finally, in (iv), we used that $\mathbb{E}\left[\theta_* \mid X,S\right]=\mathbb{E}\left[\theta_* \mid S\right]=\hat{\mu}$, 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 S, 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 $\delta\in(0,1)$, we define the following high-probability events

$$E_a=\left\{\forall x\in\mathcal{X}\,:\,|r(x,a;\theta_*)-r(x,a;\hat{\mu})| \leq \alpha(d,\delta)\|\phi(x)\|_{\hat{\Sigma}_a}\right\}\,, \qquad \forall a\in\mathcal{A}\,,$$

where $\alpha(d,\delta) = \sqrt{d + 2\sqrt{d\log\frac{1}{\delta}} + 2\log\frac{1}{\delta}}$. Then we decompose $\textsc{Bso}(\hat{\pi}_{\mathrm{G}})$ as

$$
\begin{aligned}
&\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) \leq \mathbb{E}\left[r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})\right] , \\
&\leq \mathbb{E}\left[|r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})|\right] , \\
&\leq \mathbb{E}\left[|r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})|\mathbb{1}\{E_{\pi_*(X)}\}\right] + \mathbb{E}\left[|r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})|\mathbb{1}\{\bar{E}_{\pi_*(X)}\}\right] , \\
&\leq \alpha(d,\delta)\mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] + \mathbb{E}\left[|r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})|\mathbb{1}\{\bar{E}_{\pi_*(X)}\}\right] .
\end{aligned}
$$

Now we deal with the term $\mathbb{E}\left[|r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})|\mathbb{1}\{\bar{E}_{\pi_*(X)}\}\right]$. Let $Z_a = r(X,a;\theta_*) - r(X,a;\hat{\mu})$, so that $Z_{\pi_*(X)} = r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})$. Then we have that

$$
\begin{aligned}
&\mathbb{E}\left[|r(X,\pi_*(X);\theta_*) - r(X,\pi_*(X);\hat{\mu})|\mathbb{1}\{\bar{E}_{\pi_*(X)}\}\right] = \mathbb{E}\left[|Z_{\pi_*(X)}|\mathbb{1}\Big\{|Z_{\pi_*(X)}| > \alpha(d,\delta)\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\Big\}\right] , \\
&= \mathbb{E}\left[\mathbb{E}\left[|Z_{\pi_*(X)}|\mathbb{1}\Big\{|Z_{\pi_*(X)}| > \alpha(d,\delta)\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\Big\} \mid X, S\right]\right] , \\
&\overset{(i)}{\leq} \mathbb{E}\left[\frac{2}{\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\sqrt{2\pi}}\int_{u=\alpha(d,\delta)\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}}^{\infty} u\exp\left[-\frac{u^2}{2\|\phi(X)\|^2_{\hat{\Sigma}_{\pi_*(X)}}}\right]\mathrm{d}u\right] , \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\frac{2}{\sqrt{2\pi}}\int_{u=\alpha(d,\delta)}^{\infty} u\exp\left[-\frac{u^2}{2}\right]\mathrm{d}u\right] , \\
&\overset{(iii)}{\leq} \sqrt{\frac{2}{\pi}}\exp\Big(-\frac{\alpha(d,\delta)^2}{2}\Big)\mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] ,
\end{aligned}
$$

where (i) follows from the facts that $Z_{\pi_*(X)} \mid X, S \sim \mathcal{N}(0, \|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}})$. In (ii), we use the change of variables $u \leftarrow u/\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}$. Finally, in (iii), we compute the integral. This leads to the following result

$$
\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) \leq \left(\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{\alpha(d,\delta)^2}{2}\Big) + \alpha(d,\delta)\right)\mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] . \tag{40}
$$

But this holds for any $\delta \in (0,1)$. In particular, it holds for $\delta \to 1$ leading to

$$
\textsc{Bso}(\hat{\pi}_{\mathrm{G}}) \leq \left(\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{d}{2}\Big) + \sqrt{d}\right)\mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right] , \tag{41}
$$

since $\lim_{\delta\to 1}\alpha(d,\delta) = \alpha(d,1) = \sqrt{d}$. But $\sqrt{\frac{2}{\pi}}\exp\Big(-\frac{d}{2}\Big) \leq \sqrt{d}$ which concludes the proof.

E.4 Proof of Theorem 5.3

Proof. The proof is similar to that of Theorem 5.1. Precisely, we have that

$$
\begin{aligned}
\textsc{Bmse}(\hat{r}(x,a)) &= \mathbb{E}\big[\big(\hat{r}(x,a) - r(x,a;\theta_*)\big)^2\big] , \\
&= \mathbb{E}\big[\mathbb{E}\big[\big(\hat{r}(x,a) - r(x,a;\theta_*)\big)^2 \mid S\big]\big] , \\
&= \mathbb{E}\big[\mathbb{E}\big[\big(\phi(x)^\top\hat{\mu}_a - \phi(x)^\top\theta_{*,a}\big)^2 \mid S\big]\big] , \\
&= \mathbb{E}\big[\mathbb{E}\big[\big(\phi(x)^\top\theta_{*,a} - \phi(x)^\top\hat{\mu}_a\big)^2 \mid S\big]\big] .
\end{aligned}
$$

But we assumed that the true action parameters $\theta_{*,a}$ are random and their prior distribution matches our model in (9). Therefore, $\theta_{*,a} \mid S$ have the same density as our posterior $\theta_a \mid S$, and hence $\theta_{*,a} \mid S \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$. This means that $\phi(x)^\top\theta_{*,a} \mid S \sim \mathcal{N}(\phi(x)^\top\hat{\mu}_a, \|\phi(x)\|^2_{\hat{\Sigma}_a})$. But $\mathbb{E}\big[\big(\phi(x)^\top\theta_{*,a} - \phi(x)^\top\hat{\mu}_a\big)^2 \mid S\big]$ is exactly the variance of $\phi(x)^\top\theta_{*,a} \mid S$ and thus it is equal to $\|\phi(x)\|^2_{\hat{\Sigma}_a}$. Taking the expectation concludes the proof.

□

E.5 Optimality of Greedy Policies

Here, we show that Greedy policy $\hat{\pi}_{\mathrm{G}}$ should be preferred to any other choice of policies when considering the BSO as our performance metric. This is because $\hat{\pi}_{\mathrm{G}}$ minimizes the BSO. To see this, note that by definition the Greedy policy $\hat{\pi}_{\mathrm{G}}$ is deterministic, that is for any context $x \in \mathcal{X}$, there exists $\hat{a}_{\mathrm{G}}$, such that $\hat{\pi}_{\mathrm{G}}(\hat{a}_{\mathrm{G}} \mid x) = 1$. Thus, for any context $x \in \mathcal{X}$, we simplify the notation by letting $\hat{\pi}_{\mathrm{G}}(x)$ denote the action that has a mass equal to 1. Then, we have that

$$\mathbb{E}_{A\sim\hat{\pi}_{\mathrm{G}}(\cdot|x)}\left[\mathbb{E}_{\theta_*}\left[r(x,A;\theta_*) \mid S\right]\right] = \mathbb{E}_{\theta_*}\left[r(x,\hat{\pi}_{\mathrm{G}}(x);\theta_*) \mid S\right] \geq \mathbb{E}\left[r(x,a;\theta_*) \mid S\right] \qquad \forall x, a \in \mathcal{X}\times\mathcal{A}\,. \tag{42}$$

where this follows from the definition of $\hat{r}(x,a) = \mathbb{E}\left[r(x,a;\theta) \mid S\right]$, the definition of $\hat{\pi}_{\mathrm{G}}$ and the fact that θ_* is sampled from the prior, which leads to $\mathbb{E}\left[r(x,a;\theta) \mid S\right] = \mathbb{E}\left[r(x,a;\theta_*) \mid S\right]$. Now (42) 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,

$$\mathbb{E}_{X\sim\nu, A\sim\hat{\pi}_{\mathrm{G}}(\cdot|X)}\left[\mathbb{E}_{\theta_*}\left[r(x,A;\theta_*) \mid S\right]\right] \geq \mathbb{E}_{X\sim\nu, A\sim\pi(\cdot|X)}\left[\mathbb{E}_{\theta_*}\left[r(x,A;\theta_*) \mid S\right]\right]\,. \tag{43}$$

