Co-Exploration and Co-Exploitation via Shared Structure in Multi-Task Bandits

Sumantrak Mukherjee
DFKI
sumantrak.mukherjee@dfki.de

Serafima Lebedeva
DFKI
RPTU Kaiserslautern–Landau

Jasmin Brandt
Bielefeld University

Valentin Margraf
LMU Munich
MCML

Jonas Hanselle
DFKI
LMU Munich
MCML

Kanta Yamaoka
DFKI
RPTU Kaiserslautern–Landau

Viktor Bengs
DFKI

Stefan Konigorski
German Institute of Human Nutrition
University of Potsdam

Eyke Hüllermeier
DFKI
LMU Munich
MCML

Sebastian Josef Vollmer
DFKI
RPTU Kaiserslautern–Landau

Abstract

We propose CoCo-Bandits, a Bayesian framework for efficient exploration in contextual multi-task multi-armed bandits with partially observed context and latent reward dependencies across tasks and arms. Our approach learns a global joint distribution across tasks while retaining personalised inference, and distinguishes structural uncertainty from user-specific uncertainty due to incomplete context and limited history. Concretely, we approximate the joint distribution over tasks and rewards with a particle-based log-density Gaussian Process, enabling discovery of inter-arm and inter-task dependencies without parametric latent assumptions. We show theoretically that learning the meta-prior yields sublinear cumulative multi-task regret under posterior contraction assumptions. Empirically, on synthetic benchmarks and a real-world-dataset-based environment, our method outperforms parametric baselines and shows that nonparametric latent modelling improves robustness under misspecification and complex heterogeneity.

1 Introduction

Multi-armed bandits (MABs) formalise sequential decisions in settings such as personalised recommendation [Zhou et al., 2017] and clinical decision support [Durand et al., 2018, Shrestha and Jain, 2021, Aziz et al., 2021]. In both settings a system repeatedly chooses among a small set of interventions for many users. Each user offers only a short interaction window. The central question is therefore how to transfer information from earlier users without treating superficially similar users as identical.

We study multi-task contextual bandits [Deshmukh et al., 2017] with *partial* task context. Observable features such as demographics or test results, written $\mathbf{x}_{\text{obs}}$, support personalisation and transfer. Other factors, written $\mathbf{x}_{\text{unobs}}$, stay unobserved and can change which arm is best. A useful policy must therefore learn what can be shared at the population level, and must at the same time keep uncertainty about the reward vector of the user currently served.

19th European Workshop on Reinforcement Learning (EWRL 2026).

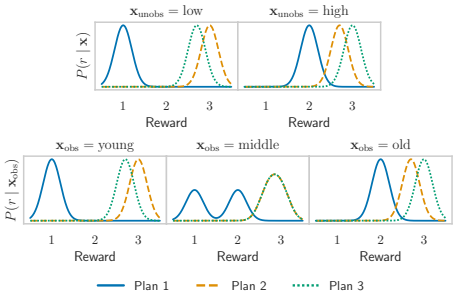


Figure 1: Reward distributions conditioned on the unobservable feature *metabolic type* (first row) and on the observable feature *age group* (second row).

	MT	PC	NPLRS	UD	DR	CI
HLINUCB	✗	✓	✗	✗	✗	✗
KMTL-UCB	✓	✗	n/a	✗	✗	✗
GRADBAND	✗	✗	n/a	✗	✗	✗
ROBUSTAGG	✓	n/a	n/a	✗	✗	✓
MTTS	✓	✗	✗	✗	✗	✓
HIERTS	✓	n/a	✗	✗	✓	✓
ROME	✓	✗	n/a	✗	✓	✓
COCO-BANDITS	✓	✓	✓	✓	✓	✓

Table 1: Comparison with representative prior methods. MT: multi-task; PC: partial context; NPLRS: nonparametric latent reward structures; UD: explicit uncertainty distinction; DR: dynamic user recruitment; CI: concurrent interaction.

Figure 1 gives the running example. A system recommends three dietary plans. It observes the age group and does not observe the metabolic type. Age is informative, but age does not determine the best plan: Plan 2 is preferred for one metabolic type and Plan 3 for another. Within an ambiguous subgroup, the reward from Plan 1 can still be valuable, because it changes the posterior over the entire reward vector. COCO-BANDITS does not try to recover a user's metabolic type. It learns the induced distribution of reward vectors conditional on age, and it uses early observations to personalise later decisions.

Related work. Our work relates to multi-task bandits [Wang et al., 2021, Deshmukh et al., 2017], hierarchical models [Hong et al., 2022], meta-learning in bandits [Bastani et al., 2022, Peleg et al., 2022, Basu et al., 2021, Cella et al., 2020, Ortega et al., 2019], mixed-effects information sharing [Aouali et al., 2023, Huch et al., 2024], and partial-observability and cold-start methods [Wang et al., 2016, Bharadhwaj, 2019, Silva et al., 2023]. These methods make different choices about the structure shared across users, the treatment of missing context, and the arrival protocol. Our contribution is a nonparametric population model that keeps the reward vector latent and conditions it on the context that is actually observed. Table 1 positions the setting relative to representative methods. Appendix Section B gives the detailed comparison.

Contributions.

- **Model.** We formulate COCO-BANDITS, a hierarchical contextual bandit in which a shared population model captures dependence across arms and users, and each user keeps a personalised posterior.
- **Decision rules.** We make the three uncertainty sources — population structure, user-specific rewards, and reward noise — operational in two action rules: nonparametric meta-prior Thompson sampling (NPM-TS) and global information-directed sampling (GIDS).
- **Theory.** For NPM-TS, we bound the post-alignment performance gap by the contraction of the learned population prior, in a sequential alignment construction.
- **Experiments.** Across four heterogeneous synthetic conditions and a MovieLens semi-synthetic environment, we measure both the benefit of information sharing and the trade-off between flexibility and parametric efficiency.

In Section 2, we introduce the problem setting. In Section 3, we introduce COCO-BANDITS and its two decision rules. In Section 4, we state the numerical approximation and the concurrent algorithm. In Section 5, we present the multi-task regret bound. In Section 6, we report the experiments. In Section 7, we summarise the scope of the results.

2 Problem Setting

This section defines the environment, the interaction protocol, the information available to the learner, and the performance criterion. It imposes no parametric form on the population distribution.

2.1 Environment and Interaction Protocol

Users (tasks) are indexed by $i = 1, 2, \ldots$. At the arrival of user i, the environment draws

$$(\boldsymbol{\mu}_i, \mathbf{x}_{i,\text{obs}}, \mathbf{x}_{i,\text{unobs}}) \sim P^\star,$$

where $\boldsymbol{\mu}_i = (\mu_{i,a})_{a\in[K]} \in \mathbb{R}^K$ is the vector of latent arm means, and $\boldsymbol{x}_i = (\mathbf{x}_{i,\text{obs}}, \mathbf{x}_{i,\text{unobs}})$ is a context split into an observed and an unobserved component. The draws are i.i.d. across users from one fixed and unknown population distribution $P^\star$, which describes the data-generating environment; Section 3 introduces the learner's model of it. Each user stays for the fixed and known horizon T_{user}, so neither the context nor the arm means change within a user session.

The learner has a total of T global interactions. Users arrive in batches of fixed size b, which makes concurrent transfer explicit. At global interaction t, the learner serves one user $i(t)$ from the active set I_{t-1}, using round-robin sweeps. Let U_t be the set of all users that have arrived by time t, and define

$$h_{i,t} := \sum_{s=1}^{t} \mathbb{1}\{i(s) = i\}, \qquad I_t := \{i \in U_t : h_{i,t} < T_{\text{user}}\},$$

where $h_{i,t}$ is the number of interactions with user i by time t. After a complete sweep through the active users, the environment recruits a new batch of b users, and completed users leave. Algorithm 1 gives the deployed bookkeeping. The batching convention is an arrival protocol, not a modelling assumption: the inference and the action rules do not require fixed batch sizes.

When the learner selects $a_t \in [K] := \{1, \ldots, K\}$ for user $i(t)$, the environment generates

$$r_t = \mu_{i(t),a_t} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma^2).$$

2.2 Observations

At a user's arrival the learner observes $\mathbf{x}_{i,\text{obs}}$. After it chooses an action, it observes that action and its noisy reward (a_t, r_t). It never observes $\mathbf{x}_{i,\text{unobs}}$, the latent mean vector $\boldsymbol{\mu}_i$, or the population distribution $P^\star$. We write

$$\mathcal{H}_{i,t} := \{(a_s, r_s) : s \le t,\ i(s) = i\}, \qquad \mathcal{H}_t := \{(i(s), a_s, r_s) : s \le t\}$$

for the user-specific and the global interaction history. The full observed data are $\mathcal{D}_t := \{(\mathbf{x}_{i,\text{obs}}, \mathcal{H}_{i,t}) : i \in U_t\}$.

Let $p^\star$ denote the density of $P^\star$. The conditional reward distribution that is identifiable from the observed context is the marginal

$$p^\star(\boldsymbol{\mu} \mid \mathbf{x}_{\text{obs}}) = \int p^\star(\boldsymbol{\mu} \mid \mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{unobs}})\, p^\star(\mathbf{x}_{\text{unobs}} \mid \mathbf{x}_{\text{obs}})\, d\mathbf{x}_{\text{unobs}}. \tag{1}$$

Unobserved context therefore appears as residual heterogeneity in the conditional reward distribution. In the dietary-plan example the learner conditions on age and integrates over plausible metabolic types. Equation (1) states which information about the reward distribution the observed context makes available; recovering $\mathbf{x}_{\text{unobs}}$ itself is beyond the scope of this paper.

2.3 Performance Metric

The online objective is to maximise the expected cumulative *latent-mean* reward. For an algorithm $\mathcal{A}$, define

$$V_{\mathcal{A}}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \mu_{i(t),a_t^{\mathcal{A}}}\right].$$

Maximising $V_{\mathcal{A}}$ is equivalent to minimising the cumulative Bayes regret [Russo and Van Roy, 2014, Wan et al., 2021] relative to the best arm of each realised user,

$$\text{BR}_{\mathcal{A}}(T) := \mathbb{E}\left[\sum_{t=1}^{T} \left(\max_{a\in[K]} \mu_{i(t),a} - \mu_{i(t),a_t^{\mathcal{A}}}\right)\right]. \tag{2}$$

Smaller values are better. The expectation is taken over users and latent means drawn from the fixed unknown population $P^\star$, over the observation noise that enters through the histories and the resulting actions, over the arrival and ordering protocol, and over all algorithmic randomisation. The population distribution $P^\star$ is held fixed throughout. The latent-mean definition is equivalent in expectation to a definition that uses realised rewards with zero-mean noise, and it separates regret from observation noise. Sublinear growth $\mathrm{BR}_{\mathcal{A}}(T) = o(T)$ implies $\mathrm{BR}_{\mathcal{A}}(T)/T \to 0$, so the average per-round performance approaches that of the per-user best arm.

Multi-task regret serves as an *evaluation reference*. It isolates the cost of learning shared structure by comparing $\mathcal{A}$ with *True-Prior Thompson Sampling* (TP-TS). For user i, TP-TS uses the true task-conditional prior

$$\pi_i^\star(d\boldsymbol{\mu}) := P^\star(d\boldsymbol{\mu} \mid \mathbf{x}_{i,\mathrm{obs}})$$

and updates it with *its own* action–reward history. TP-TS is therefore a fixed decision rule endowed with the true prior. The multi-task regret of $\mathcal{A}$ is

$$\mathrm{MTR}_{\mathcal{A}}(T) := V_{\text{TP-TS}}(T) - V_{\mathcal{A}}(T) = \mathrm{BR}_{\mathcal{A}}(T) - \mathrm{BR}_{\text{TP-TS}}(T). \tag{3}$$

Multi-task regret is therefore the part of the performance gap that is attributable to a learned rather than a true population prior, with the within-task regret of TP-TS removed. The figures label empirical estimates of these cumulative quantities CBR and CMTR.

2.4 Problem Statement

Given an environment that generates data as in Section 2.1, and observations restricted as in Section 2.2, our objective is to construct a policy that minimises the cumulative Bayes regret (2) over the global horizon T. Admissible policies are those whose action at interaction t is measurable with respect to $\sigma(\mathcal{D}_{t-1}, \mathbf{x}_{i(t),\mathrm{obs}})$ and the algorithm's own randomisation. We report the multi-task regret (3) alongside Bayes regret, in order to separate the cost of learning the population prior from the within-task exploration cost.

3 Method

Section 2 specifies the environment and the information available to the learner, without imposing a parametric form on $P^\star$. This section defines CoCo-Bandits: a population model for the observed-data distribution, a hierarchical posterior, and two decision rules built from that posterior. Throughout, untilded quantities such as Q_t and $\pi_t^{i,f}$ denote the Bayesian objects themselves; Section 4 introduces their computable counterparts.

3.1 Population Model

Let $z = (\boldsymbol{\mu}, \mathbf{x}_{\mathrm{obs}})$ belong to the joint reward–context space $\mathcal{Z}$. We place a Gaussian-process prior Q_0 on an unnormalised log-density $f : \mathcal{Z} \to \mathbb{R}$:

$$f \sim \mathcal{GP}(m, k), \qquad p^{(f)}(z) := \frac{\exp(f(z))}{Z_f}, \qquad Z_f := \int_{\mathcal{Z}} \exp(f(z'))\, dz'.$$

Thus f is a log-*potential*, with $\log p^{(f)}(z) = f(z) - \log Z_f$, and every draw f induces a joint distribution $P^{(f)}$. The meta-prior $Q_0(df)$ therefore represents uncertainty about the population density $p^\star(\boldsymbol{\mu}, \mathbf{x}_{\mathrm{obs}})$. Modelling the joint distribution, rather than independent per-arm models, admits multimodality and arbitrary dependence between arms and observed context without fixing a parametric latent-variable family. In particular, $p^{(f)}(\boldsymbol{\mu} \mid \mathbf{x}_{i,\mathrm{obs}})$ can encode inter-arm structure: a reward from one arm updates beliefs about the arms that were not pulled, through the joint posterior over $\boldsymbol{\mu}_i$.

