Bayesian Off-Policy Evaluation and Learning for Large Action Spaces

Imad Aouali [1,2] Victor-Emmanuel Brunel [2] David Rohde [1] Anna Korba [2]

[1]Criteo AI Lab [2]CREST–ENSAE

Table of contents

Interactive Systems

Learning in Interactive Systems

Framework (Offline Contextual Bandit [8, 9, 12, 13]).

Contexts x	Actions a	Logging policy π_0
User / environment features	Items / ads / decisions	Deployed system

Logged Data. $\mathcal{D} = \{(x_i, a_i, r_i)\}_{i=1}^{n}$ with $a_i \sim \pi_0(\cdot \mid x_i)$.
Objective. Evaluate/learn a new policy π that maximizes

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

IPS vs DM in Large Action Spaces

Inverse Propensity Scoring (IPS) [5, 7, 9, 10, 14, 15]

$$\hat{V}_{\mathsf{IPS}}(\pi, S) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i \mid x_i)}{\pi_0(a_i \mid x_i)} \, r_i$$

Pros: unbiased if π_0 has full support.

Cons: high variance, biased if π_0 has deficient support.

IPS vs DM in Large Action Spaces

Inverse Propensity Scoring (IPS) [5, 7, 9, 10, 14, 15]

$$\hat{V}_{\mathsf{IPS}}(\pi, S) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(a_i \mid x_i)}{\pi_0(a_i \mid x_i)} \, r_i$$

Pros: unbiased if π_0 has full support.

Cons: high variance, biased if π_0 has deficient support.

Direct Method (DM) [4, 11]

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

Pros: low variance; does not require π_0, practical [3].

Cons: modeling bias if $\hat{r}$ is misspecified.

Structured DM (`sDM`)

Motivation: Why Structure?

Pitfall of non-structured priors. Standard Bayesian DM:

$$\theta_a \sim \mathcal{N}(\mu_a, \Sigma_a)\,,$$
$$R \mid X, A, \theta \sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2)$$

Issue: Posterior of θ_a only uses samples with $A = a$. Unseen actions revert to the prior $\Rightarrow$ inefficient when K is large.

Pitfall of non-structured priors. Standard Bayesian DM:

$$\theta_a \sim \mathcal{N}(\mu_a, \Sigma_a)\,,$$
$$R \mid X, A, \theta \sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2)$$

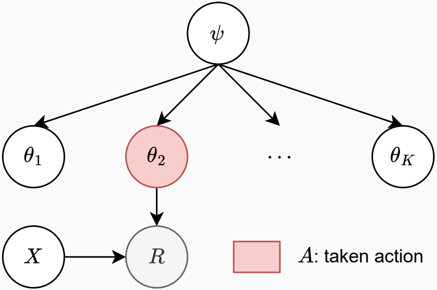
Issue: Posterior of θ_a only uses samples with $A = a$. Unseen actions revert to the prior $\Rightarrow$ inefficient when K is large.

Key idea of `sDM`. Share information across actions via latent ψ:

$$\psi \sim q\,,$$
$$\theta_a \mid \psi \sim p_a(\cdot; f_a(\psi)),$$
$$R \mid X, A, \theta \sim p(\cdot \mid X; \theta_A)$$

Effect: Observing one action updates beliefs about others.

Graphical View



- Conditional independence: $\{\theta_a\}_a$ independent given ψ.
- Structure encoded by f_a (e.g., linear mixing via W_a).
- Scales without expensive $Kd \times Kd$ posteriors.

Linear–Gaussian Instance

Model.

$$\psi \sim \mathcal{N}(\mu, \Sigma)\,,$$
$$\theta_a \mid \psi \sim \mathcal{N}(W_a \psi, \Sigma_a)\,,$$
$$R \mid X, A, \theta \sim \mathcal{N}(\phi(X)^\top \theta_A, \sigma^2)\,.$$

Closed-form posteriors.

$$\theta_a \mid \psi, S \sim \mathcal{N}(\tilde{\mu}_a, \tilde{\Sigma}_a)$$
$$\psi \mid S \sim \mathcal{N}(\bar{\mu}, \bar{\Sigma})$$

Action posterior (marginalizing ψ): $\theta_a \mid S \sim \mathcal{N}(\hat{\mu}_a, \hat{\Sigma}_a)$ with

$$\hat{\mu}_a = \tilde{\Sigma}_a(\Sigma_a^{-1} W_a \bar{\mu} + B_a), \quad \hat{\Sigma}_a = \tilde{\Sigma}_a + \tilde{\Sigma}_a \Sigma_a^{-1} W_a \bar{\Sigma} W_a^\top \Sigma_a^{-1} \tilde{\Sigma}_a$$

Plug-in reward: $\hat{r}(x, a) = \phi(x)^\top \hat{\mu}_a$.