Taking another expectation w.r.t. the sample set S and using Fubini's theorem and the tower rule leads to $\mathbb{E}\left[V(\hat{\pi}_{\mathrm{G}};\theta_*)\right] \geq \mathbb{E}\left[V(\pi;\theta_*)\right]$ 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 $\hat{\pi}_{\mathrm{G}}$ compared to any stationary policy π, in particular, compared to the policy π_{P} induced by pessimism.

E.6 Discussing the Main Assumption

In this section, we discuss our main assumption of the existence of the latent parameter ψ. Let us take the Gaussian case, where were assume that

$$\begin{aligned} \psi &\sim \mathcal{N}(\mu, \Sigma)\,, \\ \theta_a \mid \psi &\sim \mathcal{N}\Big(\mathrm{W}_a\psi,\, \Sigma_a\Big)\,, \qquad\qquad \forall a \in \mathcal{A}\,. \end{aligned} \tag{44}$$

Now we discuss if (44) is as mild as assuming that the action parameters are jointly sampled as $\theta \sim \mathcal{N}(\mu_{\mathrm{ACT}}, \Sigma_{\mathrm{ACT}})$, where $\mu_{\mathrm{ACT}} \in \mathbb{R}^{dK}$ and $\Sigma_{\mathrm{ACT}} \in \mathbb{R}^{dK\times dK}$. To do so, assume that $\theta \sim \mathcal{N}(0, \Sigma_{\mathrm{ACT}})$, where we set $\mu_{\mathrm{ACT}} = 0$ to simplify notation. Then we want to prove that there exists $\psi \sim \mathcal{N}(0, I_{d'})$, with $\mu \in \mathbb{R}^{d'}$ for some $d' \geq 1$ and $\mathrm{W} \in \mathbb{R}^{dK\times d'}$ such that $\theta \mid \psi \sim \mathcal{N}(\mathrm{W}\psi, \tilde{\Sigma})$, where $\tilde{\Sigma}$ is diagonal by $d\times d$ blocks. First, note that (θ, ψ) is also Gaussian with a mean equal to 0 and covariance $\Omega = \begin{pmatrix} \Sigma_{\mathrm{ACT}} & \Omega_2 \\ \Omega_2^\top & I_{d'} \end{pmatrix} \in \mathbb{R}^{(dK+d')\times(dK+d')}$. Then it suffices to find W such that $\theta - \mathrm{W}\psi \perp\!\!\!\perp \psi$. That is, W such that $\mathrm{cov}(\theta - \mathrm{W}\psi, \psi) = \mathrm{cov}(\theta,\psi) - \mathrm{W}\mathrm{cov}(\psi,\psi) = \Omega_2 - \mathrm{W} = 0$, which leads to $\mathrm{W} = \Omega_2$. Using this conditional independence and that $\mathrm{W} = \Omega_2$, we get that the distribution of $\theta - \mathrm{W}\psi \mid \psi$ is the same as the distribution of $\theta - \mathrm{W}\psi$ and it writes $\mathcal{N}(0, \Sigma_{\mathrm{ACT}} - \mathrm{W}\mathrm{W}^\top)$. Thus, $\theta \mid \psi \sim \mathcal{N}(\mathrm{W}\psi, \Sigma_{\mathrm{ACT}} - \mathrm{W}\mathrm{W}^\top)$. But Σ_{ACT} is positive-definite and hence there exists $\lambda > 0$ such that $\Sigma_{\mathrm{ACT}} - \lambda I_{dK}$ is symmetric and positive semi-definite. Therefore, $\Sigma_{\mathrm{ACT}} - \lambda I_{dK}$ can be written as $\Sigma_{\mathrm{ACT}} - \lambda I_{dK} = \mathrm{B}\mathrm{B}^\top$ where $\mathrm{B} \in \mathbb{R}^{dK\times d'}$ is such that $d' \geq \mathrm{rank}(\Sigma_{\mathrm{ACT}} - \lambda I_{dK})$ and setting $\mathrm{W} = \mathrm{B}$ concludes the proof.

E.7 Explicit Bound

Assumptions. Before presenting the explicit bound, we introduce some assumptions to simplify the notation, though the results hold under more general conditions.

- **(A0)** We assume $\Sigma_a = \sigma_0^2 I_d$, $\Sigma = \tau^2 I_{d'}$, $\phi(x) = x$, $\|x\|_2 \leq 1$, and the matrices W_a are normalized such that $\lambda_1(\mathrm{W}_a\mathrm{W}_a^\top) = \lambda_d(\mathrm{W}_a\mathrm{W}_a^\top) = 1$.

Assumption **(A0)** can be relaxed; the theory holds for any positive definite matrices Σ_a and Σ, bounded L_2 norm contexts x, and any matrices W_a. Additionally, we make the following necessary assumptions:

- **(A1)** Let $G = \mathbb{E}_{X\sim\nu}[XX^\top]$ with $g = \lambda_d(G)$. We assume that $g > 0$.
- **(A2)** The fourth moment of X is bounded, and X satisfies $\sqrt{\mathbb{E}[(v^\top XX^\top v)^2]} \leq \mathrm{h} v^\top G v$ for all $v \in \mathbb{R}^d$. Additionally, X and $X \mid A$ follow the same distribution.

Theorem E.4 (Explicit Bound). *Let $\pi_*(x)$ be the optimal action for context x. Under **(A0)**, **(A1)** and **(A2)**, the BSO of* sDM *under the structured prior (9) satisfies*

$$\mathrm{BSO}(\hat{\pi}_{\mathrm{G}}) \leq 2\sqrt{\mathbb{E}_{X\sim\nu}\left[\frac{d}{\sigma^{-2}g\alpha_X+\sigma_0^{-2}}+\frac{\tau^2\sigma_0^{-4}d}{\left(\sigma^{-2}g\alpha_X+\sigma_0^{-2}\right)^2}+(\sigma_0^2+\tau^2)d\exp\Big(-\frac{n\pi_0^2(\pi_*(X)|X)}{2}\Big)\right]+\frac{2d(\sigma_0^2+\tau^2)}{n}},$$

where $\alpha_x = \lfloor\frac{\pi_0(\pi_*(x)|x)}{2}n\rfloor - 7\mathrm{h}\sqrt{\lfloor\frac{\pi_0(\pi_*(x)|x)}{2}n\rfloor(d+2\ln(n))}$ *for any* $x\in\mathcal{X}$ *and* $a\in\mathcal{A}$.