The hierarchy represents three distinct uncertainty sources [Hüllermeier and Waegeman, 2021]. *Population uncertainty* is uncertainty about the shared density, and Q_t represents it. *User-specific uncertainty* is uncertainty about $\boldsymbol{\mu}_i$, conditional on a population model and on the partial context and history of user i. *Reward noise* is the variance σ^2 that remains in an observed reward even conditional on $\boldsymbol{\mu}_i$. The next subsection makes the first two posterior layers explicit.

3.2 Hierarchical Posterior

The population posterior after t interactions is $Q_t(df) := Q_0(df \mid \mathcal{D}_t)$. Conditional on the selected actions, it factorises as

$$Q_t(df) \;\propto\; \left[\prod_{i\in U_t} p^{(f)}(\mathbf{x}_{i,\text{obs}})\, p^{(f)}(\mathcal{H}_{i,t} \mid \mathbf{x}_{i,\text{obs}})\right] Q_0(df), \tag{4}$$

where

$$p^{(f)}(\mathcal{H}_{i,t} \mid \mathbf{x}_{i,\text{obs}}) = \int \left[\prod_{(a_s,r_s)\in\mathcal{H}_{i,t}} \mathcal{N}(r_s;\mu_{a_s},\sigma^2)\right] p^{(f)}(\boldsymbol{\mu} \mid \mathbf{x}_{i,\text{obs}})\, d\boldsymbol{\mu}.$$

The context-density factor is present because the model learns the joint distribution of $(\boldsymbol{\mu}, \mathbf{x}_{\text{obs}})$. The product runs over U_t rather than over the active set I_t, so that completed users continue to inform the population model.

For a specific user and a fixed population model f, the personalised posterior is

$$\pi_t^{i,f}(\boldsymbol{\mu}) := p^{(f)}(\boldsymbol{\mu} \mid \mathbf{x}_{i,\text{obs}}, \mathcal{H}_{i,t}) \;\propto\; p^{(f)}(\boldsymbol{\mu} \mid \mathbf{x}_{i,\text{obs}}) \prod_{(a_s,r_s)\in\mathcal{H}_{i,t}} \mathcal{N}(r_s;\mu_{a_s},\sigma^2). \tag{5}$$

Marginalising over the population posterior gives $\pi_t^i(\boldsymbol{\mu}) = \int \pi_t^{i,f}(\boldsymbol{\mu}) Q_t(df)$. Thus Q_t describes what is shared across users, and $\pi_t^{i,f}$ describes the uncertainty that remains for user i.

Sequential update. After the learner observes r_t from action a_t for $i = i(t)$, the incremental predictive likelihood is

$$p^{(f)}(r_t \mid a_t, \mathbf{x}_{i,\text{obs}}, \mathcal{H}_{i,t-1}) = \int \mathcal{N}(r_t;\mu_{a_t},\sigma^2)\, \pi_{t-1}^{i,f}(\boldsymbol{\mu})\, d\boldsymbol{\mu}.$$

Consequently $Q_t(df) \propto p^{(f)}(r_t \mid a_t, \mathbf{x}_{i,\text{obs}}, \mathcal{H}_{i,t-1}) Q_{t-1}(df)$, together with the context factor in (4) at the arrival of user i. Thus Q_t is a posterior measure over population models.

3.3 Decision Rules

We define both action rules from the posterior Q_t and the personalised belief $\pi_t^{i,f}$. NPM-TS mainly *co-exploits*: it uses data from all users to construct a better prior, and then acts from the current user's posterior. GIDS deliberately *co-explores*: it can trade one-step expected reward for information that improves later expected reward through the shared population model. Multi-task regret measures the outcome of this transfer. Section 4 states how these quantities are approximated in deployment, and the analysis in Section 5 covers NPM-TS.

Nonparametric meta-prior Thompson sampling. NPM-TS makes the two posterior draws explicit. Before serving $i = i(t)$, it computes

$$\text{(i) } f_t \sim Q_{t-1}, \qquad \text{(ii) } \widetilde{\boldsymbol{\mu}}_{i,t} \sim \pi_{t-1}^{i,f_t}, \qquad \text{(iii) } a_t \in \arg\max_{a\in[K]} \widetilde{\mu}_{i,t,a}.$$

The draw of f_t propagates uncertainty about the shared population structure. The draw of $\widetilde{\boldsymbol{\mu}}_{i,t}$, conditional on f_t, propagates uncertainty about the current user after the learner has observed $\mathbf{x}_{i,\text{obs}}$ and $\mathcal{H}_{i,t-1}$. Observation noise affects both draws only through the reward histories that enter these posteriors.

Global information-directed sampling. GIDS follows information-directed sampling [Russo and Van Roy, 2014], but it targets information about the population posterior rather than information about the current best arm alone. For a sampled f and user $i = i(t)$, let

$$\Delta_t^{i,f}(a) := \mathbb{E}_{\boldsymbol{\mu}\sim\pi_{t-1}^{i,f}}[\max_{k\in[K]} \mu_k - \mu_a]$$

denote the expected local regret of arm a. Let $p_{t-1}^{i,f}(\cdot \mid a)$ be the Gaussian posterior-predictive reward distribution induced by $\pi_{t-1}^{i,f}$, and let $Q_t^{a,r,i}$ denote the population posterior obtained after a hypothetical reward r from the pair (i, a). The expected global information gain is

$$\mathcal{I}_t^{i,f}(a) := \mathbb{E}_{r\sim p_{t-1}^{i,f}(\cdot|a)}\Big[\mathrm{Ent}(Q_{t-1}) - \mathrm{Ent}(Q_t^{a,r,i})\Big].$$

The rule chooses the action distribution

$$\rho_f^\star \in \arg\min_{\rho\in\Delta_K} \frac{\Delta_t^{i,f}(\rho)^2}{\mathcal{I}_t^{i,f}(\rho)},$$

and then samples $a_t \sim \rho_f^\star$. The rule optimises the ratio under one sampled structural hypothesis f. A regret guarantee for GIDS requires a separate information-gain analysis and is beyond the scope of this paper.

4 Inference and Implementation

Section 3 defines the statistical target and the decision rules. Computing those objects in closed form is intractable. This section states the numerical approximation used in the experiments, and the online procedure that deploys it.

4.1 Sources of Intractability

The normalising constant Z_f is an integral over a continuous joint reward–context space, and the reward component of that space has dimension K. The posterior Q_t is a measure over the infinite-dimensional function f. Both the population likelihood in (4) and the user posterior in (5) require marginalisation of the latent vector $\boldsymbol{\mu}$. These operations have no closed form in general. For GIDS, the information gain additionally requires hypothetical posterior updates and posterior entropies for every candidate action and every candidate reward. Both the posterior and the action computation therefore require numerical approximation.

4.2 Finite Representation and Particle Approximation

We discretise $\mathcal{Z}$ on a finite grid, replace Z_f by the corresponding finite sum, and normalise the grid values with a softmax. We represent f by J terms of a truncated Karhunen–Loève expansion [Qi et al., 2012]. Sequential Monte Carlo (SMC) [Chopin and Papaspiliopoulos, 2020] then maintains S weighted particles over this finite representation:

$$\widetilde{Q}_t := \sum_{s=1}^{S} w_t^{(s)} \delta_{f_t^{(s)}} \;\approx\; Q_t.$$

For each particle, the same grid and Gaussian likelihood give a discrete personalised posterior $\widetilde{\pi}_t^{i,f_t^{(s)}}$ that approximates $\pi_t^{i,f_t^{(s)}}$. The experiments use a Nyström approximation with M_{ind} inducing points, which reduces the Gaussian-process cost. Deployed NPM-TS therefore samples a particle from $\widetilde{Q}_t$ and then samples from the associated $\widetilde{\pi}_t^{i,f}$, and deployed GIDS evaluates its regret and information terms with the same particles and grid. Appendix Sections D.2, D.3, and D.5 give the implementation details.

4.3 Algorithm

Algorithm 1 is the deployed concurrent version of CoCo-Bandits. Its state consists of the shared approximate posterior $\widetilde{Q}_t$, the set of active users with their contexts and interaction counts, and the per-user histories. At every interaction the algorithm draws a population particle, forms the personalised approximate posterior of the user being served, selects an action with approximate NPM-TS or approximate GIDS, and folds the observed reward back into $\widetilde{Q}_t$ before the next interaction. Users leave the active set once they reach the horizon T_{user}, and a new batch is recruited after each complete sweep. The algorithm returns the population posterior together with the actions taken.

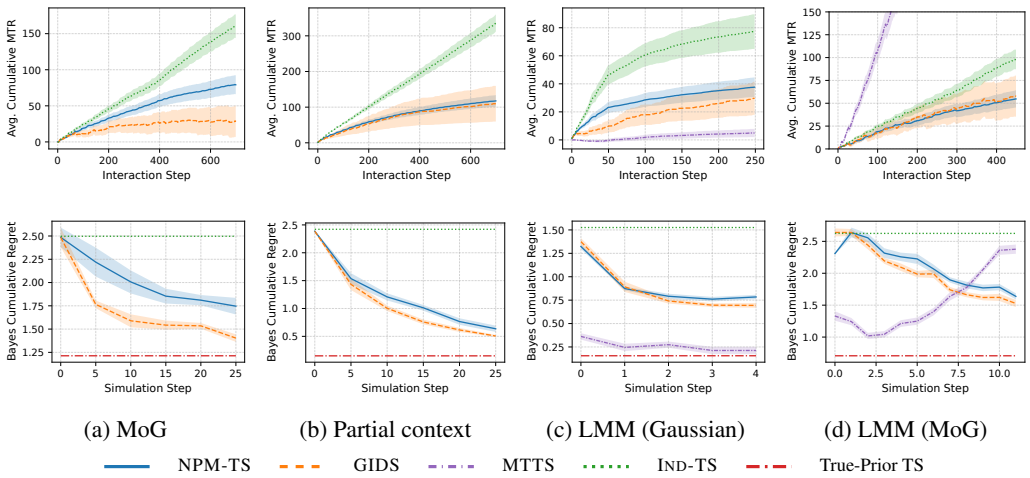


Figure 2: Four experimental conditions. MoG denotes a mixture of Gaussians, and LMM denotes a linear mixed model. Each **column** is one condition, and the label (a)–(d) below the column applies to both of its panels. The **top row** shows online empirical cumulative multi-task regret (CMTR) as users interact; the **bottom row** shows cumulative Bayes regret (CBR) for the same condition, measured on a fixed evaluation batch. Lower is better in both rows. Curves are means over 10 seeds, with ± 1 standard-error bands. IND-TS appears in conditions (a) and (b); MTTS appears additionally in (c) and (d), where its assumptions can be matched. True-Prior TS is the evaluation reference of Section 2.3. The x-axis counts interaction rounds.

Algorithm 1: CoCo-Bandits

Input: approximate meta-prior $\widetilde{Q}_0$ of the meta-prior Q_0; known user horizon T_{user}; batch size b; reward noise σ; action rule (NPM-TS or GIDS)

Initialise: $t \leftarrow 0$; recruit $U \leftarrow I \leftarrow \{1, \ldots, b\}$; observe $\mathbf{x}_{i,\text{obs}}$ and set $h_i \leftarrow 0$ for each $i \in I$

Output: population posterior $\widetilde{Q}_t$ and the sequence of selected actions

while the global interaction budget remains **do**
  $I_{\text{sweep}} \leftarrow I$ ▷ Snapshot of the active set
  for each $i \in I_{\text{sweep}}$ **do**
    if t is within the interaction budget **then**
      Draw $f_t \sim \widetilde{Q}_{t-1}$ and form $\widetilde{\pi}^{i,f_t}_{t-1}$ from $\mathbf{x}_{i,\text{obs}}, \mathcal{H}_{i,t-1}$
      Select a_t with the approximate action rule; observe r_t
      Update $\mathcal{H}_t, \mathcal{D}_t, \widetilde{Q}_t$; set $h_i \leftarrow h_i + 1$ and $t \leftarrow t + 1$
      if $h_i = T_{\text{user}}$ **then**
        $I \leftarrow I \setminus \{i\}$ ▷ User i has completed its session
      end if
    end if
  end for
  if t is within the interaction budget **then**
    Recruit a new batch $\mathcal{B}$ of b users; observe $\boldsymbol{x}_{j,\text{obs}}$ and set $h_j \leftarrow 0$ for $j \in \mathcal{B}$
    $I \leftarrow I \cup \mathcal{B}, \quad U \leftarrow U \cup \mathcal{B}$
  end if
end while

5 Theoretical Analysis

Scope of the analysis. Algorithm 1 is concurrent, continuously updated, and numerically approximate. To obtain a clean prior-sensitivity argument, the theorem analyses a more conservative *sequential* construction. In that construction N users arrive one at a time, each user has horizon T_{user},

and the first K interactions pull every arm once for alignment. The remaining horizon is

$$H := T'_{\text{user}} := T_{\text{user}} - K.$$

The alignment data construct the learned prior that post-alignment Thompson sampling uses. Alignment yields aligned, non-adaptive data, for which standard posterior contraction results apply. Contraction under fully adaptive bandit feedback, and the analysis of Algorithm 1 itself, are beyond the scope of this paper.

For task $i = n + 1$, let $\pi_i^\star(d\boldsymbol{\mu}) = P^\star(d\boldsymbol{\mu} \mid \mathbf{x}_{i,\text{obs}})$, and let $\hat{\pi}_{n,i}$ be the conditional prior learned from the preceding n aligned tasks. Let $p_{\min} > 0$ lower-bound the true observed-context marginal density on the contexts considered in the proof. The learned joint density $\hat{p}_n$ contracts in Hellinger distance at rate ε_n: for a constant $C_{\text{ctr}} > 0$, the event $h(\hat{p}_n, p^\star) \le C_{\text{ctr}}\varepsilon_n$ has an exponentially decaying complement probability. Appendix Assumption 6 supplies the pointwise conditional control that the proof applies at each observed context. The per-task quantity MTR_i is the gap between learned-prior and True-Prior TS, as in (3).

