Diffusion Models Meet Contextual Bandits

Imad Aouali

Criteo AI Lab
CREST, ENSAE, IP Paris

Table of contents

Motivation

Why diffusion priors for bandits?

Problem

- Large-K contextual bandits: independent posteriors (LinUCB [8, 10]/LinTS [1, 2, 12]) become statistically inefficient; joint posteriors are computationally intractable.
- Many real systems exhibit *correlated actions* [5]: learning one action informs many others.

Why diffusion priors for bandits?

Problem

- Large-K contextual bandits: independent posteriors (LinUCB [8, 10]/LinTS [1, 2, 12]) become statistically inefficient; joint posteriors are computationally intractable.
- Many real systems exhibit *correlated actions* [5]: learning one action informs many others.

Key idea

- Use a **pre-trained diffusion model** as an expressive *Bayesian prior* over action parameters.
- Design **Diffusion Thompson sampling (dTS)** with efficient posterior updates and sampling.

Algorithm

Contextual bandit with diffusion prior

At $t \in [n]$, observe X_t, choose $A_t \in [K]$, get $Y_t \sim P(\cdot \mid X_t; \theta_{A_t})$.

Per-action (disjoint) parameters: $\theta_a \in \mathbb{R}^d$, GLM reward with mean $g(x^\top \theta_a)$.

Diffusion-derived prior (hierarchical):

$$\begin{aligned} \psi_L &\sim \mathcal{N}(0, \Sigma_{L+1}), \\ \psi_{\ell-1} \mid \psi_\ell &\sim \mathcal{N}\big(f_\ell(\psi_\ell), \Sigma_\ell\big), \quad \ell \in [L] \setminus \{1\}, \\ \theta_a \mid \psi_1 &\sim \mathcal{N}\big(f_1(\psi_1), \Sigma_1\big), \quad a \in [K]\,, \\ Y_t \mid X_t, A_t, \theta &\sim P\big(\cdot \mid X_t, \theta_{A_t}\big). \end{aligned}$$

Shared-parameter variant (Other possible setting)

If $r(x, a; \theta) = g(\varphi(x, a)^\top \theta)$ with shared $\theta \in \mathbb{R}^d$:

$$
\begin{aligned}
\psi_L &\sim \mathcal{N}(0, \Sigma_{L+1}), \\
\psi_{\ell-1} \mid \psi_\ell &\sim \mathcal{N}\big(f_\ell(\psi_\ell), \Sigma_\ell\big), \\
\theta \mid \psi_1 &\sim \mathcal{N}\big(f_1(\psi_1), \Sigma_1\big), \\
Y_t \mid X_t, A_t, \theta &\sim P\big(\cdot \mid \varphi(X_t, A_t)^\top \theta\big).
\end{aligned}
$$

All posterior formulas adapt verbatim; K-independent regret becomes attainable if φ is known.

Hierarchical sampling via recursion

Posterior factorization:

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

dTS (one round):

1. Sample $\psi_{t,L} \sim p(\psi_L \mid H_t)$,
2. Descend to $\psi_{t,1}$ via $\psi_{t,1} \sim p(\psi_{\ell-1} \mid \psi_{t,\ell}, H_t)$.
3. For each $a \in [K]$, sample $\theta_{t,a} \sim p(\theta_a \mid \psi_{t,1}, H_{t,a})$ (conditionally independent).
4. Play $A_t = \arg\max_a r(X_t, a; \theta_t)$; observe Y_t and update.