Scaling with n. Note that

$$\alpha_X = \lfloor\frac{\pi_0(\pi_*(x)|x)}{2}n\rfloor - 7\mathrm{h}\sqrt{\lfloor\frac{\pi_0(\pi_*(x)|x)}{2}n\rfloor(d+2\ln(n))}\,, \qquad \forall(x,a)\in\mathcal{X}\times\mathcal{A}\,. \tag{45}$$

In particular, with enough data n large enough, there exists a constant $c>0$ such that

$$\alpha_x \geq c\pi_0(\pi_*(x)|x)n\,, \qquad \forall(x,a)\in\mathcal{X}\times\mathcal{A}\,.$$

Therefore, for n large enough, the bound could be expressed as

$$\begin{aligned}\mathrm{BSO}(\hat{\pi}_{\mathrm{G}}) &= \mathcal{O}\left(\sqrt{\mathbb{E}_{X\sim\nu}\left[\frac{d}{\pi_0(\pi_*(X)|X)n+1}+\frac{d}{(\pi_0(\pi_*(X)|X)n+1)^2}+d\exp\Big(-n\pi_0^2(\pi_*(X)|X)\Big)\right]}+\frac{d}{n}\right),\\ &= \mathcal{O}\left(\sqrt{\mathbb{E}_{X\sim\nu}\left[\frac{d}{\pi_0(\pi_*(X)|X)n+1}+d\exp\Big(-n\pi_0^2(\pi_*(X)|X)\Big)\right]}+\frac{d}{n}\right),\end{aligned}$$

where we omitted constants and used that $\frac{d}{(\pi_0(\pi_*(X)|X)n+1)^2} = \mathcal{O}(\frac{d}{\pi_0(\pi_*(X)|X)n+1})$ leads to the scaling provided in Theorem 5.2.

Proof idea. From Theorem 5.1, we establish that

$$\mathrm{BSO}(\hat{\pi}_{\mathrm{G}}) \leq 2\sqrt{d}\mathbb{E}\left[\|X\|_{\hat{\Sigma}_{\pi_*(X)}}\right],$$

where we assume that $\phi(x)=x$. Thus we need to control the scaling of the posterior covariance matrix $\hat{\Sigma}_{\pi_*(X)}$ with n. First, note that for a fixed context x, the only randomness in $\hat{\Sigma}_{\pi_*(x)}$ originates from the design matrix $\hat{G}_{\pi_*(x)}$. Previous works simplify this by assuming the data is well-explored [Hong et al., 2023], i.e., $\hat{G}_a \succ \gamma nG$ for all $a\in\mathcal{A}$, where $G=\mathbb{E}_{X\sim\nu}\left[XX^\top\right]$. We instead only need to control $\hat{G}_{\pi_*(x)}$, and instead of assuming it satisfies $\hat{G}_{\pi_*(x)}\succ\gamma nG$, we will consider the events where this happens and the events where this is violated. We can do this based on when $\hat{G}_{\pi_*(x)}$ is the sum of n out-products. The problem, however, is that in this setting, $\hat{G}_{\pi_*(x)}$ is the sum of N_a outer products, where the random variable N_a of the number of times the optimal action $\pi_*(X)$ appears in the sample set. To address this, we need to control N_a. Fortunately, we notice that $N_a|X$ follows a Binomial distribution with carefully chosen parameters. Using Hoeffding's inequality, we can bound the tails of this distribution, ensuring that the optimal action appears sufficiently often in the dataset. Once the optimal action counts are bounded, the results from Oliveira [2016] can be adapted to bound $\hat{G}_{\pi_*(x)}$. Matrix manipulations combined with eigenvalue inequalities then lead to the desired result.

Proof. From Theorem 5.1, we have that

$$\mathrm{BSO}(\hat{\pi}_{\mathrm{G}}) \leq 2\sqrt{d}\mathbb{E}\left[\|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}}\right], \tag{46}$$

$$= 2\sqrt{d}\mathbb{E}\left[\|X\|_{\hat{\Sigma}_{\pi_*(X)}}\right], \tag{47}$$

$$\leq 2\sqrt{d\,\mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}}\right]}, \tag{48}$$

where we used that the simplifying assumption that $\phi(x) = x$ in the second equality, and Cauchy-Schwartz in the third inequality. 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

$$\lambda_1(\hat{\Sigma}_a) \leq \sigma_0^2 + \tau^2 \,, \qquad \forall a \in \mathcal{A} \,, \tag{49}$$

since $\lambda_1(\mathrm{W}_a \mathrm{W}_a^\top) = 1$. These are obtained using Weyl's inequalities.

Now we introduce some quantities:

- N_a : number of samples where $A_i = a$.
- $\epsilon_{x,a} = \mathbb{P}(A = a \mid x) = \pi_0(a \mid x)$ is the probability of choosing the action a given x.

Remark that $n - N_a \mid X \sim \mathrm{Bin}(n, 1 - \epsilon_{X,a})$. Therefore, via Hoeffding inequality, we can prove that for any $t > 0$ and $a \in \mathcal{A}$

$$\mathbb{P}(n - N_a - n(1 - \epsilon_{X,a}) > t \mid X) \leq \exp(-2t^2/n) \,,$$

which simplifies to

$$\mathbb{P}(N_a < n\epsilon_{X,a} - t \mid X) \leq \exp(-2t^2/n) \,.$$

In particular, for $t = n\epsilon_{X,a}/2$, we have that

$$\mathbb{P}(N_a < n\epsilon_{X,a}/2 \mid X) \leq \exp(-n\epsilon_{X,a}^2/2) \,.$$

If we let $\gamma_{x,a} = \epsilon_{x,a}/2$, we get that

$$\mathbb{P}(N_a < \gamma_{X,a} n \mid X) \leq \exp(-2n\gamma_{X,a}^2) \,.$$

Now, we define $\Omega_{x,a} = \{N_a \geq \gamma_{x,a} n\}$, then we have that

$$\mathbb{P}(\bar{\Omega}_{X,a} \mid X) \leq \exp(-2n\gamma_{X,a}^2) \,. \tag{50}$$

In particular,

$$\mathbb{P}(\bar{\Omega}_{X,\pi_*(X)} \mid X) \leq \exp(-2n\gamma_{X,\pi_*(X)}^2) \,. \tag{51}$$

First, we have that

$$\begin{aligned}
\sqrt{\mathbb{E}\big[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}}\big]} &= \sqrt{\mathbb{E}\big[\mathbb{E}\big[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mid X\big]\big]} \,, \\
&= \sqrt{\mathbb{E}\big[\mathbb{E}\big[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{I}\{\Omega_{X,\pi_*(X)}\} \mid X\big]\big] + \mathbb{E}\big[\mathbb{E}\big[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{I}\{\bar{\Omega}_{X,\pi_*(X)}\} \mid X\big]\big]} \,, \\
&= \sqrt{\mathbb{E}\big[I_1\big] + \mathbb{E}\big[I_2\big]} \,.
\end{aligned} \tag{52}$$

where $I_1 = \mathbb{E}\big[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{I}\{\Omega_{X,\pi_*(X)}\} \mid X\big]$ and $I_2 = \mathbb{E}\big[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{I}\{\bar{\Omega}_{X,\pi_*(X)}\} \mid X\big]\big]$.