Theorem 1 (Cumulative multi-task regret of NPM-TS)**.** *Assume $\mu_{i,a} \in [0, B]$, i.i.d. sequential tasks from $P^\star$, the aligned data-collection process described above, log-Gaussian-process posterior contraction at rate ε_n, and the local conditional-stability condition in Appendix Assumption 6. Let ℓ be the finite burn-in index after which the contraction bound is non-vacuous. Then, for positive constants c_1, c_2, C_{ctr}, the full-horizon regret gap satisfies*

$$\sum_{i=1}^{N} \mathbb{E}[\text{MTR}_i] \le BT_{\text{user}}\ell + BKN + \frac{8BH^2C_{\text{ctr}}}{p_{\min}} \sum_{n=\ell}^{N-1} \varepsilon_n + BH \sum_{n=\ell}^{\infty} c_1 e^{-c_2 n\varepsilon_n^2}, \tag{6}$$

where B bounds the mean rewards, $p_{\min}$ lower-bounds the observed-context marginal density, and c_1, c_2, C_{ctr} depend on the prior and on the contraction rate but not on N.

The four terms have distinct sources. The term $BT_{\text{user}}\ell$ is the burn-in cost of the first ℓ tasks, for which the contraction bound is vacuous. The term BKN arises from the coupling of the learned-prior and True-Prior TS histories in the proof, and it cancels against a True-Prior TS reference that is assigned the same alignment phase. The third term is the post-alignment population-learning cost, and with fixed H it is sublinear whenever $\sum_{n \le N} \varepsilon_n = o(N)$; this is the sublinearity claim of the theorem. The fourth term is the contraction tail. The H^2 dependence is inherited from the Thompson-sampling prior-sensitivity bound of Simchowitz et al. [2021]. A corresponding guarantee for GIDS, and an end-to-end bound on the numerical approximation error, are beyond the scope of this paper. Appendix Sections E and E.2 give the full assumptions and the proof.

6 Experiments

6.1 Experimental Setup

We evaluate CoCo-Bandits with NPM-TS and GIDS on four synthetic conditions. Each condition isolates a different part of the claim: a multimodal population (MoG); a partially observed mixed-type context; a linear mixed-model (LMM) population that matches a Gaussian mixed-effects prior; and an LMM population whose random effect is mixture-distributed and therefore violates that prior. The first two conditions use dynamic recruitment. The matched LMM condition uses a fixed cohort, which mirrors the MTTS protocol. Appendix Section F gives the data-generating processes. Appendix Sections G.2 and G.1 report the MovieLens 1M semi-synthetic experiment and the ablations.

We compare against independent Thompson sampling with a uniform prior (Ind-TS) in the first two conditions. In the LMM conditions we compare additionally against the Gaussian linear mixed-effects version of MTTS [Wan et al., 2021], which is the information-sharing baseline most directly compatible with the matched protocol. Other contextual, clustering, and hierarchical methods would require changes to the model or to the arrival protocol before they could accommodate the same partial-context population; adapting them is beyond the scope of this paper.

The experiments test a deliberately conditional claim: flexible sharing should help when the population structure is learnable but not well described by a restrictive parametric family, whereas a well-specified parametric method should be more statistically efficient. All curves report means over 10 random seeds with ± 1 standard-error bands. Appendix Sections F.4 and F.5 report the compute resources, the random seeds, and the hyperparameters.

6.2 Results

Figure 2 reports the four conditions, one per column. The top row measures online CMTR as users interact. The bottom row evaluates the learned prior of the same condition on a fixed batch over 10 Thompson-sampling pulls and reports CBR, so each bottom panel is a held-out complement to the live-interaction curve directly above it. Lower is better in both rows.

In the multimodal MoG condition (Figure 2a), both nonparametric rules narrowed their gap to the True-Prior TS reference over the plotted budget. GIDS was the stronger of the two here. An arm that is locally suboptimal can identify a population subgroup and thereby improve later personalisation, which is the mechanism illustrated by the dietary-plan example.

In the partial-context condition (Figure 2b), the two nonparametric rules separated less. The hidden variables create heterogeneity, but they do not create sharply separated groups with different optimal arms, so the additional global exploration of GIDS bought less. Both rules nevertheless improved the held-out prior over the observed interaction budget.

The Gaussian LMM condition (Figure 2c) matches the parametric assumptions and the fixed-cohort protocol of MTTS. In this matched case MTTS had the strongest early performance, which shows the statistical efficiency that correct structure makes available. The nonparametric rules stayed competitive with IND-TS over the plotted budget.

In the mixture-LMM condition (Figure 2d), users arrive dynamically and the Gaussian random-effect specification is misspecified. The MTTS curves degraded after the initial users, which is consistent with an overconfident population prior that transfers poorly to later subgroups. The nonparametric rules kept a smaller gap to True-Prior TS over the plotted horizon. Taken together, the two LMM conditions show the intended trade-off: parametric sharing can be the more efficient choice when its assumptions hold, and flexible sharing degrades less when they do not.

7 Conclusion

We introduced COCO-BANDITS, a hierarchical nonparametric framework for multi-task contextual bandits with partial context. Its central design choice separates population-level uncertainty, which is shared across users, from user-level uncertainty, which stays personalised. The experiments support a conditional message. A structured parametric model is the more efficient choice where its assumptions hold. Where the population is multimodal or carries heterogeneous residual variation, the flexible population model keeps a smaller gap to the True-Prior TS reference.

Scalability. The joint reward–context grid is the main practical bottleneck. With $G \asymp n_\mu^K n_x^{d_{\text{obs}}}$ grid points, a full Gaussian-process construction costs $\mathcal{O}(G^3)$ time and $\mathcal{O}(G^2)$ memory. A Nyström representation with $M_{\text{ind}} \ll G$ inducing points reduces this to approximately $\mathcal{O}(GM_{\text{ind}}^2 + M_{\text{ind}}^3)$ time and $\mathcal{O}(GM_{\text{ind}} + M_{\text{ind}}^2)$ memory. Maintaining J Karhunen–Loève terms and S SMC particles adds roughly $\mathcal{O}(SJG)$ work per update. These approximations make the reported small-action experiments feasible, but they do not remove the exponential dependence of a naive joint grid on the number of arms and on the observed-context dimension.

Scope of the results. Theorem 1 covers NPM-TS under the sequential alignment construction, and it is conditional on the local stability condition of Appendix Assumption 6. It therefore isolates the statistical cost of learning the population prior. Extending it to GIDS, to the grid, Karhunen–Loève, Nyström, and finite-particle implementation, and to a comparison against algorithms that share no information is beyond the scope of this paper. The reported experiments use three arms and low-dimensional observed context, so the empirical claims are restricted to that regime.

Future work. Scaling to larger action and context spaces calls for sparse or low-rank density representations, partial factorisations across arms, and an analysis of approximation error under adaptive feedback. A second direction is to characterise when the added flexibility of a nonparametric population model repays its computational cost. Appendix Sections G.1 and I give a longer discussion.

Acknowledgments

This work was funded by the Ministerium für Wissenschaft und Gesundheit des Landes Rheinland-Pfalz through the project *AI4Nof1* under grant 724-0019#2024/0001-150115404.

References

Rabaa Alabdulrahman, Herna Viktor, and Eric Paquet. Active learning and user segmentation for the cold-start problem in recommendation systems. In *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019) - KDIR*, pages 113–123. INSTICC, SciTePress, 2019. ISBN 978-989-758-382-7. doi: 10.5220/0008162901130123.

Imad Aouali, Branislav Kveton, and Sumeet Katariya. Mixed-effect Thompson sampling. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2087–2115. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/aouali23a.html.

Maryam Aziz, Emilie Kaufmann, and Marie-Karelle Riviere. On multi-armed bandit designs for dose-finding trials. *Journal of Machine Learning Research*, 22(14):1–38, 2021.

Yikun Ban, Yunzhe Qi, and Jingrui He. Neural contextual bandits for personalized recommendation. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1246–1249, 2024a.

Yikun Ban, Yunzhe Qi, Tianxin Wei, Lihui Liu, and Jingrui He. Meta clustering of neural bandits. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 95–106, 2024b.

Hamsa Bastani, David Simchi-Levi, and Ruihao Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 68(3):1865–1881, 2022.

Soumya Basu, Branislav Kveton, Manzil Zaheer, and Csaba Szepesvári. No regrets for learning the prior in bandits. In *Advances in Neural Information Processing Systems*, 2021.

Homanga Bharadhwaj. Meta-learning for user cold-start recommendation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2019. doi: 10.1109/IJCNN.2019.8852100.

Craig Boutilier, Chih-wei Hsu, Branislav Kveton, Martin Mladenov, Csaba Szepesvari, and Manzil Zaheer. Differentiable meta-learning of bandit policies. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2122–2134. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/171ae1bbb81475eb96287dd78565b38b-Paper.pdf.

Leonardo Cella, Alessandro Lazaric, and Massimiliano Pontil. Meta-learning with stochastic linear bandits. In *International Conference on Machine Learning*, pages 1360–1370. PMLR, 2020.

Yutian Chen, Aja Huang, Ziyu Wang, Ioannis Antonoglou, Julian Schrittwieser, David Silver, and Nando de Freitas. Bayesian optimization in AlphaGo, 2018. URL https://arxiv.org/abs/1812.06855.

Nicolas Chopin and Omiros Papaspiliopoulos. *An introduction to sequential Monte Carlo*, volume 4. Springer, 2020.

Aniket Anand Deshmukh, Urun Dogan, and Clay Scott. Multi-task learning for contextual bandits. In *Advances in Neural Information Processing Systems*, pages 4848–4856, 2017.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Machine Learning for Healthcare Conference*, pages 67–82. PMLR, 2018.

Dylan J Foster, Claudio Gentile, Mehryar Mohri, and Julian Zimmert. Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems*, 33, 2020.

Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR, 2014.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

F Maxwell Harper and Joseph A Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):1–19, 2015.

Joey Hong, Branislav Kveton, Manzil Zaheer, and Mohammad Ghavamzadeh. Hierarchical Bayesian bandits. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7724–7741. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/hong22c.html.

Chih-Wei Hsu, Branislav Kveton, Ofer Meshi, Martin Mladenov, and Csaba Szepesvari. Empirical Bayes regret minimization. *arXiv preprint arXiv:1904.02664*, 2019.

Ruiquan Huang, Weiqiang Wu, Jing Yang, and Cong Shen. Federated linear contextual bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27057–27068. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Easton Huch, Jieru Shi, Madeline R Abbott, Jessica Golbus, Alexander Moreno, and Walter Dempsey. RoME: A robust mixed-effects bandit algorithm for optimizing mobile health interventions. *Advances in Neural Information Processing Systems*, 37:128280–128329, 2024.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Sanath Kumar Krishnamurthy, Vitor Hadad, and Susan Athey. Adapting to misspecification in contextual bandits with offline regression oracles. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5805–5814. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/krishnamurthy21a.html.

Branislav Kveton, Mikhail Konobeev, Manzil Zaheer, Chih-Wei Hsu, Martin Mladenov, Craig Boutilier, and Csaba Szepesvari. Meta-Thompson sampling. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5884–5893. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/kveton21a.html.

Branislav Kveton, Martin Mladenov, Chih-Wei Hsu, Manzil Zaheer, Csaba Szepesvari, and Craig Boutilier. Meta-learning bandit policies by gradient ascent, 2021b.

Claire Lacour. Rates of convergence for nonparametric deconvolution. *Comptes Rendus Mathematique*, 342(11):877–882, 2006.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

Shuai Li, Alexandros Karatzoglou, and Claudio Gentile. Collaborative filtering bandits. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–548, 2016.

Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 485–492. JMLR Workshop and Conference Proceedings, 2010.

Pedro A Ortega, Jane X Wang, Mark Rowland, Tim Genewein, Zeb Kurth-Nelson, Razvan Pascanu, Nicolas Heess, Joel Veness, Alex Pritzel, Pablo Sprechmann, Siddhant M Jayakumar, Tom McGrath, Kevin Miller, Mohammad Azar, Ian Osband, Neil Rabinowitz, András György, Silvia Chiappa, Simon Osindero, Yee Whye Teh, Hado van Hasselt, Nando de Freitas, Matthew Botvinick, and Shane Legg. Meta-learning of sequential strategies. *arXiv preprint arXiv:1905.03030*, 2019.

Hongju Park. *Partially Observable Stochastic Contextual Bandits*. PhD thesis, University of Georgia, 2024.

Hongju Park and Mohamad Kazem Shirani Faradonbeh. Analysis of Thompson sampling for partially observable contextual multi-armed bandits. *IEEE Control Systems Letters*, 6:2150–2155, 2022. doi: 10.1109/LCSYS.2021.3137269.

Michael Pearce, Janis Klaise, and Matthew Groves. Practical Bayesian optimization of objectives with conditioning variables, 2020. URL https://arxiv.org/abs/2002.09996.

Amit Peleg, Naama Pearl, and Ron Meir. Metalearning linear bandits by prior update. In *International Conference on Artificial Intelligence and Statistics*, pages 2885–2926. PMLR, 2022.

Yuan Qi, Bo Dai, and Yao Zhu. EigenGP: Sparse Gaussian process models with data-dependent eigenfunctions. *arXiv preprint arXiv:1204.3972*, 2012.

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

Sulthana Shams, Daron Anderson, and Douglas Leith. Cluster-based bandits: Fast cold-start for recommender system new users. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1613–1616, 2021.

Chengshuai Shi and Cong Shen. Federated multi-armed bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9603–9611, 2021.

Sama Shrestha and Sonia Jain. A Bayesian-bandit adaptive design for N-of-1 clinical trials. *Statistics in Medicine*, 40(7):1825–1844, 2021.

Nicollas Silva, Thiago Silva, Heitor Werneck, Leonardo Rocha, and Adriano Pereira. User cold-start problem in multi-armed bandits: when the first recommendations guide the user's experience. *ACM Transactions on Recommender Systems*, 1(1):1–24, 2023.

Max Simchowitz, Christopher Tosh, Akshay Krishnamurthy, Daniel J Hsu, Thodoris Lykouris, Miro Dudik, and Robert E Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 34:26382–26394, 2021.

Aad W Van der Vaart and J Harry Van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Runzhe Wan, Lin Ge, and Rui Song. Metadata-based multi-task bandits with Bayesian hierarchical models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29655–29668. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f7cfdde9db36af8e0d9a6d123d5c385e-Paper.pdf.

Huazheng Wang, Qingyun Wu, and Hongning Wang. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pages 1633–1642, 2016.

Lu Wang, Chengyu Wang, Keqiang Wang, and Xiaofeng He. Biucb: A contextual bandit algorithm for cold-start and diversified recommendation. In *2017 IEEE International Conference on Big Knowledge (ICBK)*, pages 248–253. IEEE, 2017.

