Mixed-Effect Thompson Sampling

Imad Aouali [1] Branislav Kveton [2] Sumeet Katariya [2]

[1]Criteo AI Lab [2]Amazon

Table of contents

Motivation

Contextual Bandit Recap

Framework (Contextual Bandit [5, 6, 7]).

Contexts x	Actions a	Reward y
User / environment features	Items / ads / decisions	Stochastic, depends on (x, a)

Objective. Maximize expected cumulative reward; balance exploration/exploitation (UCB [4], TS [8]).

Why Structure? Three Examples

Movie recommendation

Movies share themes; learn category effects ψ_ℓ and action params θ_i. A: single param per category (biased). B: hierarchical (movie around category). C: *multi-category* per movie (our setting).

Why Structure? Three Examples

Movie recommendation

Movies share themes; learn category effects ψ_ℓ and action params θ_i. A: single param per category (biased). B: hierarchical (movie around category). C: *multi-category* per movie (our setting).

Ad placement (slates)

$K \approx L^M$ slates but only L items. Parameterize $\theta_i = \sum_\ell b_{i,\ell}\,\psi_\ell + \epsilon_i$ to share information via items/positions.

Why Structure? Three Examples

Movie recommendation

Movies share themes; learn category effects ψ_ℓ and action params θ_i. A: single param per category (biased). B: hierarchical (movie around category). C: *multi-category* per movie (our setting).

Ad placement (slates)

$K \approx L^M$ slates but only L items. Parameterize $\theta_i = \sum_\ell b_{i,\ell}\,\psi_\ell + \epsilon_i$ to share information via items/positions.

Drug design

Drugs are mixtures; dosage $b_{i,\ell}$ mixes component effects ψ_ℓ. Enables fast learning across candidates.

Model

Two-Level Graphical Model

Generative process.

$$\begin{aligned} \Psi_* &\sim Q_0, \\ \theta_{*,i} \mid \Psi_* &\sim P_{0,i}(\cdot \mid \Psi_*), \quad i \in [K], \\ Y_t \mid X_t, \theta_{*,A_t} &\sim P(\cdot \mid X_t; \theta_{*,A_t}). \end{aligned}$$

- $\Psi_* = (\psi_{*,\ell})_{\ell \leq L} \in \mathbb{R}^{Ld}$: *effect* parameters.
- $\Theta_* = (\theta_{*,i})_{i \leq K} \in \mathbb{R}^{Kd}$: *action* parameters.
- Structure via missing edges $\psi_{*,\ell} \nrightarrow \theta_{*,i}$.

Linearity in Effects (Common, Tractable)

Assume known mixing weights $b_i = (b_{i,\ell})_{\ell \le L}$ and

$$\theta_{*,i} \mid \Psi_* \sim P_{0,i}\Big(\cdot \Big| \sum_{\ell=1}^{L} b_{i,\ell}\, \psi_{*,\ell} \Big).$$

Instances.

- **Linear Gaussian** (closed-form posteriors): $\Psi_* \sim \mathcal{N}(\mu_\Psi, \Sigma_\Psi)$, $\theta_{*,i} \mid \Psi_* \sim \mathcal{N}(\sum_\ell b_{i,\ell}\psi_{*,\ell}, \Sigma_{0,i})$, $Y_t \mid X_t, \theta \sim \mathcal{N}(X_t^\top \theta, \sigma^2)$.
- **GLM** (Laplace approx): same priors, $Y_t \mid X_t, \theta \sim P$ in exp. family with mean $f(X_t^\top \theta)$ (e.g., logistic).

Algorithm

Hierarchical Sampling

Key idea. Sample effects then actions (conditional independence given Ψ).

Algorithm 1 *meTS* : Mixed-Effect Thompson Sampling

Input: Q_0, $\{P_{0,i}\}_{i \leq K}$; initialize $Q_1 \leftarrow Q_0$, $P_{1,i} \leftarrow P_{0,i}$
for $t = 1, \ldots, n$ **do**
    Sample $\Psi_t \sim Q_t$
    For each $i \in [K]$, sample $\theta_{t,i} \sim P_{t,i}(\cdot \mid \Psi_t)$
    $A_t \leftarrow \arg\max_{i \in [K]} \mathbb{E}\left[Y \mid X_t; \theta_{t,i}\right]$
    Observe $Y_t \sim P(\cdot \mid X_t; \theta_{*,A_t})$
    Update Q_{t+1} and $\{P_{t+1,i}\}$ using $H_t = (X_{1:t-1}, A_{1:t-1}, Y_{1:t-1})$
end for

Closed-Form Posteriors: Linear Case

Let $G_{t,i} = \sigma^{-2} \sum_{\ell \in S_{t,i}} X_\ell X_\ell^\top$, $B_{t,i} = \sigma^{-2} \sum_{\ell \in S_{t,i}} Y_\ell X_\ell$.

