## 1. Prior-Aligned Meta-RL: Thompson Sampling with Learned Priors and Guarantees in Finite-Horizon MDPs

**Accession number:** 20250491480
**Authors:** Zhou, Runlin (1); Chen, Chixiang (2); Chen, Elynn (3)
**Author affiliation:** (1) Department of Statistics, University of Science and Technology of China, China; (2) Department of Epidemiology and Public Health, University of Maryland, Baltimore, United States; (3) Department of Technology, Operations, and Statistics, Stern School of Business, New York University, United States
**Corresponding author:** Chen, Elynn(elynn.chen@stern.nyu.edu)
**Source title:** arXiv
**Abbreviated source title:** arXiv
**Issue date:** October 6, 2025
**Publication year:** 2025
**Language:** English
**E-ISSN:** 23318422
**Document type:** Preprint (PP)
**Repository:** arXiv
**Abstract:** We study meta-reinforcement learning in finite-horizon MDPs where related tasks share similar structures in their optimal action-value functions. Specifically, we posit a linear representation Q#h(s, a) = Φh(s, a) θh(k) and place a Gaussian meta-prior N(θh#, Σ#h) over the task-specific parameters θh(k). Building on randomized value functions, we propose two Thompson-style algorithms: (i) MTSRL, which learns only the prior mean and performs posterior sampling with the learned mean and known covariance; and (ii) MTSRL+, which additionally estimates the covariance and employs prior widening to control finite-sample estimation error. Further, we develop a prior-alignment technique that couples the posterior under the learned prior with a meta-oracle that knows the true prior, yielding meta-regret guarantees: we match prior-independent Thompson sampling in the small-task regime and strictly improve with more tasks once the prior is learned. Concretely, for known covariance we obtain Õ(H4S3/2 √ANK) meta-regret, and with learned covariance Õ(H4S3/2 √AN3K); both recover a better behavior than prior-independent after K Õ(H2) and K Õ(N2H2), respectively. Simulations on a stateful recommendation environment (with feature and prior misspecification) show that after brief exploration, MTSRL/MTSRL+ track the meta-oracle and substantially outperform prior-independent RL and bandit-only meta-baselines. Our results give the first meta-regret guarantees for Thompson-style RL with learned Q-priors, and provide practical recipes (warm-start via RLSVI, OLS aggregation, covariance widening) for experiment-rich settings. © 2025, CC BY.
**Number of references:** 0
**Main heading:** Structural optimization
**Controlled terms:** Machine learning
**Uncontrolled terms:** Bayesian - Bayesian RL - Finite horizons - Finite-horizon MDP - Learned prior - Meta-regret - Meta-reinforcement learning - Reinforcement learnings - Thompson samplings - Value functions
**Classification code:** 1101.2 Machine Learning - 1201.7 Optimization Techniques
**Numerical data indexing:** Electrical conductance 4.00E+00S, Inductance 2.00E+00H, Temperature 3.00E+00K
**DOI:** 10.48550/arXiv.2510.05446
**Compendex references:** YES
**Preprint ID:** 2510.05446v1
**Preprint source website:** https://arxiv.org
**Preprint ID type:** ARXIV
**Database:** Compendex
**Data Provider:** Engineering Village