Zhi Wang, Chicheng Zhang, Manish Kumar Singh, Laurel Riek, and Kamalika Chaudhuri. Multitask bandit learning through heterogeneous feedback aggregation. In *International Conference on Artificial Intelligence and Statistics*, pages 1531–1539. PMLR, 2021.

Zhiyong Wang, Jize Xie, Xutong Liu, Shuai Li, and John Lui. Online clustering of bandits with misspecified user models. *Advances in Neural Information Processing Systems*, 36:3785–3818, 2023.

Qingyun Wu, Huazheng Wang, Quanquan Gu, and Hongning Wang. Contextual bandits in a collaborative environment. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 529–538, 2016.

Jiaqi Yang, Wei Hu, Jason D. Lee, and Simon S. Du. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2021.

Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.

Qian Zhou, XiaoFang Zhang, Jin Xu, and Bin Liang. Large-scale bandit approaches for recommender systems. In *International Conference on Neural Information Processing*, pages 811–821. Springer, 2017.

Shuang Zhou, Debdeep Pati, Tianying Wang, Yun Yang, and Raymond J Carroll. Gaussian processes with errors in variables: Theory and computation. *Journal of Machine Learning Research*, 24(87): 1–53, 2023.

Appendix

A Notation and Abbreviations

A.1 Core Notation

Convention. P denotes a probability distribution and p its density. Untilded Q_t and $\pi_t^{i,f}$ are the Bayesian quantities of Section 3; tilded quantities are the finite grid, KL, and SMC approximations used by the deployed algorithm.

Symbol	Meaning
Environment and interaction protocol	
$i, t, i(t), h_{i,t}$	User index, global interaction index, user served at t, and number of interactions with user i by time t.
T, b, U_t, I_t	Global interaction budget, incoming batch size, all arrived users, and active users.
$K, T_{\mathrm{user}}, H = T'_{\mathrm{user}} = T_{\mathrm{user}} - K$	Number of arms, full user horizon, and post-alignment horizon. In the proof construction, alignment lasts K pulls (one per arm).
$\boldsymbol{x}_i = (\mathbf{x}_{i,\mathrm{obs}}, \mathbf{x}_{i,\mathrm{unobs}})$	Full context; only $\mathbf{x}_{i,\mathrm{obs}}$ is observed by the learner.
$\boldsymbol{\mu}_i, r_t, \epsilon_t, \sigma$	Latent arm-mean vector, observed reward, observation noise, and its standard deviation.
$\mathcal{H}_{i,t}, \mathcal{H}_t, \mathcal{D}_t$	User history, global interaction history, and all observed contexts and histories.
$P^\star, p^\star$	Fixed true population distribution and its observed-data density $p^\star(\boldsymbol{\mu}, \mathbf{x}_{\mathrm{obs}})$, after marginalising $\mathbf{x}_{\mathrm{unobs}}$.
Model and posteriors	
$f, m, k, Z_f, p^{(f)}, P^{(f)}$	GP log-potential, GP mean function, kernel, normaliser, induced density, and induced distribution.
$Q_0(df), Q_t(df)$	Meta-prior and posterior measure over population models.
$\pi_t^{i,f}, \pi_t^i$	Personal posterior conditional on f, and its mixture over Q_t.
Numerical approximation	
$G, J, S, M_{\mathrm{ind}}$	Joint grid size, retained KL rank, number of SMC particles, and Nyström inducing-point count.
$\widetilde{Q}_t, \widetilde{\pi}_t^{i,f}, w_t^{(s)}$	Particle approximation to Q_t, discrete personal posterior, and particle weights; $\widetilde{Q}_t = \sum_{s=1}^{S} w_t^{(s)} \delta_{f_t^{(s)}}$.
$\widetilde{L}_t^{(s)}, \mathcal{R}, \epsilon$	Incremental particle predictive likelihood, numerical reward quadrature for GIDS, and its positive numerical stabiliser.
$\widetilde{I}_t^i, \widetilde{\Delta}_t^{i,f}, \widetilde{\rho}_t^{i,f}$	Approximate structural information gain, expected instantaneous regret, and GIDS action distribution.
ESS, η, τ	Effective sample size, Langevin step size, and SMC resampling threshold.
Performance criteria and analysis	
$V_{\mathcal{A}}, \mathrm{BR}_{\mathcal{A}}, \mathrm{MTR}_{\mathcal{A}}$	Expected latent-mean value, Bayes regret, and learned-prior versus True-Prior TS gap.
$\mathrm{CMTR}, \mathrm{CBR}$	Empirical estimates of cumulative multi-task and Bayes regret.
$N, n, \ell, \mathcal{A}_n$	Number of sequential proof tasks, preceding aligned tasks, proof burn-in index, and proof-only aligned-data set.
$\pi_i^\star, \hat{\pi}_{n,i}, \hat{P}_n, \hat{p}_n, \hat{p}_n^{\mathrm{approx}}$	True task-conditional prior, learned prior and distribution after n aligned tasks, learned density, and its numerical approximation.
$B, p_{\min}, \varepsilon_n, C_{\mathrm{ctr}}, c_1, c_2$	Latent-mean bound, observed-context marginal-density lower bound, contraction rate, contraction constant, and tail constants.
$h, \mathrm{TV}, \mathcal{G}_n, \mathcal{B}_n, \mathcal{C}_n$	Hellinger distance, total variation, and proof good/bad/contraction events.

Table 2: Notation used in the main paper and the appendix.

A.2 Abbreviations

Abbreviation	Meaning
MAB	Multi-armed bandit
GP	Gaussian process
KL	Karhunen–Loève
SMC	Sequential Monte Carlo
MCMC	Markov chain Monte Carlo
ESS	Effective sample size
IDS	Information-directed sampling
GIDS	Global information-directed sampling
TS	Thompson sampling
TP-TS	True-Prior Thompson sampling
NPM-TS	Nonparametric meta-prior Thompson sampling
Ind-TS	Independent Thompson sampling
MTB	Metadata-based bandit baseline
LMM	Linear mixed model
MTTS	Multi-task Thompson-sampling baseline
CMTR	Cumulative multi-task regret
CBR	Cumulative Bayes regret
RKHS	Reproducing kernel Hilbert space

Table 3: Abbreviations used throughout the paper and appendix.

B Extended Related Work

The multi-armed bandit (MAB) framework formalises sequential decisions under uncertainty. The learner balances exploration against exploitation in order to minimise cumulative regret. The contextual variant adds side information, such as user or item features, and conditions the action choice on it [Li et al., 2010]. Contextual models generalise across contexts, which is why they are used for personalised recommendation and adaptive decision systems. This section groups prior work by the structure it shares across tasks and by the way it treats missing context.

Contextual Bandits and Partial Observability. Contextual bandits have been studied under linear [Lu et al., 2010] and nonlinear [Zhou et al., 2020, Ban et al., 2024a] reward assumptions. Linear models admit regret guarantees and interpretable parameters. Neural methods scale to high-dimensional contexts. Both families usually assume realisability, that is, the observed context determines the expected reward [Foster et al., 2020, Krishnamurthy et al., 2021]. Our model instead retains randomness in the reward after conditioning on the observed metadata, which is the property that the partial-context setting of Section 2 requires.

A second group of methods studies partial-contextual settings, in which only some features are observable, for example user features without item data. These methods infer the missing information from statistical patterns shared across tasks [Wang et al., 2016, Park and Shirani Faradonbeh, 2022, Park, 2024]. We share that starting point and model the task-specific reward parameters as latent variables conditioned on the observed metadata. The distinction is that we place no parametric family on the conditional distribution of those latent variables.

Meta and Multi-Task Bandits. Meta-learning in bandits transfers knowledge across tasks and adapts from limited per-task data [Hsu et al., 2019, Boutilier et al., 2020, Kveton et al., 2021a,b]. META-TS [Kveton et al., 2021a] assumes that reward vectors are drawn i.i.d. from a shared distribution, and estimates a prior from the observed data. Most meta-bandit methods assume a metadata-agnostic task distribution and therefore discard task metadata. Our method conditions the shared distribution on task metadata and admits arbitrary task arrival orders.

Hierarchical bandits are closely related: a shared global prior governs the task-level reward functions [Basu et al., 2021]. These models adapt from few observations per task. Our nonparametric

Bayesian model extends this line by learning the prior from data conditional on metadata, without assuming a fixed degree of inter-task similarity. Wan et al. [2021] consider context-dependent meta-priors, but assume knowledge of the latent variables, of the functional relationship between context and reward, and of a parametric family for the latent-variable distribution. Hong et al. [2022] propose a hierarchical Bayesian model that unifies meta, multi-task, and federated bandits, but their framework still requires the analyst to specify the prior distribution. The unresolved distinction for the present work is that both require a parametric specification of the shared structure, whereas the partial-context setting leaves that structure unknown.

Federated, Collaborative, and Structured Bandits. Federated and multi-agent bandits study decentralised settings in which agents, such as users or devices, act independently and synchronise occasionally [Wu et al., 2016, Huang et al., 2021, Shi and Shen, 2021, Huch et al., 2024]. These methods usually assume clustered or homogeneous agent populations, and coordinate learning through shared priors or latent factor models. Our model makes no clustering assumption; the relationships between tasks follow from the learned metadata-conditional population density.

Collaborative-filtering bandits treat rewards as a user–item matrix and exploit low-rank or clustered structure [Li et al., 2016]. Clustering methods such as the Gang of Bandits [Gentile et al., 2014, Ban et al., 2024b] partition users into latent groups and pool observations within a group. Both work when a clustering approximation is valid. Later work relaxes that assumption by modelling probabilistic task similarity or user misspecification [Wang et al., 2023]. Our population density represents a clustered population as a multimodal conditional distribution, and admits populations for which no clustering holds.

A further strand decomposes the reward function into a global shared component and a local task-specific component, following mixed-effects models [Yang et al., 2021, Huch et al., 2024]. The additive decomposition captures both population trends and per-task deviations. Our model represents the same dependence without fixing the additive form, and the LMM conditions in Appendix Section F.3 measure what that costs when the additive form is correct.

Cold-Start and Bayesian Optimisation Connections. Cold-start settings are the case in which new users or items have little history. Prior work addresses them with side information [Wang et al., 2017] , with meta-learned priors [Bharadhwaj, 2019], or with population-level summaries [Alabdulrahman et al., 2019, Shams et al., 2021]. We learn the prior directly from task metadata, and the hierarchy of Section 3.1 keeps the population uncertainty separate from the user uncertainty that dominates in the cold-start regime.

Bayesian optimisation has been extended to multi-task and meta-learning settings through conditional formulations that share information across tasks with surrogate models [Chen et al., 2018, Pearce et al., 2020]. Our method uses the same hierarchical Bayesian device for transfer across tasks, applied to a bandit rather than an optimisation objective.

All of these settings use shared structure to reduce the cost of exploration. They differ in what they assume that structure to be. Our contribution is to keep the shared structure nonparametric, to condition it on the metadata that is actually observed, and to keep the reward vector latent.

C Population Model

This section expands the population model and hierarchical posterior of Section 3. It introduces no additional environment assumptions: the target is the observed-data distribution of $(\boldsymbol{\mu}, \mathbf{x}_{\text{obs}})$, with the unobserved context already marginalised as in (1).

C.1 Gaussian Processes as Distributions over Functions

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact input domain. A GP defines a distribution over real-valued functions $f : \mathcal{X} \to \mathbb{R}$, such that for any finite collection of input points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the associated function values $(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n$ follow a multivariate Gaussian distribution:

$$f \sim \mathcal{GP}(m, k) \quad \Longleftrightarrow \quad (f(x_1), \ldots, f(x_n)) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}),$$

where the mean vector $\boldsymbol{m} \in \mathbb{R}^n$ and covariance matrix $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ are given by

$$\boldsymbol{m}_i = m(x_i), \qquad \boldsymbol{K}_{ij} = k(x_i, x_j).$$

The mean function $m : \mathcal{X} \to \mathbb{R}$ and the kernel (covariance) function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ fully specify the GP. The kernel function must be symmetric and positive semi-definite, ensuring that for any finite set of inputs, the corresponding covariance matrix $\boldsymbol{K}$ is positive semi-definite.

The choice of kernel controls the regularity properties of the sample paths of the GP. For example, the squared exponential kernel,

$$k_{\text{SE}}(x, x') = \sigma_f^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right),$$

yields smooth, infinitely differentiable functions. Matérn kernels allow more control over function roughness, depending on a smoothness parameter ν. For most standard kernels, sample paths are almost surely continuous.

Conceptually, GPs define probability measures over function spaces. While the reproducing kernel Hilbert space (RKHS) associated with k describes a subspace of functions with bounded norm, the support of the GP measure extends beyond the RKHS, including more general continuous functions.

C.2 From Function Priors to Log-Densities

To define a flexible prior over probability densities on a space $\mathcal{Z} = \mathbb{R}^K \times \mathbb{R}^{d_{\text{obs}}}$, we place a GP prior on the logarithm of the density. Specifically, we let

$$f \sim \mathcal{GP}(m, k), \quad f : \mathcal{Z} \to \mathbb{R},$$

and define a density $p^{(f)}$ on $\mathcal{Z}$ via

$$p^{(f)}(z) = \frac{\exp(f(z))}{Z_f}, \qquad Z_f = \int_{\mathcal{Z}} \exp(f(z))\, dz.$$

The density $p^{(f)}$ induces the distribution $P^{(f)}$; the normalisation constant Z_f ensures that $p^{(f)}$ integrates to one. For common kernels and compact domains, the integral Z_f exists with high probability under the GP prior. When the domain is non-compact, it is often sufficient to ensure integrability by adding a negative quadratic drift or restricting f to decay appropriately at infinity.

This transformation of GP sample paths into densities via exponentiation and normalisation ensures that $p^{(f)}$ is strictly positive and integrates to one. Since GPs generate continuous sample paths almost surely, the resulting densities are smooth and capable of modelling complex dependencies.