Effect posterior $Q_t = \mathcal{N}(\bar{\mu}_t, \bar{\Sigma}_t)$:

$$\bar{\Sigma}_t^{-1} = \Sigma_\Psi^{-1} + \sum_{i=1}^{K} b_i b_i^\top \otimes \left(\Sigma_{0,i} + G_{t,i}^{-1}\right)^{-1},$$

$$\bar{\mu}_t = \bar{\Sigma}_t \Big(\Sigma_\Psi^{-1} \mu_\Psi + \sum_{i=1}^{K} b_i \otimes \left((\Sigma_{0,i} + G_{t,i}^{-1})^{-1} G_{t,i}^{-1} B_{t,i}\right)\Big).$$

Action posterior $P_{t,i}(\cdot \mid \Psi_t) = \mathcal{N}(\tilde{\mu}_{t,i}, \tilde{\Sigma}_{t,i})$:

$$\tilde{\Sigma}_{t,i}^{-1} = \Sigma_{0,i}^{-1} + G_{t,i}, \qquad \tilde{\mu}_{t,i} = \tilde{\Sigma}_{t,i} \Big(\Sigma_{0,i}^{-1} \sum_{\ell=1}^{L} b_{i,\ell} \psi_{t,\ell} + B_{t,i}\Big).$$

GLM Case: Laplace Approximation

For action i:

$$\log \mathcal{L}_{t,i}(\theta) = \sum_{\ell \in S_{t,i}} Y_\ell X_\ell^\top \theta - A(X_\ell^\top \theta) + C(Y_\ell), \quad \dot{A} = f.$$

MLE and curvature:

$$\mu_{t,i}^{\mathsf{LAP}} = \arg\max_\theta \log \mathcal{L}_{t,i}(\theta), \quad G_{t,i}^{\mathsf{LAP}} = \sum_{\ell \in S_{t,i}} \dot{f}(X_\ell^\top \mu_{t,i}^{\mathsf{LAP}}) X_\ell X_\ell^\top.$$

Approximate $\mathcal{L}_{t,i} \approx \mathcal{N}(\mu_{t,i}^{\mathsf{LAP}}, (G_{t,i}^{\mathsf{LAP}})^{-1})$ and plug into the linear formulas with $G \leftarrow G^{\mathsf{LAP}}$, $G^{-1}B \leftarrow \mu^{\mathsf{LAP}}$.

Why Hierarchical Sampling? Complexity

Joint posterior over $\Theta_* \in \mathbb{R}^{Kd}$: space $\mathcal{O}(K^2 d^2)$, time $\mathcal{O}(K^3 d^3)$.
`meTS` with effects $\Psi \in \mathbb{R}^{Ld}$: space $\mathcal{O}((L^2 + K)d^2)$, time $\mathcal{O}((L^3 + K)d^3)$.
When $K \gg L$ (typical), hierarchical sampling is far cheaper while retaining cross-action coupling via Ψ.

Theory

Main Regret Bound (Linear Case)

Assume $\Sigma_{0,i} = \sigma_0^2 I_d$, $\Sigma_\Psi = \sigma_\Psi^2 I_{Ld}$, $\|X_t\|_2^2 \leq \kappa_x$, and define $\kappa_b = \max_i \|b_i\|_2^2$.

Theorem (Informal). For any $\delta \in (0, 1)$,

$$\mathcal{BR}(n) \ \leq \ \sqrt{2n\Big(\mathcal{R}^{\mathsf{A}}(n) + \mathcal{R}^{\mathsf{E}}(n)\Big)\log(1/\delta)} + cn\delta.$$

$\mathcal{R}^{\mathsf{A}}$: learning actions; $\mathcal{R}^{\mathsf{E}}$: learning effects. Both scale with d, K/L, and prior widths.

Simplified (set $\kappa_x = \kappa_b = \sigma = 1$):

$$\mathcal{BR}(n) = \tilde{\mathcal{O}}\Big(\sqrt{nd\left(K\sigma_0^2 + L\sigma_\Psi^2(1 + \sigma_0^2)\right)}\Big).$$

Lower priors $\Rightarrow$ lower regret; fewer parameters $(K, L, d) \Rightarrow$ easier.