Implementing the posteriors

Action posterior (given ψ_1):

$$p(\theta_a \mid \psi_1, H_{t,a}) \propto \Big[\prod_{i \in S_{t,a}} P(Y_i \mid X_i; \theta_a)\Big] \mathcal{N}\big(\theta_a; f_1(\psi_1), \Sigma_1\big).$$

Latent posteriors:

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

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

Recursions for $p(H_t \mid \psi_\ell)$:

Base: $$p(H_t \mid \psi_1) = \prod_{a=1}^{K} \int \Big[\prod_{i \in S_{t,a}} P(Y_i \mid X_i; \theta_a)\Big] \mathcal{N}(\theta_a; f_1(\psi_1), \Sigma_1)\, d\theta_a,$$

Step: $$p(H_t \mid \psi_\ell) = \int p(H_t \mid \psi_{\ell-1})\, \mathcal{N}(\psi_{\ell-1}; f_\ell(\psi_\ell), \Sigma_\ell)\, d\psi_{\ell-1}.$$

Two approximations

- **(i) Likelihood approx.** Likelihood approximated by Gaussian with MLE $\hat{B}_{t,a}$ as its mean and Hessian $\hat{G}_{t,a}^{-1}$ as its covariance [9].

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

- **(ii) Diffusion approx.** start from exact *linear* diffusion solutions [3] and replace linear maps by $f_\ell(\cdot)$ to obtain closed-form Gaussian conditionals with data-dependent means/covariances.

Overall (important). The resulting global posterior is *not* Gaussian. Our construction preserves the diffusion hierarchy but replaces each layer's conditional with a Gaussian whose mean and covariance are *updated and data-dependent*.

Approximate action posterior

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

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

Precision-additivity; mean is precision-weighted average of prior mean $f_1(\psi_1)$ and MLE $\hat{B}_{t,a}$.

Approximate latent posteriors

For $\ell \in [L+1] \setminus \{1\}$:

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

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

with base/step recursions

$$\bar{G}_{t,1} = \sum_{a=1}^{K} \big(\Sigma_1^{-1} - \Sigma_1^{-1}\hat{\Sigma}_{t,a}\Sigma_1^{-1}\big), \quad \bar{B}_{t,1} = \Sigma_1^{-1} \sum_{a=1}^{K} \hat{\Sigma}_{t,a}\hat{G}_{t,a}\hat{B}_{t,a},$$
$$\bar{G}_{t,\ell} = \Sigma_\ell^{-1} - \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\Sigma_\ell^{-1}, \quad \bar{B}_{t,\ell} = \Sigma_\ell^{-1}\bar{\Sigma}_{t,\ell-1}\bar{B}_{t,\ell-1}.$$

Theory

Linear-Gaussian intuition

Assume linear links $f_\ell(\psi_\ell) = W_\ell \psi_\ell$ and Gaussian rewards. Approximation becomes exact.

Informal Bayes regret:

$$\tilde{\mathcal{O}}\left(\sqrt{n\left(dK\sigma_1^2 + d\sum_{\ell=1}^{L}\sigma_{\ell+1}^2\,\sigma_{\max}^{2\ell}\right)}\right), \quad \sigma_{\max}^2 = \max_{\ell\in[L+1]}\left(1+\frac{\sigma_\ell^2}{\sigma^2}\right).$$

Refinement with sparsity: if W_ℓ has $d_\ell \ll d$ active columns, replace d by d_ℓ in latent terms.

Takeaways: informative (possibly sparse) priors reduce regret; dependence on K enters only via σ_1^2.

Complexity

Complexity and statistical benefits

- Maintaining a full $dK \times dK$ joint posterior:
  - $\mathcal{O}(K^3 d^3)$ time
  - $\mathcal{O}(K^2 d^2)$ space
- `dTS` stores only $L + K$ many $d \times d$ covariances:
  - $\mathcal{O}((L+K)d^3)$ time
  - $\mathcal{O}((L+K)d^2)$ space
- Compared to `LinTS`: similar cost but *uses correlations*, lowering regret especially for large K.