To bound I_2, we use the assumption that $\|x\|_2 \leq 1$ for any context x and (49) leads to

$$\|x\|^2_{\hat{\Sigma}_a} \leq \sigma_0^2 + \tau^2 \,, \qquad \forall x, a \,.$$

In particular, we have that

$$\|x\|^2_{\hat{\Sigma}_{\pi_*(x)}} \leq \sigma_0^2 + \tau^2 \,.$$

Let $c_1 = \sigma_0^2 + \tau^2$. Then we have that

$$\begin{aligned} I_2 &= \mathbb{E}\left[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{I}\{\bar{\Omega}_{X,\pi_*(X)}\} \mid X\right], \\ &\leq \mathbb{E}\left[c_1 \mathbb{I}\{\bar{\Omega}_{X,\pi_*(X)}\} \mid X\right], \\ &= c_1 \mathbb{E}\left[\mathbb{I}\{\bar{\Omega}_{X,\pi_*(X)}\} \mid X\right], \\ &= c_1 \mathbb{P}(\bar{\Omega}_{X,\pi_*(X)} \mid X), \\ &\leq c_1 \exp(-2n\gamma^2_{X,\pi_*(X)}) \end{aligned} \tag{53}$$

Lemma E.5 (Oliveira [2016]). *Assume $X_1, \ldots, X_n \in \mathbb{R}^d$ are i.i.d. random variables with bounded fourth moments. Define $G \equiv \mathbb{E}\left[X_1 X_1^\top\right]$ and $\widehat{G}_n \equiv \sum_{i=1}^n X_i X_i^\top$. Let $\mathrm{h} \in (1, +\infty)$ be such that $\sqrt{\mathbb{E}\left[\left(v^T X_1 X_1^\top v\right)^2\right]} \leq \mathrm{h} v^T G v$ for all $v \in \mathbb{R}^d$. Then for any $\delta \in (0,1)$:*

$$\mathbb{P}\left(\widehat{G}_n \succeq \alpha(n,\delta) G\right) \geq 1 - \delta,$$

where $\alpha(n,\delta) = n - 7\mathrm{h}\sqrt{n(d + 2\ln(2/\delta))}$. In particular, for any fixed positive definite matrix Λ, we have that

$$\mathbb{P}\left((\widehat{G}_n + \Lambda)^{-1} \preceq (\alpha(n,\delta) G + \Lambda)^{-1}\right) \geq 1 - \delta, \tag{54}$$

Let us focus on the firm term I_1. Thanks to multiplication with $\mathbb{I}\{\Omega_{X,\pi_*(X)}\}$, we know that given X, $N_{\pi_*(X)} \geq \gamma_{X,\pi_*(X)}\, n$. Therefore, given X, we have that

$$\hat{G}_{\pi_*(X)} = \sigma^{-2} \sum_{i \in [n]} \mathbb{I}_{\{A_i = \pi_*(X)\}} X_i X_i^\top \succeq \sigma^{-2} \sum_{i=1}^{\lfloor \gamma_{X,\pi_*(X)}\, n \rfloor} X_i' X_i'^\top, \qquad \forall a \in \mathcal{A}, \tag{55}$$

where the X_i' are copies of X_i (assuming that $X|A$ and X have the same law). To ease exposition, let $\alpha(\lfloor \gamma_{x,\pi_*(x)} n \rfloor, \delta) = \alpha_x$. Then, applying (54) in Lemma E.5 with $\Lambda = \Sigma_0^{-1} = \sigma_0^{-2} I_d$ and δ leads to

$$\mathbb{P}\left((\hat{G}_{\pi_*(X)} + \sigma_0^{-2} I_d)^{-1} \mathbb{I}\{\Omega_{X,\pi_*(X)}\} \preceq (\sigma^{-2} \alpha_X G + \sigma_0^{-2} I_d)^{-1} \mid X\right) \geq 1 - \delta, \tag{56}$$

where $G = \mathbb{E}_{X \sim \nu}\left[X X^\top\right]$. Noticing that $\tilde{\Sigma}_{\pi_*(X)} = (\hat{G}_{\pi_*(X)} + \sigma_0^{-2} I_d)^{-1}$ leads to

$$\mathbb{P}\left(\tilde{\Sigma}_{\pi_*(X)} \mathbb{I}\{\Omega_{X,\pi_*(X)}\} \leq (\sigma^{-2} \alpha_X G + \sigma_0^{-2} I_d)^{-1} \mid X\right) \geq 1 - \delta. \tag{57}$$

In particular, we bound its maximum eigenvalue as

$$\mathbb{P}\left(\lambda_1(\tilde{\Sigma}_{\pi_*(X)}) \mathbb{I}\{\Omega_{X,\pi_*(X)}\} \leq \frac{1}{\sigma^{-2} g \alpha_X + \sigma_0^{-2}} \mid X\right) \geq 1 - \delta. \tag{58}$$

where $g = \lambda_d(G)$.

Moreover, we know 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 thus

$$\lambda_1(\hat{\Sigma}_a) = \lambda_1(\tilde{\Sigma}_a) + \sigma_0^{-4} \tau^2 \lambda_1(\tilde{\Sigma}_a)^2,$$

where we use Weyl's inequality and the fact that for any matrix A and any positive semi-definite matrix B such that the product $\mathrm{A}^\top \mathrm{B} \mathrm{A}$ exists, the following inequality holds $\lambda_1(\mathrm{A}^\top \mathrm{B} \mathrm{A}) \leq \lambda_1(\mathrm{B}) \lambda_1(\mathrm{A}^\top \mathrm{A})$. Therefore, from (58), we have that

$$\mathbb{P}\left(\lambda_1(\hat{\Sigma}_{\pi_*(X)}) \mathbb{I}\{\Omega_{X,\pi_*(X)}\} \leq \frac{1}{\sigma^{-2} g \alpha_X + \sigma_0^{-2}} + \frac{\sigma_0^{-4} \tau^2}{\left(\sigma^{-2} g \alpha_X + \sigma_0^{-2}\right)^2} \mid X\right) \geq 1 - \delta. \tag{59}$$

Using the law of total expectation, we get that

$$I_1 = \mathbb{E}\big[\|X\|^2_{\hat{\Sigma}_{\pi_*(X)}} \mathbb{I}\{\Omega_{X,\pi_*(X)}\} \mid X\big]\,,$$
$$\leq \frac{1}{\sigma^{-2} g \alpha_X + \sigma_0^{-2}} + \frac{\sigma_0^{-4}\tau^2}{\big(\sigma^{-2} g \alpha_X + \sigma_0^{-2}\big)^2} + c_1 \delta\,.$$

Setting $\delta = 2/n$ leads to

$$I_1 \leq \frac{1}{\sigma^{-2} g \alpha_X + \sigma_0^{-2}} + \frac{\sigma_0^{-4}\tau^2}{\big(\sigma^{-2} g \alpha_X + \sigma_0^{-2}\big)^2} + \frac{2c_1}{n}\,. \tag{60}$$

Combining (53) and (60) yields

$$\mathbb{E}\big[\|X\|_{\hat{\Sigma}_{\pi_*(X)}}\big] \leq \sqrt{\mathbb{E}\big[I_1\big] + \mathbb{E}\big[I_2\big]}\,, \tag{61}$$
$$\leq \sqrt{\mathbb{E}_{X\sim\nu}\left[\frac{1}{\sigma^{-2} g \alpha_X + \sigma_0^{-2}} + \frac{\sigma_0^{-4}\tau^2}{\big(\sigma^{-2} g \alpha_X + \sigma_0^{-2}\big)^2}\right] + \frac{2c_1}{n} + c_1 \mathbb{E}_{X\sim\nu}\left[\exp(-2n\gamma^2_{X,\pi_*(X)})\right]}\,,$$

Replacing $c_1 = \sigma_0^2 + \tau^2$ leads to

$$\mathbb{E}\big[\|X\|_{\hat{\Sigma}_{\pi_*(X)}}\big] \tag{62}$$
$$\leq \sqrt{\mathbb{E}_{X\sim\nu}\left[\frac{1}{\sigma^{-2} g \alpha_X + \sigma_0^{-2}} + \frac{\sigma_0^{-4}\tau^2}{\big(\sigma^{-2} g \alpha_X + \sigma_0^{-2}\big)^2}\right] + \frac{2(\sigma_0^2+\tau^2)}{n} + (\sigma_0^2+\tau^2) \mathbb{E}_{X\sim\nu}\left[\exp(-2n\gamma^2_{X,\pi_*(X)})\right]}\,,$$

Finally, noticing that $\gamma_{X,\pi_*(X)} = \frac{\pi_0(\pi_*(X)|X)}{2}$ concludes the proof. □

F ADDITIONAL EXPERIMENTS

As mentioned in Appendix F, 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.