C.3 Hierarchical Posterior

The GP meta-prior is the measure $Q_0(df)$. Each draw $f \sim Q_0$ defines a joint distribution $P^{(f)}$, and posterior inference yields the meta-posterior $Q_t(df) = Q_0(df \mid \mathcal{D}_t)$. As in (4), the observed-data likelihood includes the context factor when a user arrives and the reward-history likelihood for every user in U_t, including completed users. Conditional on a fixed f, this induces the personalised posterior $p^{(f)}(\boldsymbol{\mu} \mid \mathbf{x}_{i,\text{obs}}, \mathcal{H}_{i,t}) = \pi_t^{i,f}(\boldsymbol{\mu})$ from (5). Thus, the population and user-specific uncertainty layers remain distinct.

C.4 Properties of the Log-Gaussian-Process Prior

Modelling the log-density with a Gaussian process performs density estimation without a parametric family for the density itself. The exponential-and-normalise transformation guarantees that every draw is a valid density, and the kernel expresses the inductive bias. Two comparisons are relevant to the present setting. Against kernel density estimators, the log-Gaussian-process model admits structured priors and returns a posterior rather than a point estimate, which the hierarchy of Section 3.1 requires. Against parametric latent-variable models, it represents multimodal and nonlinear conditional structure without committing to a latent-variable family; the LMM conditions in Appendix Section F.3 measure the statistical price of that generality when a parametric family is in fact correct.

D Inference Details

This section expands the approximation and deployment layer in Section 4. Grid discretisation, KL truncation, SMC, and Nyström compression are numerical approximations of Q_t and $\pi_t^{i,f}$; the

statistical model is the one defined in Section 3. An end-to-end bound on the numerical error is beyond the scope of this paper.

D.1 Sources of Intractability

A Gaussian process gives a nonparametric prior over log-densities, but three computations block its direct use. This subsection states each of them.

First, to define a density from a draw $f \sim \mathcal{GP}(m, k)$, one must compute the normalisation constant

$$Z_f = \int_{\mathcal{Z}} \exp(f(z))\, dz,$$

This requires evaluating the function f over a discretised grid covering $\mathcal{Z}$. In practice, we approximate Z_f using numerical quadrature or Monte Carlo integration, which involves computing $f(z_j)$ for a dense set of grid points $\{z_j\}_{j=1}^{G}$. This step alone incurs $\mathcal{O}(G)$ cost per sample, where G grows exponentially with the input dimension due to the curse of dimensionality.

To obtain such evaluations, one must first draw a sample $f \sim \mathcal{GP}(m, k)$. Since GPs are infinite-dimensional stochastic processes, sampling a full function is only possible on a finite set of evaluation points. Let $\mathcal{Z}_G = \{z_j\}_{j=1}^{G} \subset \mathcal{Z}$ denote a grid of inputs. The restriction of f to $\mathcal{Z}_G$ defines a multivariate Gaussian distribution:

$$(f(z_1), \ldots, f(z_G))^\top \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}),$$

with $\boldsymbol{m}_j = m(z_j)$ and $\boldsymbol{K}_{ij} = k(z_i, z_j)$. Sampling from this distribution requires drawing a vector from a Gaussian with covariance matrix $\boldsymbol{K} \in \mathbb{R}^{G \times G}$, which costs $\mathcal{O}(G^3)$ for Cholesky decomposition and $\mathcal{O}(G^2)$ for each sample.

Therefore, for each function sample:

- Sampling the function over G grid points requires $\mathcal{O}(G^3)$ preprocessing (or lower cost with structured kernels or inducing points),
- Evaluating $\exp(f(z))$ over all $z \in \mathcal{Z}_G$ incurs $\mathcal{O}(G)$ cost,
- Approximating the integral Z_f adds another $\mathcal{O}(G)$,
- This must be repeated for every model draw under the posterior $Q_t(df)$ during inference or planning.

More fundamentally, Bayesian inference requires marginalising over an infinite-dimensional posterior measure. In the notation of the main paper,

$$Q_t(df) \propto p^{(f)}(\mathcal{D}_t)\, Q_0(df),$$

and predictions require evaluating expectations of the form

$$\mathbb{E}_{f \sim Q_t}\left[\Phi(f)\right],$$

for functionals $\Phi(f)$ such as expected rewards or optimal arm probabilities.

Since this outer integral is over an infinite-dimensional space of functions, exact inference is intractable. The combination of:

1. integration over $\mathcal{Z}$ for each function sample to evaluate likelihoods,
2. integration over function space for Bayesian inference,

renders exact computation computationally prohibitive, particularly in online or sequential settings such as bandits.

To overcome these issues, we employ a truncated Karhunen–Loève expansion of the GP prior, projecting f into a low-dimensional linear span of orthonormal eigenfunctions:

$$f(z) \approx \sum_{j=1}^{J} \theta_j \phi_j(z), \quad \theta_j \sim \mathcal{N}(0, \lambda_j),$$

where $\{(\lambda_\ell, \phi_\ell)\}$ are the leading eigenpairs of the covariance operator induced by k. Although this approximation still requires evaluating the expansion on the discretised grid over the support of $\boldsymbol{\mu}$, it is substantially more tractable than updating the full GP posterior, whose computational cost also grows with the number of observations and would quickly become infeasible in a streaming patient setting. We therefore perform posterior inference over the finite-dimensional coefficient vector $\boldsymbol{\theta} \in \mathbb{R}^J$ using Sequential Monte Carlo (SMC), which allows efficient sampling and updating over time while preserving much of the expressive modelling capacity of the full GP prior. Further implementation details are provided in Appendix D.3.

D.2 Karhunen–Loève Expansion and Particle Approximation

To make inference over Gaussian Process-defined log-densities computationally feasible, we adopt a finite-dimensional approximation based on the Karhunen–Loève (KL) expansion. This enables us to project the infinite-dimensional function space of the GP onto a lower-dimensional basis of orthonormal eigenfunctions, and reduces the problem of sampling functions to sampling standard Gaussian vectors in Euclidean space.

Formally, given a zero-mean Gaussian Process $f \sim \mathcal{GP}(0, k)$ defined over a compact domain $\mathcal{Z} \subset \mathbb{R}^{K+d_{\text{obs}}}$, the KL expansion of f takes the form:

$$f(z) = \sum_{m=1}^{\infty} \sqrt{\lambda_m}\, \xi_m\, \phi_m(z),$$

where $\{\phi_m\}_{m=1}^{\infty}$ are the orthonormal eigenfunctions of the covariance operator associated with k, and $\{\lambda_m\}_{m=1}^{\infty}$ are the corresponding eigenvalues. The coefficients $\{\xi_m\}$ are independent standard normal random variables, $\xi_m \sim \mathcal{N}(0, 1)$.

In practice, we truncate this expansion to the top J components:

$$f(z) \approx \sum_{j=1}^{J} \sqrt{\lambda_j}\, \xi_j\, \phi_j(z),$$

yielding a finite-dimensional approximation of the function space. The GP prior thus induces a distribution over the coefficients $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_J)$, which now parameterise the function f.

To numerically construct the empirical KL basis, we discretise $\mathcal{Z}$ with a grid of points $\{z_g = (\boldsymbol{\mu}_g, \boldsymbol{x}_g)\}_{g=1}^{G}$, and compute the kernel matrix

$$K_{gg'} = k(z_g, z_{g'}).$$

Performing eigendecomposition $K = \Phi \Lambda \Phi^\top$ gives us the top J eigenvalues $\lambda_1, \ldots, \lambda_J$ and the corresponding empirical eigenvectors $\phi_j \in \mathbb{R}^G$, which approximate the continuous eigenfunctions at the grid points.

Each function sample $f^{(s)}$ can then be constructed by drawing a particle $\boldsymbol{\xi}^{(s)} \sim \mathcal{N}(0, I_J)$ and computing:

$$f^{(s)}(z) = \sum_{j=1}^{J} \sqrt{\lambda_j}\, \xi_j^{(s)}\, \phi_j(z).$$

This representation allows us to sample functions from the GP by sampling low-dimensional Gaussian vectors. Since all dependence on z is encoded in the basis functions $\phi_m(z)$, this approximation decouples function complexity from inference dimensionality.

Particle Approximation and Inference. Using this KL representation, the implementation maintains S weighted coefficient particles $\{\boldsymbol{\xi}^{(s)}, w_t^{(s)}\}_{s=1}^{S}$, each corresponding to a log-density sample $f_t^{(s)}$. They form the empirical meta-posterior

$$\widetilde{Q}_t = \sum_{s=1}^{S} w_t^{(s)} \delta_{f_t^{(s)}} \approx Q_t.$$

The density induced by one particle is obtained by exponentiating and normalising:

$$p^{(f_t^{(s)})}(z) = \frac{\exp(f_t^{(s)}(z))}{\int \exp(f_t^{(s)}(z'))dz'}.$$

These densities are used to compute predictive likelihoods, update particle weights, and propagate $\widetilde{Q}_t$ over time using Sequential Monte Carlo (SMC). The particle-based KL approximation therefore implements the posterior of Section 3.2.

Solving the Inverse Problem. In the context of density estimation, the fundamental inverse problem is to infer a function f such that $\exp(f(z))/Z_f$ best explains the observed data under a generative model. The KL expansion transforms this into a search over a finite-dimensional coefficient space $\mathbb{R}^J$, making posterior inference feasible. By optimising or sampling over $\boldsymbol{\xi} \in \mathbb{R}^J$, we indirectly explore the space of plausible log-densities, while avoiding the intractability of directly estimating infinite-dimensional functionals.

Computational Efficiency. Let G denote the number of grid points in the joint reward–context discretisation, J the number of retained KL basis functions, and S the number of SMC particles. Without low-rank structure, GP kernel construction and eigendecomposition cost $\mathcal{O}(G^3)$ time. With a Nyström approximation using M_{ind} inducing points, these costs reduce to approximately $\mathcal{O}(GM_{\text{ind}}^2 + M_{\text{ind}}^3)$. Evaluating one truncated KL sample over the grid costs $\mathcal{O}(JG)$, so maintaining S particles adds roughly $\mathcal{O}(SJG)$ work per sequential update. This makes online particle filtering feasible in moderately high-dimensional spaces while making explicit the approximation trade-offs described in the main paper.

The reported experiments retain $J = 80$ Karhunen–Loève components, and the ablation in Appendix Section G.1 varies J over $\{40, 80, 120\}$. Under the squared-exponential kernel used here, the eigenvalues decay fast enough that this range covers the functional variation the population model needs; the ablation reports how regret changes across it.

D.3 Sequential Monte Carlo for the Population Posterior

The deployed algorithm approximates the population posterior with SMC. It uses the same observed data $\mathcal{D}_t$ as the posterior in (4): each observed context contributes once at arrival, and reward data from both active and completed users remain in the population posterior. This is separate from the personalised posterior used to serve one current user.

Particle representation. Following Chopin and Papaspiliopoulos [2020], SMC approximates Q_t by $\widetilde{Q}_t = \sum_{s=1}^{S} w_t^{(s)} \delta_{f_t^{(s)}}$, where each $f_t^{(s)}$ is represented by its J-dimensional KL coefficient vector and $w_t^{(s)}$ is an importance weight. The weighted particles are updated as data arrive, then resampled and rejuvenated when necessary.

KL representation. Using the Karhunen–Loève (KL) expansion described in Appendix D.2, we express each GP sample as

$$f^{(s)}(z) = \sum_{j=1}^{J} \sqrt{\lambda_j}\, \xi_j^{(s)}\, \phi_j(z),$$

where $\boldsymbol{\xi}^{(s)} \in \mathbb{R}^J$ are standard Gaussian coefficients.

Incremental population update. When a new user i arrives, particle s first receives the context factor $p^{(f_t^{(s)})}(\mathbf{x}_{i,\text{obs}})$. After action a_t produces reward r_t, it receives the incremental predictive likelihood

$$\widetilde{L}_t^{(s)} := \sum_{\ell=1}^{G} \mathcal{N}\big(r_t; \mu_{\ell,a_t}, \sigma^2\big)\, \widetilde{\pi}_{t-1}^{i,f_t^{(s)}}(\boldsymbol{\mu}_\ell).$$

Thus $w_t^{(s)} \propto w_{t-1}^{(s)} \widetilde{L}_t^{(s)}$, followed by normalisation. The discrete personalised posterior $\widetilde{\pi}_t^{i,f_t^{(s)}}$ is defined below. This incremental update accounts for each reward once rather than repeatedly multiplying a user's complete history into an already updated particle weight.

Resampling. As more data accumulates, particle degeneracy can occur, where most particles have negligible weight. To address this, we monitor the effective sample size (ESS),

$$\text{ESS} = \left(\sum_{s=1}^{S} (w_t^{(s)})^2 \right)^{-1},$$

and trigger a resampling step if ESS $< \tau S$, for some threshold $\tau \in (0, 1)$. Resampling replicates high-weight particles and discards low-weight ones, ensuring that computational effort is focused on promising regions of the parameter space.

Rejuvenation via Langevin dynamics. Resampling creates duplicate particles and reduces diversity. We therefore apply an MCMC rejuvenation step. Instead of a random-walk proposal, we use a Langevin proposal, which uses the gradient of the log-target:

$$\theta' = \theta + \frac{\eta}{2} \nabla_\theta \log \varrho_t(\theta) + \sqrt{\eta}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$

where θ denotes the KL coefficients, $\varrho_t(\theta) \propto p(\theta) p(\mathcal{D}_t \mid \theta)$ is the target density induced by all observed data, and η is a step size. The proposal is accepted or rejected using a Metropolis–Hastings rule to correct for discretisation error. This Langevin-based rejuvenation allows efficient exploration of the finite-dimensional approximation to the population posterior.

Tempered transitions. Multimodal or sharply peaked posteriors can collapse the particle system in a single update. We therefore interpose a sequence of intermediate targets indexed by a temperature parameter $\beta \in [0, 1]$:

$$\varrho_\beta(\theta) \propto p(\theta) \cdot p(\mathcal{D}_t \mid \theta)^\beta.$$

Increasing β from 0 to 1 moves the particles from the prior to the full posterior in steps. This *tempering* scheme reduces the chance of early particle collapse.

Why SMC over repeated MCMC. The generative model $f \sim \mathcal{GP}(0, k)$ is fixed, but its posterior changes at every interaction. Re-running a batch MCMC sampler after every event repeats work already done. SMC instead transports one weighted approximation through the sequence of observed-data likelihood factors, and resamples and rejuvenates only when the weights degenerate. This is what makes the update usable online, at the same statistical model.