Experiments

Synthetic: true diffusion prior

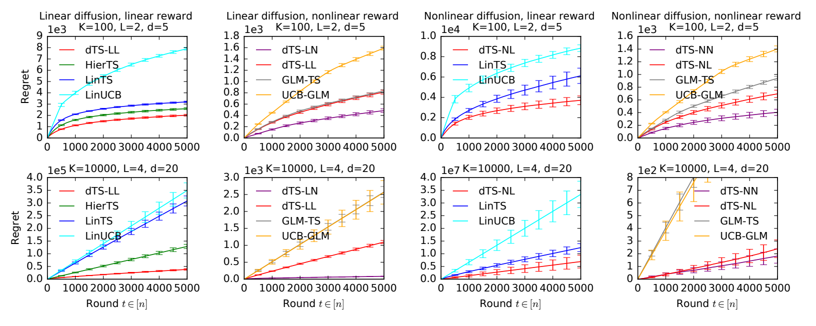
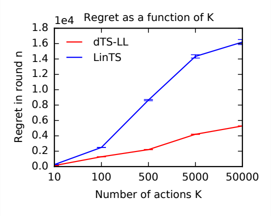
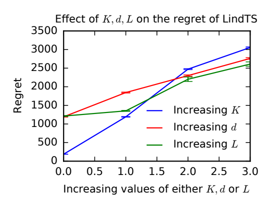


Figure 1: Regret of *dTS* across linear/nonlinear diffusion and rewards; varying d, K, L.

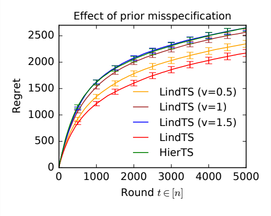
Scaling and misspecification



Gap vs K

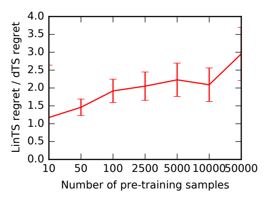


Scaling with K, d, L

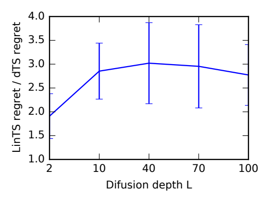


Prior misspecification

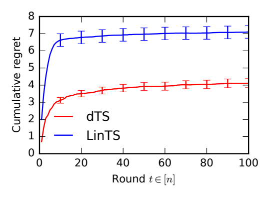
True prior is not diffusion



Pretrain sample size



Diffusion depth L



MovieLens regret

Discussion

Limitations and scope

Limits

- Theory: formal guarantees only shown for linear-Gaussian case. PAC-Bayes theory could be used (e.g., proved successful in offline contextual bandits [4, 7]).
- Approximations: (i) GLM likelihood Gaussianization; (ii) diffusion linearization; did not quantify the error.

Extensions

- *Best-Arm Identification (BAI)*. The diffusion prior gives a sample-efficient structure for BAI; plug our posterior sampler into Bayesian fixed-budget BAI procedures [11].
- *Off-Policy Learning (OPE/OPL)*. The same hierarchy can regularize large-action OPE/OPL objectives [6].

Conclusion

Conclusion

- **dTS** : efficient Thompson sampling with diffusion priors.
- Tractable approximate posteriors and Bayes regret insight.
- Strong empirical performance across regimes.

When to use dTS

- Large-scale settings where offline data exists to pretrain the diffusion prior.
- If actions are unstructured or data are extremely scarce, *LinTS* or *HierTS* can suffice.

Code:
github.com/imadaouali/diffusion-thompson-sampling

References

[1] Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceeding of the 25th Annual Conference on Learning Theory*, pages 39.1–39.26, 2012.

[2] 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.

References

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

[4] 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.

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

[6] Imad Aouali, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Bayesian off-policy evaluation and learning for large action spaces. *arXiv preprint arXiv:2402.14664*, 2024.

[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. In *Uncertainty in Artificial Intelligence*, pages 88–109. PMLR, 2024.

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

[9] Sarah Filippi, Olivier Cappe, Aurelien Garivier, and Csaba Szepesvari. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23*, pages 586–594, 2010.

References

[10] 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.

[11] Nicolas Nguyen, Imad Aouali, András György, and Claire Vernade. Prior-dependent allocations for bayesian fixed-budget best-arm identification in structured bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 379–387. PMLR, 2025.

[12] Steven Scott. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639 – 658, 2010.