F.1 Implementation Details of Baselines

We implement the baselines as follows.

- **IPS.** For OPE, we use

$$\hat{V}(\pi, S) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}} r_i\,, \tag{63}$$

  and for OPL, we use

$$\operatorname*{argmax}_{\pi} \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}} r_i\,, \tag{64}$$

  where $\tau \in [0, 1]$ is a hyper-parameter.

- **snIPS.** For OPE, we use

$$\hat{V}(\pi, S) = \frac{1}{\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}} \sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} r_i\,, \tag{65}$$

  and for OPL, we use

$$\operatorname*{argmax}_{\pi} \frac{1}{\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)}} \sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\pi_0(a_i|x_i)} r_i\,, \tag{66}$$

- **MIPS.** We cluster actions into L groups and let $\phi(a)$ be the cluster of action a. Let c_i be the cluster of action a_i, then For OPE, we use

$$\hat{V}(\pi, S) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(c_i|x_i)}{\pi_0(c_i|x_i)} r_i \,, \tag{67}$$

where $c_i = \phi(a_i)$ for any $i \in [n]$, and $\pi(c|x) = \sum_{a\in\mathcal{A}} \mathbb{1}\left[\phi(a) = c\right] \pi(a|x)$. For OPL, we use

$$\operatorname*{argmax}_{\pi} \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(c_i|x_i)}{\pi_0(c_i|x_i)} r_i \,, \tag{68}$$

- **PC.** We use the Knn implementation of `PC`. Let $\psi(a, k)$ be the set of k-nearest neighbors of a, then For OPE, we use

$$\hat{V}(\pi, S) = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{a\in\mathcal{A}} \mathbb{1}\left[a \in \psi(a_i, k)\right] \pi(a|x)}{\sum_{a\in\mathcal{A}} \mathbb{1}\left[a \in \psi(a_i, k)\right] \pi_0(a|x)} r_i \,, \tag{69}$$

$$\operatorname*{argmax}_{\pi} \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{a\in\mathcal{A}} \mathbb{1}\left[a \in \psi(a_i, k)\right] \pi(a|x)}{\sum_{a\in\mathcal{A}} \mathbb{1}\left[a \in \psi(a_i, k)\right] \pi_0(a|x)} r_i \,. \tag{70}$$

- **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

$$\hat{r}(x, a) = \phi(X)^\top \hat{\mu}_a \,, \tag{71}$$

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

$$\begin{aligned} \theta_a &\sim \mathcal{N}(\mu_a, \Sigma_a) \,, && \forall a \in \mathcal{A} \,, \\ R \mid \theta, X, A &\sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2) \,, \end{aligned} \tag{72}$$

Under this prior, each action a has an associated parameter θ_a. Given the prior in (2), the posterior distribution of an action parameter follows a multivariate Gaussian: $\theta_a \mid S \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$, 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

$$\hat{r}(x, a) = \phi(X)^\top \hat{\mu}_a \,, \tag{73}$$

- **DR.** For OPE, we use

$$\hat{V}(\pi, S) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}} \left(r_i - \hat{r}(a_i, x_i)\right) + \mathbb{E}_{a\sim\pi(\cdot|x_i)}\left[\hat{r}(x_i, a)\right] \,, \tag{74}$$

with $\tau \in [0, 1]$ and $\hat{r}$ is the reward model obtained using `DM (Freq)`. For OPL, we use

$$\operatorname*{argmax}_{\pi} \frac{1}{n}\sum_{i=1}^{n} \frac{\pi(a_i|x_i)}{\max\{\pi_0(a_i|x_i), \tau\}} \left(r_i - \hat{r}(a_i, x_i)\right) + \mathbb{E}_{a\sim\pi(\cdot|x_i)}\left[\hat{r}(x_i, a)\right] \,, \tag{75}$$

F.2 Additional Results

Additional Synthetic Experiments. We provide additional OPE and OPL results. We consider the same synthetic setting as in Section 6 and we vary $K \in \{100, 1000\}$ and $d' \in \{5, 10, 20\}$.

For OPE, we fix a target policy π and assess algorithm performance in evaluating the value function of that target policy using the mean squared error (MSE) averaged over 50 problem instances sampled from the prior. This can be seen as a proxy for our theoretical metric, Bayesian MSE (BMSE). The target policy is defined as ϵ-greedy with $\epsilon = 0.5$; it chooses the best action with probability 0.5 and a random action with probability 0.5.

Results are shown in Fig. 6. While `sDM` maintains its advantage over baselines, particularly in low-data settings, an interesting observation emerged. IPS-based methods (`snIPS`, `DR` and `MIPS`) outperformed standard direct methods (`DM (Bayes)` and `DM (Freq)`) in the OPE experiments, which was not the case in OPL experiments. The only direct method that outperformed them in OPE is `sDM`. Also, `MIPS` and `PC` outperform other IPS-variants when the number of action is 1000, but fail to outperform standard methods when $K = 100$.

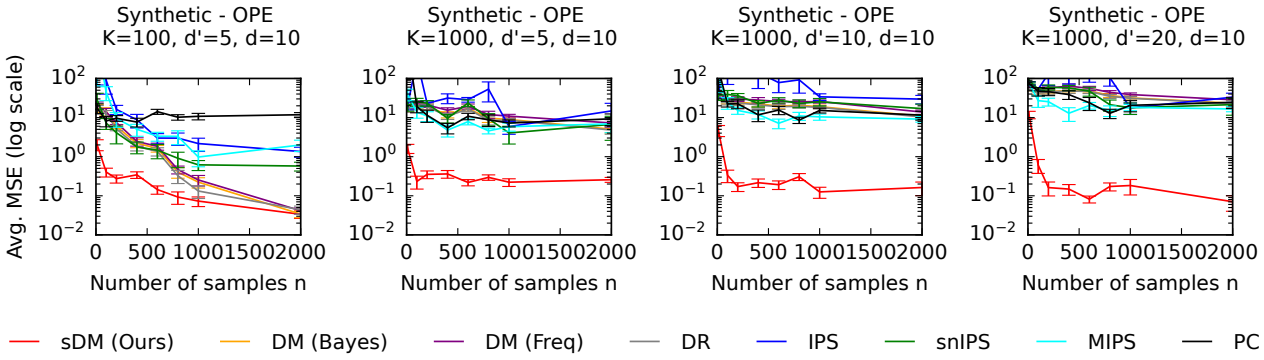

Figure 6: The average MSE of an ϵ-greedy target policy on **synthetic problems** with varying n, K and d'.

For OPL, the gap in performance between `sDM` and all other baselines widens when K increases. It also widens when the problem becomes higher-dimensional (e.g., increased value of d').