Computational complexity. The main computational costs in SMC arise from three steps:

- **Likelihood evaluation:** For each particle $\theta^{(s)}$, evaluating $f^{(s)}(z)$ over a grid of size G costs $\mathcal{O}(JG)$.
- **Weight update:** Computing log-likelihoods for new observations across S particles takes $\mathcal{O}(SJG)$ per time step.
- **Langevin MCMC:** Each rejuvenation step requires computing gradients and log-densities, typically $\mathcal{O}(JG)$ per particle per step.

The initial eigendecomposition of the kernel matrix over the grid (to construct KL components) has one-time cost $\mathcal{O}(G^3)$, but is shared across all users and reused throughout.

In summary, the SMC stage maintains the weighted particle approximation $\widetilde{Q}_t$ of the population posterior, updates it by one incremental predictive likelihood per observed reward, resamples when the effective sample size falls below τS, and restores particle diversity with tempered Langevin rejuvenation. It approximates the posterior of Section 3.2; an accuracy guarantee for the approximation is beyond the scope of this paper.

D.4 Discrete Personalised Posterior and Predictive Likelihood

For a fixed particle $f_t^{(s)}$, the population weight $w_t^{(s)}$ and a user's posterior over means play different roles. The former represents uncertainty about the shared population model; the latter conditions that

Algorithm 2: Incremental SMC update used by Algorithm 1

Require: $\widetilde{Q}_{t-1} = \sum_{s=1}^{S} w_{t-1}^{(s)} \delta_{f_{t-1}^{(s)}}$, event $(i, \mathbf{x}_{i,\text{obs}}, a_t, r_t)$, history $\mathcal{H}_{i,t-1}$, grid $\{\boldsymbol{\mu}_\ell\}_{\ell=1}^{G}$, and threshold τ

if i has just arrived **then**
  for $s = 1, \dots, S$ **do**
    $w_{t-1}^{(s)} \leftarrow w_{t-1}^{(s)} p^{(f_{t-1}^{(s)})}(\mathbf{x}_{i,\text{obs}})$
  end for
  Normalize the particle weights
end if
for $s = 1, \dots, S$ **do**
  Form $\widetilde{\pi}_{t-1}^{i, f_{t-1}^{(s)}}$ from (D.4)
  $\widetilde{L}_t^{(s)} \leftarrow \sum_{\ell=1}^{G} \mathcal{N}(r_t; \mu_{\ell,a_t}, \sigma^2) \widetilde{\pi}_{t-1}^{i, f_{t-1}^{(s)}}(\boldsymbol{\mu}_\ell)$
  $w_t^{(s)} \leftarrow w_{t-1}^{(s)} \widetilde{L}_t^{(s)}$
end for
Normalize the particle weights and set $\widetilde{Q}_t \leftarrow \sum_{s=1}^{S} w_t^{(s)} \delta_{f_t^{(s)}}$
if $\mathrm{ESS}(\widetilde{Q}_t) < \tau S$ **then**
  Resample particles, reset their weights to $1/S$, and rejuvenate their KL coefficients with a target based on $\mathcal{D}_t$
end if
return $\widetilde{Q}_t$

model on the current user's context and reward history. Let $\{\boldsymbol{\mu}_\ell\}_{\ell=1}^{G}$ be the mean-vector grid. For user i, define

$$q_{i,t,\ell}^{(s)} := \exp\big(f_t^{(s)}(\boldsymbol{\mu}_\ell, \mathbf{x}_{i,\text{obs}})\big) \prod_{(a,r)\in\mathcal{H}_{i,t}} \mathcal{N}\big(r; \mu_{\ell,a}, \sigma^2\big), \qquad \widetilde{\pi}_t^{i, f_t^{(s)}}(\boldsymbol{\mu}_\ell) := \frac{q_{i,t,\ell}^{(s)}}{\sum_{\ell'=1}^{G} q_{i,t,\ell'}^{(s)}}.$$

This is the grid counterpart of the personalised posterior in Section 3.2. In particular, it uses the user's complete available history, whereas the SMC update above uses only the one newly observed reward through the marginal likelihood $\widetilde{L}_t^{(s)}$.

Link to the global posterior. For a new user, the context factor is included once in the global particle update. Conditional on a particle, the predictive probability of a new reward is the grid average in $\widetilde{L}_t^{(s)}$; this is the quantity that updates $w_t^{(s)}$. The global approximation therefore retains evidence from all users, including users who have completed their own horizons, while $\widetilde{\pi}_t^{i, f_t^{(s)}}$ is used only to choose actions for user i.

D.5 Implementation of NPM-TS

This is the numerical counterpart of the decision rule in Section 3.3. At a decision for user $i = i(t)$, draw

$$s_t \sim \mathrm{Categorical}\Big(w_{t-1}^{(1)}, \dots, w_{t-1}^{(S)}\Big), \qquad \widetilde{\boldsymbol{\mu}}_{i,t} \sim \widetilde{\pi}_{t-1}^{i, f_{t-1}^{(s_t)}}, \qquad a_t \in \arg\max_{a\in[K]} \widetilde{\mu}_{i,t,a}.$$

The discrete posterior in (D.4) conditions the selected population particle on both $\mathbf{x}_{i,\text{obs}}$ and $\mathcal{H}_{i,t-1}$. The Thompson action is therefore driven by a mean vector sampled from the user-conditional posterior. The finite grid and particle system approximate the draw $f_t \sim Q_{t-1}$, $\widetilde{\boldsymbol{\mu}}_{i,t} \sim \pi_{t-1}^{i,f_t}$ of Section 3.3, at the same modelling assumptions.

D.6 Implementation of GIDS

The deployed GIDS rule uses particle-weight entropy as a surrogate for the structural-information criterion in Section 3.3. A regret guarantee for GIDS, for the concurrent algorithm, and for the grid, KL, and SMC approximation is beyond the scope of this paper.

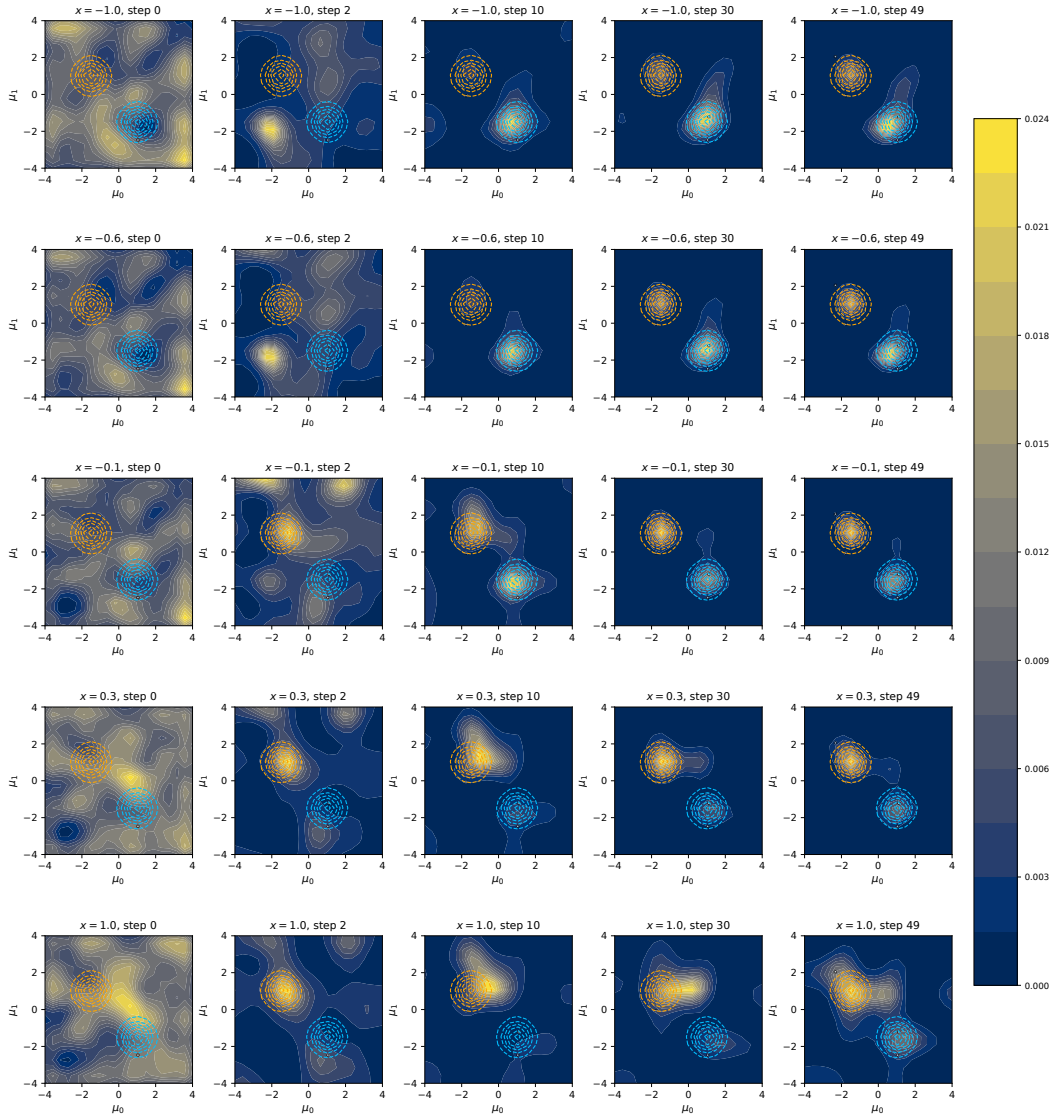


Figure 3: Sequential approximation of a context-dependent mixture of Gaussians using our KL-expanded SMC-GP framework. The underlying target is a mixture of two Gaussians with context-dependent weights: for each context $x \in [-1, 1]$, the mixture weight is given by $\sigma(x) = \frac{1}{1+\exp(-x)}$. We visualise the marginal distributions over the mean rewards μ_0 and μ_1 at five contexts and five time steps $t \in \{0, 2, 10, 30, 49\}$. At $t = 0$, particles are drawn from the GP prior; as data accumulates, the posterior increasingly concentrates around the true multimodal density.

Predictive reweighting. For each candidate arm a and candidate outcome r, compute each particle's user-specific predictive likelihood

$$\widetilde{L}_t^{(s)}(r \mid a) = \sum_{\ell=1}^{G} \widetilde{\pi}_{t-1}^{i,f_{t-1}^{(s)}}(\boldsymbol{\mu}_\ell) \mathcal{N}\big(r; \mu_{\ell,a}, \sigma^2\big)\,, \qquad \widetilde{w}_t^{(s)}(r, a) \propto w_{t-1}^{(s)} \widetilde{L}_t^{(s)}(r \mid a).$$

The entropy of the particle approximation is

$$\mathrm{Ent}(\widetilde{Q}_{t-1}) := -\sum_{s=1}^{S} w_{t-1}^{(s)} \log w_{t-1}^{(s)}.$$

Using a numerical reward quadrature $\mathcal{R}$, the corresponding estimated information gain is

$$\widetilde{I}_t^i(a) = \mathrm{Ent}(\widetilde{Q}_{t-1}) - \sum_{r\in\mathcal{R}} \widetilde{p}_t(r \mid a)\ \mathrm{Ent}\Big(\widetilde{Q}_t^{r,a}\Big)\,,$$

where $\widetilde{Q}_t^{r,a}$ uses the normalized weights $\widetilde{w}_t^{(s)}(r,a)$, and $\widetilde{p}_t(r \mid a)$ is the particle-mixture predictive probability.

Action distribution. For a particle drawn from $\widetilde{Q}_{t-1}$, let

$$\widetilde{\Delta}_t^{i,f}(a) = \sum_{\ell=1}^{G} \widetilde{\pi}_{t-1}^{i,f}(\boldsymbol{\mu}_\ell)\left(\max_{a'\in[K]} \mu_{\ell,a'} - \mu_{\ell,a}\right).$$

The implementation selects an action distribution

$$\widetilde{\rho}_t^{i,f} \in \arg\min_{\rho\in\Delta_K} \frac{\Big(\sum_{a=1}^K \rho_a \widetilde{\Delta}_t^{i,f}(a)\Big)^2}{\sum_{a=1}^K \rho_a \widetilde{I}_t^i(a) + \epsilon},$$

and samples $a_t \sim \widetilde{\rho}_t^{i,f}$. Tildes mark every quantity that is computed on the finite grid and particle system.

E Proofs for Section 5

This appendix gives the proof details for the result in Section 5. It analyses the *sequential* alignment construction of Section 5. Guarantees for GIDS, for grid, KL, SMC, and Nyström approximation error, and for fully adaptive concurrent feedback are beyond the scope of this paper.

Proof setting and alignment. There are N sequential tasks. Each task has the full horizon T_{user}. Its first K pulls visit all K arms once in a fixed order; this proof-only alignment phase yields the post-alignment horizon

$$H := T'_{\mathrm{user}} := T_{\mathrm{user}} - K.$$

The aligned data from preceding tasks construct the learned prior used by Thompson sampling in the remaining H rounds. Alignment is a device for the contraction argument; Algorithm 1 omits it.

Tasks have bounded mean vectors $\boldsymbol{\mu}_i \in [0,B]^K$ and observed contexts $\mathbf{x}_{i,\mathrm{obs}}$. They are i.i.d. from the observed-data population distribution $P^\star(d\boldsymbol{\mu}, d\mathbf{x}_{\mathrm{obs}})$; the unobserved context has already been marginalised as in (1). For task $i = n+1$,

$$\pi_i^\star(d\boldsymbol{\mu}) = P^\star(d\boldsymbol{\mu} \mid \mathbf{x}_{i,\mathrm{obs}}), \qquad \hat{\pi}_{n,i}(d\boldsymbol{\mu}) = \hat{P}_n(d\boldsymbol{\mu} \mid \mathbf{x}_{i,\mathrm{obs}})$$

denote the True-Prior TS conditional prior and the learned conditional prior after n aligned tasks.

Metric and proof idea. The per-task MTR is the latent-mean value gap between learned-prior TS and True-Prior TS, equivalently the corresponding Bayes-regret gap as defined in (3). The argument combines a Thompson-sampling prior-sensitivity bound [Simchowitz et al., 2021] with contraction of the log-GP posterior [Van der Vaart and Van Zanten, 2008]. The full statement retains the linear BKN alignment cost, which cancels against a True-Prior TS reference assigned the same forced phase. The post-alignment population-learning term is sublinear when the contraction rates sum sublinearly.