Applications of the Structure

- **Mixed-effects:** $W_a = w_a^\top \otimes I_d$, $\psi = (\psi_j)_{j \leq J}$; sparsity via $w_{a,j} = 0$.
- **Low-rank:** $d' \ll d$, W_a low-rank $\Rightarrow$ shared latent factors across actions.
- **Practical:** Movies/items clustered; W_a encodes theme mixture.

OPE/OPL with `sDM`

OPE (DM plug-in).

$$\hat{V}_{\mathsf{DM}}(\pi, S) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi(a \mid X_i)\,\hat{r}(X_i, a), \quad \hat{r}(x, a) = \mathbb{E}\left[r(x, a; \theta) \mid S\right]$$

OPE (DM plug-in).

$$\hat{V}_{\mathsf{DM}}(\pi, S) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \pi(a \mid X_i)\, \hat{r}(X_i, a), \quad \hat{r}(x, a) = \mathbb{E}\left[r(x, a; \theta) \mid S\right]$$

OPL (Greedy on $\hat{r}$).

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

Greedy beats pessimism under our Bayesian metric (next).

Main Results (Informal)

Thm (Covariance-dependent bound).

$$\mathsf{Bso}(\hat{\pi}_{\mathsf{G}}) \;\lesssim\; \mathbb{E}\big[\, \|\phi(X)\|_{\hat{\Sigma}_{\pi_*(X)}} \,\big]\,,$$

where Bso is the *suboptimality* on average, with expectation taken over S and $\theta_* \sim$ prior.

- ***sDM***'s Bayes suboptimality is smaller when posterior uncertainty of the optimal action along $\phi(X)$ is small.

Main Results (Informal)

Thm (Scaling in n).
$\textsc{Bso}(\hat{\pi}_{\mathsf{G}}) = \mathcal{O}(\,1/\sqrt{n}\,)$ with constants that depend explicitly on $\pi_0(\pi_*(X) \mid X)$.

- Avoids "well-explored dataset" assumptions; and only depends on π_0's exploration of the optimal action π_*.

Experiments

Synthetic and MovieLens

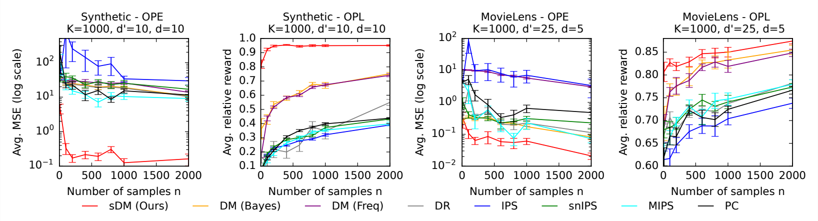


Figure 1: OPE/OPL performance: *sDM* vs. DM baselines and IPS-variants (MIPS, PC).

Scaling with K

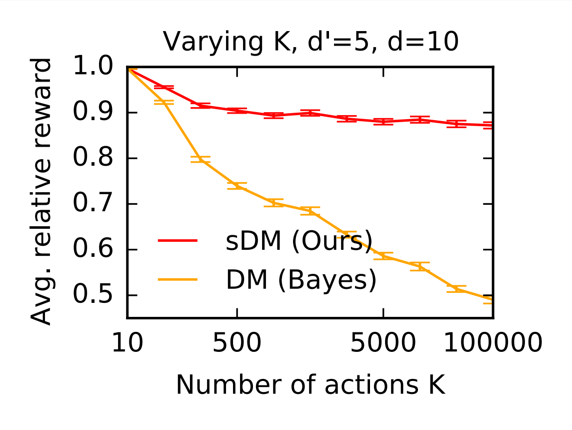


Figure 2: *sDM* vs. standard Bayesian DM as number of actions K increases.

Conclusion

Conclusion

- *sDM*: Bayesian DM with structured priors to share information across actions.
- Closed-form linear–Gaussian instance; scalable to large K.
- New Bayesian metric (BSO); greedy preferred to pessimism under BSO.
- Strong empirical results; robust to moderate misspecification.

Limitations: prior misspecification theory; non-linear hierarchies, neural networks.

Extentions: We extended these ideas to online bandits [1, 2, 6], large-scale rec sys [3, 4].

References

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

[2] Imad Aouali. Diffusion models meet contextual bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

References

[3] Imad Aouali, Amine Benhalloum, Martin Bompaire, Achraf Ait Sidi Hammou, Sergey Ivanov, Benjamin Heymann, David Rohde, Otmane Sakhi, Flavian Vasile, and Maxime Vono. Reward optimizing recommendation using deep learning and fast maximum inner product search. In *proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 4772–4773, 2022.

[4] 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.

References

[5] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Exponential Smoothing for Off-Policy Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 984–1017. PMLR, 2023.

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

References

[7] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Unified pac-bayesian study of pessimism for offline policy learning with regularized importance sampling. *UAI 2024*, 2024.

[8] 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.

References

[9] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1097–1104, 2011.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.