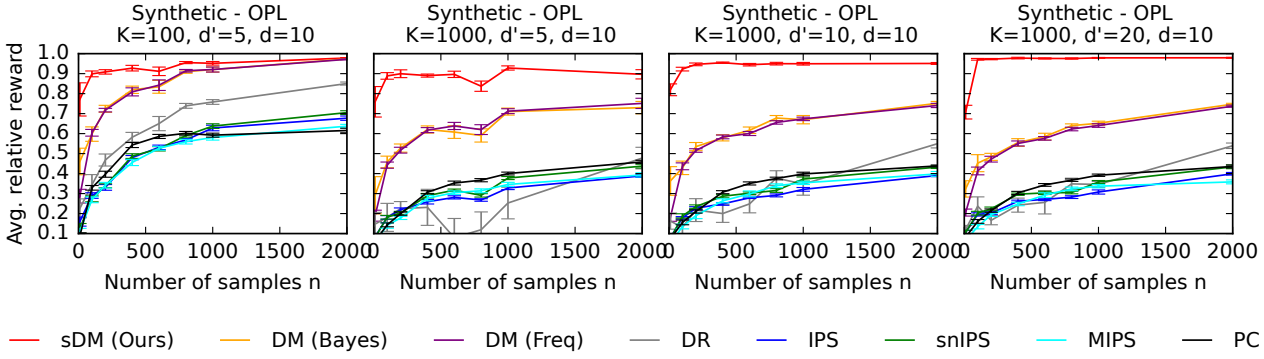

Figure 7: The average relative reward of the learned policy using one of the baselines on **synthetic problems** with varying n, K and d'.

Additional MovieLens Experiments. Similar conclusions are drawn as in the main text.

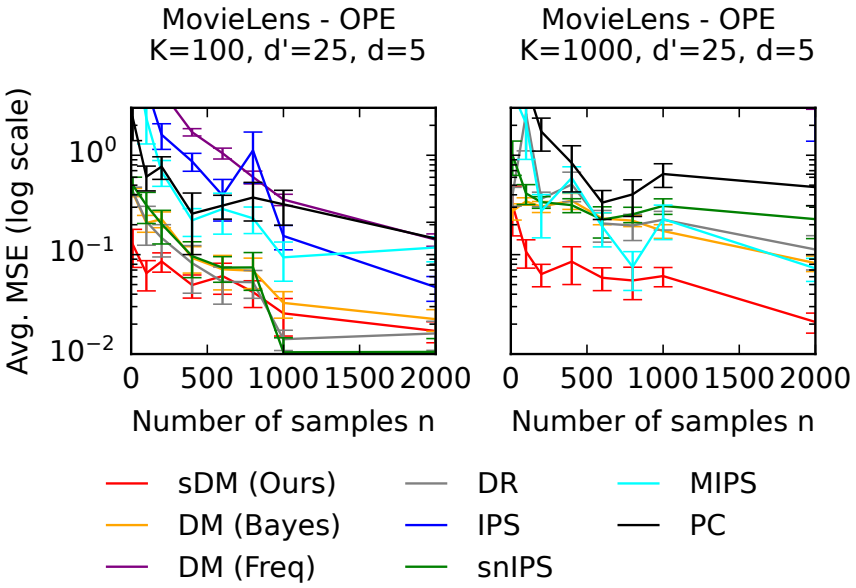

Figure 8: The average MSE of an ϵ-greedy target policy on **MovieLens problems** with varying n, K and d'.

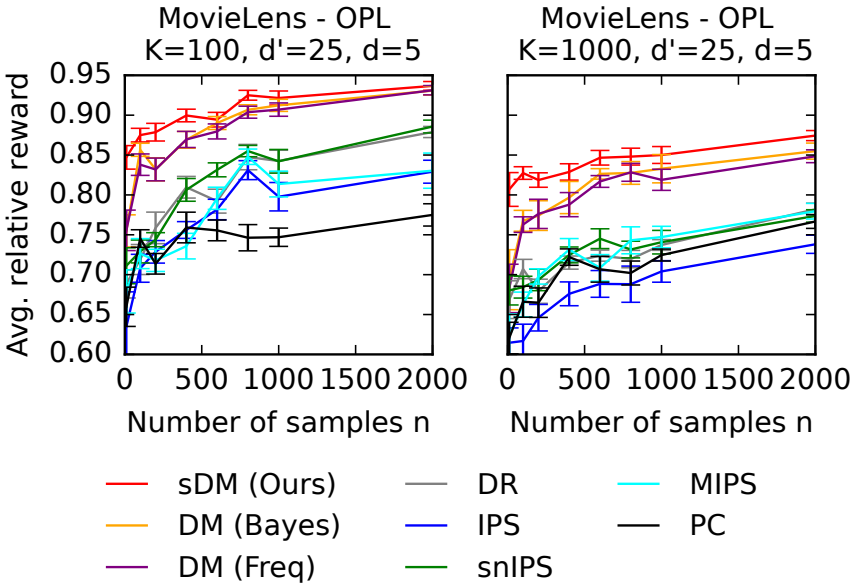

Figure 9: The average relative reward of the learned policy using one of the baselines on **MovieLens problems** with varying n, K and d'.

F.3 Different Logging Policy: ϵ-Greedy

Here we consider the setting in Section 6 and provide additional results with the ϵ-greedy logging policy, where $\epsilon = 0.5$. We consider both synthetic and MovieLens datasets. The results are shown in Fig. 10. The conclusions are similar to those in Section 6, except that `IPS` outperforms the other variants `DR` and `snIPS` when the logging policy is performing well. Also, the results for the MovieLens problems are given in Fig. 11 and the conclusions are similar to those made in Section 6.

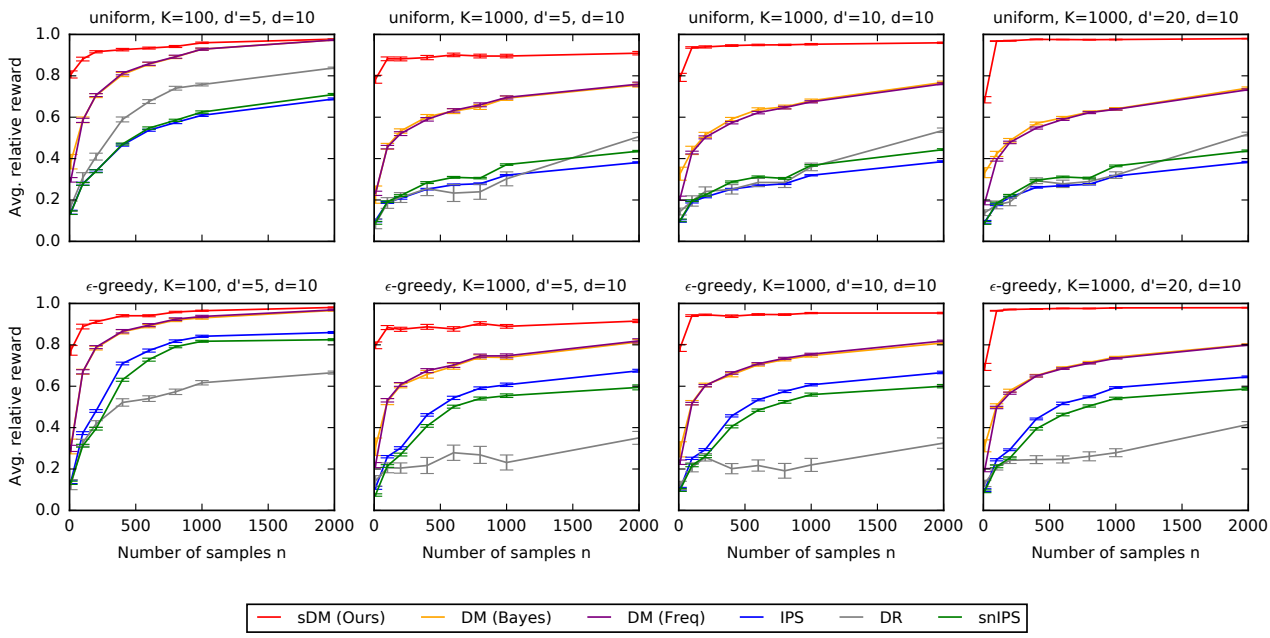

Figure 10: The relative reward of the learned policy on **synthetic problems** with varying n, K and d' and varying logging policies.