E.1 Assumptions and Proof Construction

Assumption 1 (Bounded Rewards)**.** *All mean rewards lie in* $[0,B]$.

Assumption 2 (Fixed Horizon)**.** *Each task has a fixed horizon* T_{user}.

Assumption 3 (i.i.d. Tasks)**.** *Tasks are drawn i.i.d. from* $P^\star$.

Assumption 4 (Log-GP Model)**.** *The joint density of* (μ, x_{obs}) *is modeled by a log-Gaussian process with squared exponential kernel, and* $\log p^\star$ *lies in its support.*

Algorithm 3: Sequential alignment construction for the proof

Input: meta-prior $Q_0(df)$; user horizon T_{user}; number of arms K; reward noise σ
Initialise: aligned-data set $\mathcal{A}_0 \leftarrow \emptyset$. Tasks arrive one at a time and each is served for exactly T_{user} rounds.

for $i = 1, \ldots, N$ **do**
  Observe $\mathbf{x}_{i,\text{obs}}$, set $\mathcal{H}_{i,0} \leftarrow \emptyset$, and construct $Q_{i-1}(df) := Q_0(df \mid \mathcal{A}_{i-1})$
  Form learned conditional prior $\hat{\pi}_{i-1,i}(d\boldsymbol{\mu}) := \int P^{(f)}(d\boldsymbol{\mu} \mid \mathbf{x}_{i,\text{obs}}) Q_{i-1}(df)$
  for $s = 1, \ldots, T_{\text{user}}$ **do**
    if $s \le K$ **then** ▷ forced alignment: every arm once
      $a_{i,s} \leftarrow s$
    else ▷ learned-prior Thompson sampling
      Draw $\widetilde{\boldsymbol{\mu}}_{i,s}$ from the posterior obtained from $\hat{\pi}_{i-1,i}$ and $\mathcal{H}_{i,s-1}$
      $a_{i,s} \in \arg\max_{a \in [K]} \widetilde{\mu}_{i,s,a}$
    end if
    Play $a_{i,s}$, observe $r_{i,s}$, and set $\mathcal{H}_{i,s} \leftarrow \mathcal{H}_{i,s-1} \cup \{(a_{i,s}, r_{i,s})\}$
  end for
  $\mathcal{A}_i \leftarrow \mathcal{A}_{i-1} \cup \{(\mathbf{x}_{i,\text{obs}}, \mathcal{H}_{i,K})\}$ ▷ use only aligned data for the next learned prior
end for

Assumption 5 (Context Regularity)**.** *There exists $p_{\min} > 0$ such that $p^\star(x_{\text{obs}}) \ge p_{\min}$.*

Lemma 1 (Single-Task Prior-Gap Regret Bound)**.** *Assume $\mu_{i,a} \in [0, B]$. Fix task $i = n+1$, condition on a common K-pull alignment history, and consider the remaining horizon $H = T_{\text{user}} - K$. Let $\pi_i^\star$ and $\hat{\pi}_{n,i}$ be the True-Prior and learned conditional priors. The post-alignment multi-task regret satisfies*

$$\begin{aligned} \text{MTR}_i &:= V_i(\text{TP-TS}) - V_i(\text{TS}(\hat{\pi}_{n,i})) \\ &= \text{BR}_i(\text{TS}(\hat{\pi}_{n,i})) - \text{BR}_i(\text{TP-TS}) \\ &\le 2BH^2 \,\text{TV}(\pi_i^\star, \hat{\pi}_{n,i})\,. \end{aligned} \tag{7}$$

Proof. The alignment history is common to the two compared TS rules, so it fixes the same conditioning information. Applying the prior-sensitivity bound of Simchowitz et al. [2021] to the remaining H-round task, with B-bounded means, yields (7). The equality follows from the latent-mean value and Bayes-regret definitions in Section 2.3. □

Remark 1 (Bayes Regret Gap)**.** *Since the optimal-arm term in Bayes regret is algorithm-independent, the learned-prior minus True-Prior Bayes-regret gap equals MTR:*

$$\text{BR}_{\text{TS}(\hat{\pi}_{n,i})} - \text{BR}_{\text{TP-TS}} = \text{MTR}_i. \tag{8}$$

Lemma 2 (Vacuity Threshold)**.** *The bound $2BH^2\epsilon_i$ is informative only when $\epsilon_i < 1/(2H)$. If $\epsilon_i \ge 1/(2H)$, it exceeds the trivial BH post-alignment bound and becomes vacuous.*

Contraction ingredient and conditioning scope. The next result applies to the observed-data density constructed from the aligned task data, and supplies the contraction ingredient of the proof. Contraction for the deployed particle approximation is beyond the scope of this paper.

Lemma 3 (Log-GP Posterior Contraction)**.** *Let $X_1, \ldots, X_n \sim p_0$. Let W be a zero-mean, tight Gaussian random element in $\ell^\infty(X)$ and define $p_w(x) := e^{w(x)} / \int_X e^{w(y)} d\nu(y)$. Assume $w_0 := \log p_0 \in \text{supp}(W)$. Let $\mathcal{H}$ be the RKHS of W and define the concentration function:*

$$\phi_{w_0}(\varepsilon) := \inf_{h \in \mathcal{H}: \|h - w_0\|_\infty < \varepsilon} \|h\|_{\mathcal{H}}^2 - \log \Pr(\|W\|_\infty < \varepsilon).$$

Let $\varepsilon_n \downarrow 0$ satisfy $\phi_{w_0}(\varepsilon_n) \le n\varepsilon_n^2$.

(A) Sieve and probability inequalities. *For any $C > 1$, there exists a measurable set $B_n \subset \ell^\infty(X)$ such that:*

$$\log N(3\varepsilon_n, B_n, \|\cdot\|_\infty) \le 6Cn\varepsilon_n^2, \tag{9}$$

$$\Pr(W \notin B_n) \le e^{-Cn\varepsilon_n^2}, \tag{10}$$

$$\Pr(\|W - w_0\|_\infty < 2\varepsilon_n) \ge e^{-n\varepsilon_n^2}. \tag{11}$$

(B) Posterior contraction. *For sufficiently large $C_{\mathrm{ctr}} > 0$,*

$$\mathbb{E}_{p_0}[\Pi_n(p_w : h(p_w, p_0) > C_{\mathrm{ctr}}\varepsilon_n \mid X_{1:n})] \to 0.$$

Remark 2 (How (A) implies (B))**.** *Part (A) is precisely Theorem 2.1 in Van der Vaart and Van Zanten [2008] applied with the Banach space $B = \ell^\infty(X)$ and norm $\|\cdot\|_\infty$ (equations* (9)*–*(11) *correspond to their bounds (2.2)–(2.4)). Part (B) is Theorem 3.1 in Van der Vaart and Van Zanten [2008], obtained by combining these sieve/probability bounds with a general posterior contraction theorem (their proof uses Ghosal et al. [2000, Thm. 2.1]) and the metric comparisons in their Lemma 3.1 that relate Hellinger/KL to $\|w - w_0\|_\infty$ for log-densities.*

Lemma 4 (Hellinger controls total variation)**.** *For any two probability measures P and Q,*

$$\mathrm{TV}(P, Q) \le h(P, Q).$$

Proof. This is the standard comparison between total variation and Hellinger distance. Writing densities p and q,

$$\tfrac{1}{2}\int |p - q| = \tfrac{1}{2}\int |\sqrt{p} - \sqrt{q}|(\sqrt{p} + \sqrt{q}),$$

and applying Cauchy–Schwarz gives

$$\mathrm{TV}(P, Q) \le \sqrt{\tfrac{1}{2}\int (\sqrt{p} - \sqrt{q})^2}\sqrt{\tfrac{1}{2}\int (\sqrt{p} + \sqrt{q})^2}.$$

Since $\frac{1}{2}\int(\sqrt{p} + \sqrt{q})^2 \le 2$, we obtain $\mathrm{TV}(P, Q) \le h(P, Q)$. □

Assumption 6 (Contextwise conditioning stability)**.** *For every observed context $\mathbf{x}_{obs}$ considered in the proof, $p^\star(\mathbf{x}_{obs}) \ge p_{\min} > 0$. On the contraction event for a learned observed-data density $\hat{p}_n$, assume the associated conditional prior satisfies*

$$\mathrm{TV}\Big(\hat{P}_n(d\boldsymbol{\mu} \mid \mathbf{x}_{obs}), P^\star(d\boldsymbol{\mu} \mid \mathbf{x}_{obs})\Big) \le \frac{4}{p_{\min}}\, h(\hat{p}_n, p^\star).$$

Remark 3 (Why the condition is stated locally)**.** *A global Hellinger or total-variation bound for joint densities controls their marginals in an integrated sense. Equation* (6) *states separately the pointwise conditional control that the proof uses at each context, and the theorem is conditional on it for the contexts under consideration. An expected-context variant, in which* (6) *is replaced by an integrated conditional-stability bound, is beyond the scope of this paper.*

Lemma 5 (Good-event MTR bound)**.** *Fix task $i = n + 1$ and its post-alignment horizon $H = T_{\mathrm{user}} - K$. Assume rewards lie in $[0, B]$, the contextwise stability condition in Assumption 6, and define*

$$\mathcal{G}_n(\mathbf{x}_{i,obs}) := \left\{ h(\hat{p}_n, p^\star) \le C_{\mathrm{ctr}}\varepsilon_n, \quad \frac{4C_{\mathrm{ctr}}}{p_{\min}}\varepsilon_n \le \frac{1}{2H} \right\}.$$

Then, on $\mathcal{G}_n(\mathbf{x}_{i,obs})$,

$$\begin{aligned} \mathrm{MTR}_i &\le 2BH^2\mathrm{TV}(\pi_i^\star, \hat{\pi}_{n,i}) \\ &\le \frac{8BH^2C_{\mathrm{ctr}}}{p_{\min}}\varepsilon_n. \end{aligned} \tag{12}$$

Consequently,

$$\mathbb{E}[\mathrm{MTR}_i \mathbf{1}\{\mathcal{G}_n(\mathbf{x}_{i,obs})\}] \le \frac{8BH^2C_{\mathrm{ctr}}}{p_{\min}}\varepsilon_n.$$

Proof. The second condition in $\mathcal{G}_n(\mathbf{x}_{i,\text{obs}})$ makes the prior-sensitivity bound non-vacuous by Lemma 2. Lemma 1, followed by Assumption 6, gives

$$\mathrm{TV}(\pi_i^\star, \hat{\pi}_{n,i}) \le \frac{4}{p_{\min}} h(\hat{p}_n, p^\star) \le \frac{4C_{\text{ctr}}}{p_{\min}} \varepsilon_n.$$

Substitution proves (12); the expectation statement follows because the displayed bound holds pointwise on the event. □

Lemma 6 (Contraction-tail contribution)**.** *Fix task $i = n+1$ and its post-alignment horizon H. Let*

$$\mathcal{B}_n(\mathbf{x}_{i,obs}) := \left\{ \mathrm{TV}(\pi_i^\star, \hat{\pi}_{n,i}) > \frac{1}{2H} \right\}, \qquad \mathcal{C}_n := \{ h(\hat{p}_n, p^\star) \le C_{\text{ctr}} \varepsilon_n \}.$$

If $\frac{4C_{\text{ctr}}}{p_{\min}} \varepsilon_n \le 1/(2H)$, Assumption 6 gives

$$\mathcal{B}_n(\mathbf{x}_{i,obs}) \subseteq \mathcal{C}_n^c.$$

Thus, if the contraction tail satisfies

$$\mathbb{P}(\mathcal{C}_n^c) \le c_1 e^{-c_2 n \varepsilon_n^2},$$

then the post-alignment contribution on the complement of the good event is bounded by

$$\mathbb{E}[\mathrm{MTR}_i \mathbf{1}\{\mathcal{C}_n^c\}] \le BH c_1 e^{-c_2 n \varepsilon_n^2}.$$

Proof. Since the post-alignment latent-mean value difference is at most BH, it remains only to prove the event inclusion. On $\mathcal{C}_n$, Assumption 6 gives

$$\mathrm{TV}(\pi_i^\star, \hat{\pi}_{n,i}) \le \frac{4C_{\text{ctr}}}{p_{\min}} \varepsilon_n \le \frac{1}{2H};$$

hence $\mathcal{B}_n(\mathbf{x}_{i,\text{obs}})$ cannot occur. The tail bound then yields the claimed expectation bound. □

E.2 Proof of the Main Multi-Task Regret Bound

The following restates and proves Theorem 1. Its scope is the sequential alignment construction above and the local conditioning condition in Assumption 6.

Theorem 2 (Cumulative MTR with explicit alignment cost)**.** *Assume bounded means $\mu_{i,a} \in [0, B]$, i.i.d. sequential tasks from $P^\star$, the forced K-pull alignment construction, log-GP contraction at rate ε_n, and Assumption 6. Let ℓ be a finite burn-in index such that, for all $n \ge \ell$,*

$$\frac{4C_{\text{ctr}}}{p_{\min}} \varepsilon_n \le \frac{1}{2H}, \qquad H = T_{\text{user}} - K,$$

and suppose the contraction-tail constants $c_1, c_2 > 0$ satisfy the bound in Lemma 6. Then

$$\sum_{i=1}^{N} \mathbb{E}[\mathrm{MTR}_i] \le BT_{\text{user}}\ell + BKN + \frac{8BH^2 C_{\text{ctr}}}{p_{\min}} \sum_{n=\ell}^{N-1} \varepsilon_n + BH \sum_{n=\ell}^{\infty} c_1 e^{-c_2 n \varepsilon_n^2}. \tag{13}$$

Proof. For the first ℓ tasks, use the full-horizon bound BT_{user} per task. For every task, the proof coupling charges at most BK for the forced alignment phase. For task $i = n+1$ with $n \ge \ell$, Lemma 5 bounds the post-alignment contribution on $\mathcal{C}_n$ by

$$\frac{8BH^2 C_{\text{ctr}}}{p_{\min}} \varepsilon_n,$$

and Lemma 6 bounds its complement by

$$BH c_1 e^{-c_2 n \varepsilon_n^2}.$$

Summing these terms over tasks gives (13). □

Interpretation. The expression contains the linear BKN alignment charge, which cancels once the True-Prior TS reference is given the same forced alignment phase. With fixed H, the remaining population-learning term is sublinear whenever $\sum_{n\le N}\varepsilon_n = o(N)$. As in the main text, extensions to GIDS, to concurrent deployment, and to numerical approximation error are beyond the scope of this paper.