Benefits of Structure

- **If Ψ_* known** $(\sigma_\Psi = 0)$: $\tilde{\mathcal{O}}(\sqrt{ndK\sigma_0^2})$ (no L term).
- **If perfect linear tie** $(\sigma_0 = 0)$: $\tilde{\mathcal{O}}(\sqrt{ndL\sigma_\Psi^2})$ (no K term).
- **No structure modeled:** marginalize Ψ; prior width inflates to $\sigma_0^2 + \sigma_\Psi^2 \Rightarrow$ regret $\tilde{\mathcal{O}}(\sqrt{ndK(\sigma_0^2 + \sigma_\Psi^2)})$.

When $K \gg L$ and effects are uncertain ($\sigma_\Psi \gg \sigma_0$), `meTS` gains $\sim \sqrt{K/L}$.

Proof Sketch

1. Russo&Van Roy decomposition: reduce to bounding $\sum_t \|X_t\|^2_{\hat{\Sigma}_{t,A_t}}$.
2. Total covariance decomposition with mixing: for $\Gamma_i = b_i^\top \otimes I_d$, $\hat{\Sigma}_{t,i} = \tilde{\Sigma}_{t,i} + \tilde{\Sigma}_{t,i}\Sigma_{0,i}^{-1}\Gamma_i\bar{\Sigma}_t\Gamma_i^\top\Sigma_{0,i}^{-1}\tilde{\Sigma}_{t,i}$.
3. Control via eigenvalues of $\Gamma_i\Gamma_i^\top \preceq \|b_i\|_2^2 I$; sum information gains over rounds.

Structure Learning

Proxy Structure from Offline Embeddings

Given offline $\hat{\theta}_i$ (e.g., MF embeddings), fit GMM with L clusters:

- Cluster centers μ_{ψ_ℓ}, covariances Σ_{ψ_ℓ} $\Rightarrow$ effect prior mean/cov.
- Membership probs $\Rightarrow$ mixing weights $b_{i,\ell}$.

Plugs into priors of `meTS` *; bridges offline representation learning with online exploration.*

Experiments

Synthetic: Linear Logistic

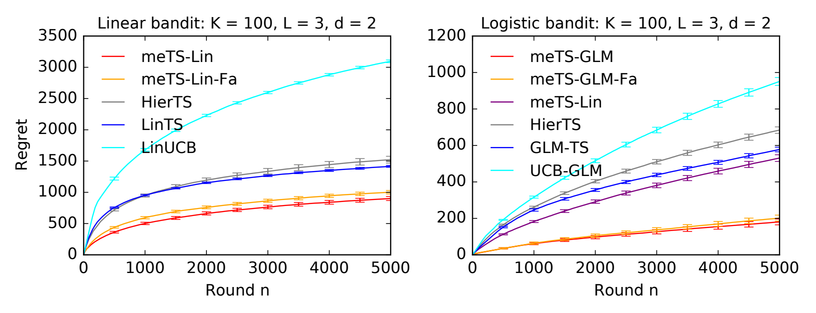

Figure 1: *meTS* (and factored variant) vs. structure-agnostic baselines and hierarchical TS with single effect.

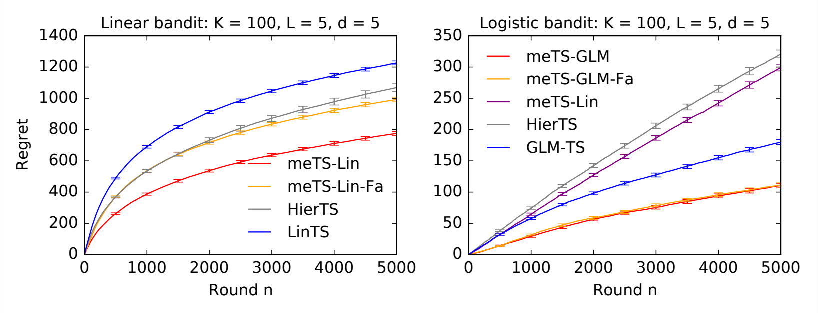

Figure 2: Proxy structure via GMM on movie embeddings; *meTS* wins under both Gaussian and logistic rewards.

Conclusion

Conclusion

- **Model:** actions depend on *multiple* shared effects.
- **Algorithm:** *meTS* with hierarchical TS; closed-form linear posteriors; Laplace for GLM.
- **Theory:** regret splits into action+effect learning; shows structure benefits.
- **Practice:** competitive and scalable; proxy structures from offline data.

Limitations: prior/mixing misspecification; beyond-Gaussian posteriors; learned $b_{i,\ell}$ dynamics.

Extensions: we extended this work to deep hierarchies [1], to diffusion models [2], and off-policy learning [3].

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, Victor-Emmanuel Brunel, David Rohde, and Anna Korba. Bayesian off-policy evaluation and learning for large action spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR, 2025.

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

[5] 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

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

[7] 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.

[8] 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.