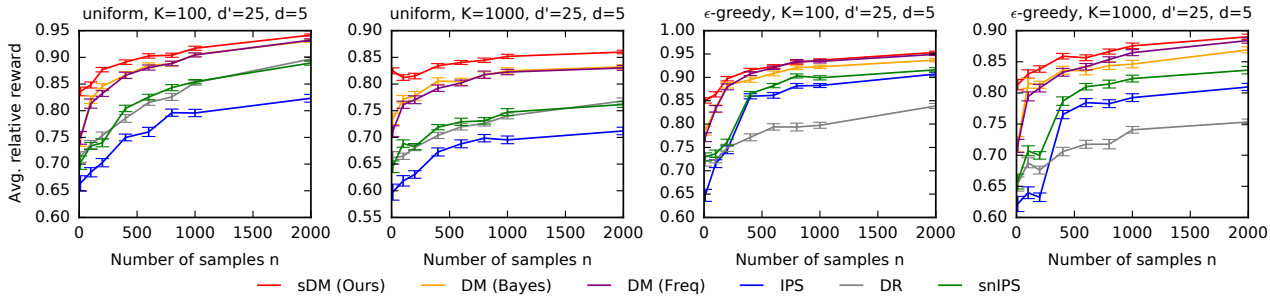

Figure 11: The relative reward of the learned policy on **MovieLens problems** with varying n and K and varying logging policies.

F.4 Robustness to Likelihood Misspecification

We strengthened our evaluation by assessing `sDM`'s robustness to likelihood misspecification below (robustness to prior misspecification is provided in Appendix F.5). In these experiments, the true data-generating process (same as the synthetic experiments in Section 6) differed from `sDM`'s assumptions in two different ways: either the likelihood is misspecified

Misspecified likelihood (Figs. 12 and 13). 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, we make the same observations as those made in the main text. Likelihood misspecification affect `sDM`'s performance on OPE, but `sDM` still outperforms all methods by a large margin in OPL.

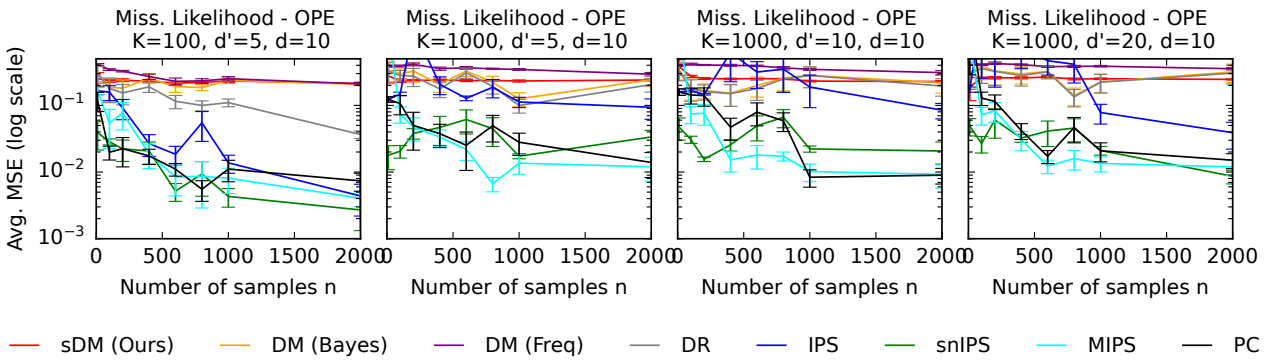

Figure 12: Effect of prior mean and covariance misspecification: The average MSE of an ϵ-greedy target policy on synthetic problems using both misspecified prior means and covariances with varying n and K.

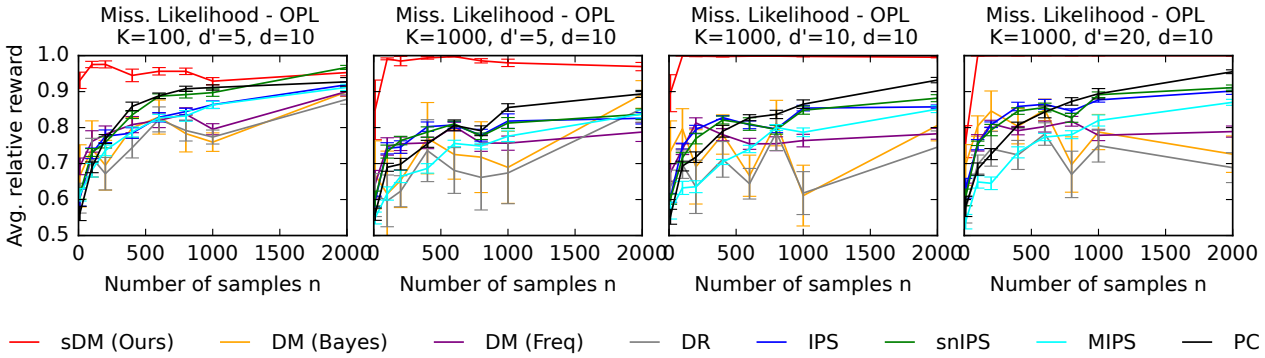

Figure 13: Effect of likelihood misspecification: The relative reward of the learned policy on synthetic problems using misspecified likelihood with varying n and K.

F.5 Robustness to Prior Misspecification

Misspecified prior means and covariances(Fig. 14). 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.

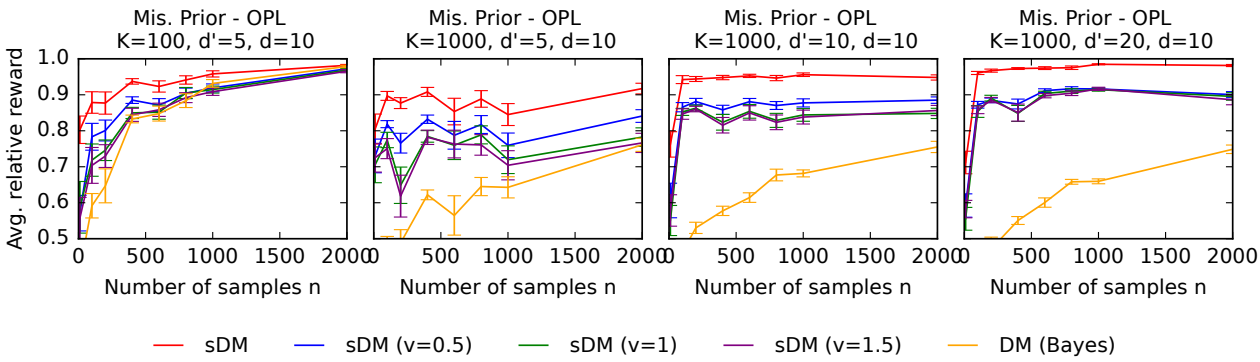

Figure 14: Effect of prior mean and covariance misspecification: The average relative reward of the learned policy on synthetic problems using both misspecified prior means and covariances with varying n and K and d'.

F.6 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 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

$$\hat{\pi}_{\mathrm{G}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b \in \mathcal{A}} \hat{r}(x, b)\}\,, \tag{76}$$

while a pessimistic one writes

$$\hat{\pi}_{\mathrm{P}}(a \mid x) = \mathbb{1}\{a = \operatorname*{argmax}_{b \in \mathcal{A}} \hat{r}(x, b) - u(x, a)\}\,, \tag{77}$$

where $u(x,a) = \alpha(d,\delta)\|\phi(X)\|_{\hat{\Sigma}_a}$ with $\alpha(d,\delta) = \sqrt{d + 2\sqrt{d\log\frac{1}{\delta}} + 2\log\frac{1}{\delta}}$ and this is derived in (28) in Appendix E.2. As predicted by our theory, the results show that the greedy policy has better average performance over multiple bandit instances sampled from the prior.

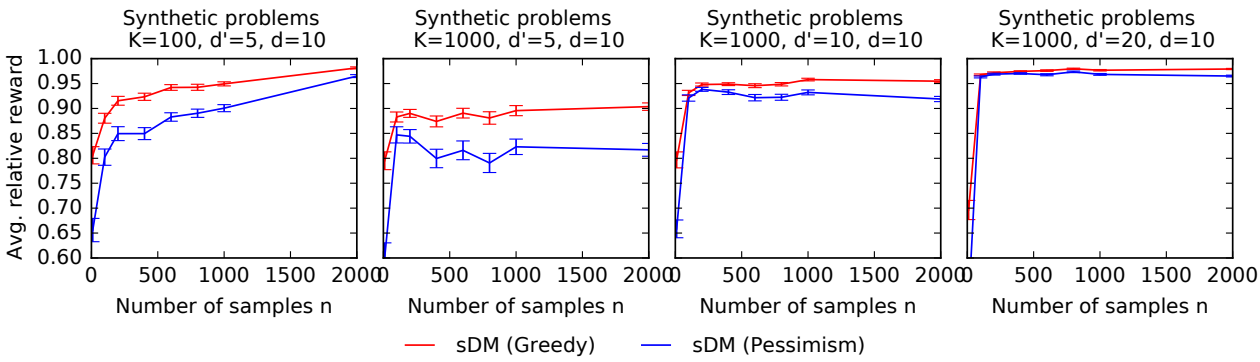

Figure 15: Comparison of `sDM` with greedy policy and `sDM` with pessimistic policy in OPL: The average MSE of an ϵ-greedy policy on synthetic problems with varying n, K, and d'.

F.7 KuaiRec

Setting. We use the KuaiRec dataset [Gao et al., 2022] to evaluate our method. This dataset contains two matrices:

1. `big_matrix`: A sparse matrix containing user-item interactions (watch ratios for various videos).
2. `small_matrix`: A smaller, fully observed dataset with complete user-item interactions. This matrix is only used for data collection and algorithm evaluation.

The `big_matrix` is used to obtain low-rank factorization embeddings through Singular Value Decomposition (SVD), representing users (contexts) and videos (actions). The `small_matrix` is used to collect the sample set and to evaluate the algorithms.

Before applying SVD, we preprocess `big_matrix` as follows. First, we clipped the watch ratio scores to a maximum value of 10 to mitigate the impact of outliers and extreme values. Second, we normalized the interaction scores per user by dividing each user's interaction values by their maximum watch ratio. This scaling ensured that all user interaction scores ranged between 0 and 1, providing a uniform scale across different users.

After preprocessing, SVD was applied to the normalized matrix, resulting in video embeddings and user embeddings. The user embeddings are provided as input to the algorithms (serving as contexts), while the video embeddings are used as prior information by Bayesian algorithms that incorporate such priors. The dataset also includes category information for each video, which is beneficial for methods that rely on item categories.

Results. Since the rewards are collected from the `small_matrix`, the true reward distribution is unknown, and there is no explicit prior from which videos are sampled. This characteristic makes the dataset a challenging benchmark for Bayesian algorithms like `sDM`, as it does not conform to the typical assumptions made by these methods. Despite this challenge, the strong performance of `sDM` in Fig. 16 is particularly noteworthy. Specifically, `sDM` and `MIPS` demonstrate comparable and superior performance to other baselines in OPL, while `sDM` outperforms the baselines in OPE. It is important to note that results for OPE are hard to read since the performance of IPS-based methods fluctuated significantly when the number of samples n changes in this scenario.

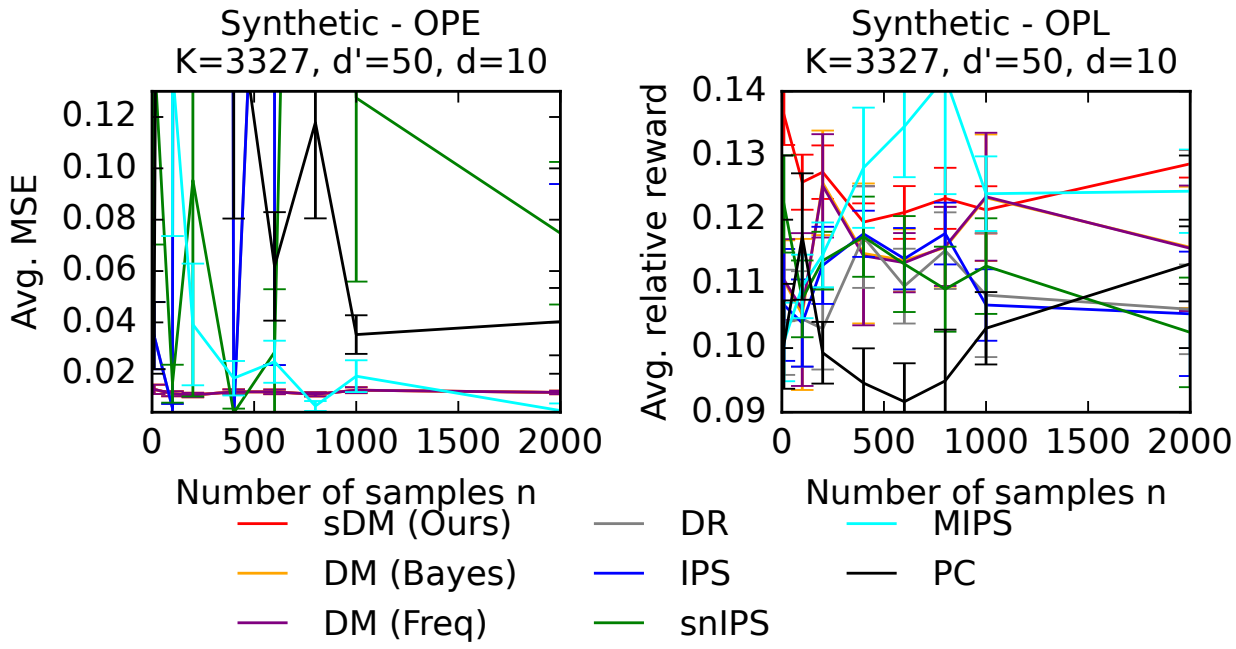

Figure 16: Comparison of `sDM` with baselines on the KuaiRec dataset for both OPE and OPL tasks.

G BROADER IMPACT

This work contributes to the development and analysis of practical algorithms for offline learning to act under uncertainty. Our generic setting and algorithms have broad potential use, practitioners will therefore need to specifically address possible social impacts with respect to the relevant application.

H AMOUNT OF COMPUTATION REQUIRED

Our experiments were conducted on internal machines with 30 CPUs and 1 GPU. Therefore, they required a moderate amount of computation.