E.3 Background: Contraction under Measurement Error

This background concerns an errors-in-variables observation model, and is included for reference. Theorem 2 and its proof are independent of it. Contraction for the noisy bandit-feedback and particle-based setting of this paper is beyond the scope of the section.

We work in an *errors-in-variables* (noise-in-variables) setting, where the latent density of interest is not observed directly but only through additive measurement noise. The result below follows by combining classical deconvolution theory for nonparametric density estimation in the super-smooth signal / super-smooth noise regime under measurement error [Lacour, 2006] with the Gaussian-process–based testing and posterior contraction machinery developed for models with errors in variables [Zhou et al., 2023]. In this setting, posterior contraction rates are governed by the existence of exponentially consistent tests in the convolution experiment, rather than by the direct-observation minimax benchmark.

Theorem 3 (Posterior contraction for super-smooth deconvolution with a log-GP prior)**.** *Let $W_1,\dots,W_n$ be i.i.d. noisy measurements generated by*

$$W_i = X_i + U_i, \qquad i = 1,\dots,n,$$

where X_i has unknown density p_0 on a compact support $\mathcal{X}$, and $U_i \sim \mathcal{N}(0,\sigma_\varepsilon^2)$ is independent additive noise. Write $f_0 = \log p_0$, and for any measurable function f define

$$p_f(x) = \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(u)}\,du}, \qquad q_f = p_f * \varphi_{\sigma_\varepsilon},$$

where $\varphi_{\sigma_\varepsilon}$ is the centered Gaussian density with variance σ_ε^2. Assume that:

1. *the true density p_0 is super-smooth as in Assumption 7;*
2. *p_0 is bounded away from zero and infinity as in Assumption 8;*
3. *the prior on f is a zero-mean Gaussian process with squared-exponential kernel C^ℓ, and f_0 lies in its support;*
4. *the testing bound of Lemma 8, the KL reduction of Lemma 9, and the prior-mass/sieve conditions of Lemma 10 hold.*

Let h_n be chosen so that the deconvolution bias and stochastic scales balance, $B_n \asymp V_n$, and define

$$\varepsilon_n \asymp B_n + V_n.$$

Then, for sufficiently large $M > 0$,

$$\Pi(\|f - f_0\|_1 > M\varepsilon_n \,|\, W_1,\dots,W_n) \to 0.$$

Equivalently, the posterior contracts around the true log-density at rate ε_n in L^1.

In the Gaussian-type super-smooth signal / Gaussian noise case, corresponding to $r = 2$, balancing B_n and V_n gives

$$h_n^2 \asymp \frac{2a + \sigma_\varepsilon^2}{\log n}, \qquad \varepsilon_n \asymp n^{-\frac{a}{2a+\sigma_\varepsilon^2}} (\log n)^{\frac{a-\delta\sigma_\varepsilon^2}{2(2a+\sigma_\varepsilon^2)}}.$$

Thus the posterior contracts at the polynomial rate

$$\Pi\left(\|f - f_0\|_1 > M\, n^{-\frac{a}{2a+\sigma_\varepsilon^2}} (\log n)^{\frac{a-\delta\sigma_\varepsilon^2}{2(2a+\sigma_\varepsilon^2)}} \,\middle|\, W_1,\dots,W_n \right) \to 0,$$

up to the usual asymptotic-equivalence constants. This rate is slower than in the noiseless setting, reflecting the ill-posedness introduced by super-smooth measurement error.

E.3.1 Deconvolutional Contraction Ingredients

To connect the generic log-GP contraction result above to the noisy-observation setting used in our theory discussion, we record the main ingredients of a testing-based deconvolution argument. We observe noisy measurements

$$W_i = X_i + U_i, \qquad i = 1, \dots, n,$$

where X_i has latent density p_0, the noise variables U_i are independent of X_i, and the observed density is the convolution

$$g_0 = p_0 * q.$$

For a generic log-density f, we define the induced latent density and observed density by

$$p_f(x) = \frac{e^{f(x)}}{\int_{\mathcal{X}} e^{f(u)}\, du}, \qquad q_f = p_f * q.$$

The goal is to prove posterior contraction for f around $f_0 = \log p_0$ from the noisy sample $W_{1:n}$.

Assumption 7 (Super-smooth signal)**.** *The true density p_0 is super-smooth in the sense that its Fourier transform satisfies*

$$\int_{\mathbb{R}} |p_0^*(t)|^2 (1+t^2)^{\delta} e^{2a|t|^r}\, dt \le L$$

for some constants $a, r, L > 0$ and $\delta \ge 0$.

Assumption 8 (Positivity and boundedness)**.** *There exist constants $0 < m_0 \le M_0 < \infty$ such that*

$$0 < m_0 \le p_0(x) \le M_0 < \infty$$

on the support $\mathcal{X}$.

Consider the standard Fourier deconvolution estimator

$$\hat{p}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K_n\left(\frac{x - W_i}{h_n}\right),$$

whose kernel is defined through the Fourier multiplier

$$K_n^*(t) = \frac{\mathbf{1}_{\{|t| \le 1\}}}{q^*(t/h_n)}.$$

Write

$$B_n := h_n^{\delta} e^{-a/h_n^r}, \qquad V_n := n^{-1/2} h_n^{-1/2} e^{\sigma_\varepsilon^2/(2h_n^2)}$$

for the bias and fluctuation scales in the Gaussian super-smooth case. Standard results in super-smooth deconvolution imply the following bounds [Lacour, 2006].

Lemma 7 (Bias and fluctuation scales)**.** *Under Assumption 7, the deconvolution estimator satisfies*

$$\|\mathbb{E}\hat{p}_n - p_0\|_1 \lesssim B_n, \qquad \mathbb{E}\,\|\hat{p}_n - \mathbb{E}\hat{p}_n\|_1 \lesssim V_n.$$

In particular, if the bandwidth is chosen so that $B_n \asymp V_n$, then the natural testing scale is $\varepsilon_n \asymp B_n + V_n$.

Lemma 8 (Testing concentration inequality)**.** *Suppose the deconvolution bandwidth is calibrated so that $B_n \asymp V_n$. Then there exists a sequence $\varepsilon_n \asymp B_n + V_n$ and constants $c, C > 0$ such that*

$$\mathbb{P}(\|\hat{p}_n - p_0\|_1 > C\varepsilon_n) \le e^{-cn\varepsilon_n^2}.$$

In particular, this concentration bound is obtained by applying McDiarmid's inequality to $\|\hat{p}_n - \mathbb{E}\hat{p}_n\|_1$, using the bounded-differences property induced by the deconvolution kernel. Thus one obtains exponentially consistent tests in the convolution experiment at scale ε_n.

Proof sketch. Write

$$F_n := \|\hat{p}_n - \mathbb{E}\hat{p}_n\|_1.$$

If only the ith observation is changed, the estimator changes by

$$\Delta_i(x) = \frac{1}{nh_n}\left\{K_n\left(\frac{x - W_i}{h_n}\right) - K_n\left(\frac{x - W_i'}{h_n}\right)\right\},$$

so

$$\|\Delta_i\|_1 \le \frac{2}{n}\|K_n\|_1.$$

Under the same kernel regularity used in the deconvolution argument from `contraction.tex`, $\|K_n\|_1$ is uniformly bounded, hence F_n satisfies a bounded-differences condition of order $1/n$. McDiarmid's inequality therefore gives

$$\mathbb{P}\big(F_n - \mathbb{E}F_n > t\big) \le e^{-c_1 n t^2}.$$

Combining this with Lemma 7 and choosing $t \asymp V_n$ yields

$$\mathbb{P}(\|\hat{p}_n - p_0\|_1 > C(B_n + V_n)) \le e^{-cn(B_n+V_n)^2},$$

and since $B_n \asymp V_n$, this gives the stated bound with $\varepsilon_n \asymp B_n + V_n$. □

Lemma 9 (KL reduction for the observed model)**.** *If $\|f - f_0\|_\infty \le \delta$, then the observed densities satisfy*

$$\mathrm{KL}(q_{f_0}, q_f) \lesssim \delta^2,$$

and the same control holds for the second KL moment up to constants.

Lemma 10 (Prior mass and sieve conditions for the log-GP prior)**.** *Let the log-density prior be $f \sim \mathcal{GP}(0, C^\ell)$ with concentration function ϕ_{f_0}. If $\phi_{f_0}(\varepsilon_n) \le n\varepsilon_n^2$, then the induced prior on q_f assigns at least $e^{-n\varepsilon_n^2}$ mass to a KL neighbourhood of q_{f_0}. Moreover, for a sieve of the form*

$$\mathcal{F}_n = M_n\mathbb{H}_1 + \varepsilon_n\mathbb{B}_1, \qquad M_n^2 \asymp n\varepsilon_n^2,$$

the prior mass outside $\mathcal{F}_n$ is bounded by $e^{-n\varepsilon_n^2}$.

Remark 4 (Contraction template in the deconvolution model)**.** *Lemmas 8–10 provide the standard ingredients required by the general posterior contraction theorem of Ghosal et al. [2000]: exponentially consistent tests, KL-neighbourhood prior mass, and a sieve with exponentially small complement probability. Together with the positivity assumption, which transfers density error to log-density error, these ingredients yield posterior contraction for the log-density in the noisy observation model. The theorem above specializes this template to the standalone super-smooth / super-smooth regime described in this optional background.*

Remark 5 (On minimax rates and the role of smoothness assumptions)**.** *The testing and contraction rates derived above reflect the intrinsic difficulty of deconvolution under Gaussian (super-smooth) measurement error [Lacour, 2006]. Under* more restrictive *assumptions, sharper guarantees are possible. For instance, in settings with repeated measurements, or when the noise distribution is only* ordinary-smooth*, classical results show that minimax-optimal (polynomial) rates can be achieved for density estimation and hypothesis testing.*

In contrast, when the underlying signal is only ordinary-smooth while the measurement error is super-smooth (e.g., Gaussian noise), the minimax rates deteriorate to logarithmic forms of the type $(\log n)^{-\beta}$. Such rates are known to converge extremely slowly, reflecting a fundamental ill-posedness of the inverse problem, and in this regime additional data sharing or information pooling across samples does not improve the asymptotic order of convergence.

In this optional model, we consider a different and less pessimistic regime. Specifically, we assume that both the signal and the measurement error are super-smooth*, with the signal modeled through a squared-exponential (SE) kernel prior and the noise following a Gaussian distribution. Under these assumptions, it is possible to establish polynomial testing and posterior contraction rates. Consequently, even under relatively mild smoothness assumptions on the signal, as encoded by the SE kernel prior, posterior consistency and convergence can be guaranteed despite the presence of super-smooth measurement error.*

F Experimental Details and Environments

All main-text experiments use the concurrent procedure from Section 4; the fixed-cohort LMM condition is the stated protocol-level exception. The environments below instantiate the population distribution and observation model from Section 2. Reported CMTR and CBR are empirical estimates of the latent-mean quantities in Section 2.3, and True-Prior TS is the evaluation reference of Section 2.3.

Environment	Observed / unobserved quantities	Arrival protocol	Main comparison
MoG	Observed context and a multimodal latent mean vector; no separate hidden-context component	Dynamic batches, five pulls per user	NPM-TS, GIDS, IND-TS, TP-TS
Partial context	x_o observed; x_c, x_d hidden; means depend on the full context	Dynamic batches, five pulls per user	Same nonparametric and independent rules
LMM (Gaussian / mixture)	Metadata observed; shared effect and user deviation are latent	Fixed cohort when Gaussian; dynamic batches when mixture	Adds MTTS under its compatible protocol

Table 4: How the appendix environments connect to the problem setting and main-text comparisons.

F.1 Environment 1: Context-Dependent Mixture of Gaussians

In the first synthetic environment the population reward distribution is a context-dependent mixture of Gaussians over the arm-mean vector. This environment isolates three properties: multimodality, latent subgroups, and the case in which an informative pull targets an arm that is not currently optimal.

Reward distribution. For each user with observed context $\boldsymbol{x}_{\mathrm{obs}}$, the latent arm-mean vector $\boldsymbol{\mu} \in \mathbb{R}^K$ is sampled from a finite mixture

$$P^\star(d\boldsymbol{\mu} \mid \boldsymbol{x}_{\mathrm{obs}}) = \sum_{j=1}^{J_{\mathrm{mix}}} v_j(\boldsymbol{x}_{\mathrm{obs}})\, \mathcal{N}(d\boldsymbol{\mu} \mid \boldsymbol{m}_j(\boldsymbol{x}_{\mathrm{obs}}), \Sigma_j)\,,$$

The mixing weights $v_j(\boldsymbol{x}_{\mathrm{obs}})$ depend smoothly on the observed context, and the component means induce different arm rankings. The population therefore contains subgroups with different optimal actions, and pulls that are informative about the subgroup but locally suboptimal.

Relation to the partial-context setting. This environment has observed context but no separately modelled $\mathbf{x}_{\text{unobs}}$: all residual heterogeneity is represented directly by the multimodal conditional distribution of $\boldsymbol{\mu} \mid \boldsymbol{x}_{\mathrm{obs}}$.

Observations. Conditional on the sampled mean vector, rewards are observed as Gaussian noise-corrupted samples

$$r_{i,t}(a) \sim \mathcal{N}(\mu_{i,a}, \sigma^2),$$

with fixed observation noise σ^2. In the experiments in the main text we use three arms and an episodic recruitment scheme, so that later users benefit from the structure learned from earlier users.

Recruitment protocol. Users arrive in batches of 5 for 30 recruitment rounds, and each recruited user interacts with the system for 5 decision steps before exiting. This setup mimics episodic deployments in real-world personalised recommendation or healthcare settings, where users enter in cohorts and interact over a short duration.

F.2 Environment 2: Partially Observed Linear Contextual Bandit

This environment evaluates the agent's ability to reason under uncertainty when only partial information about the user's context is available. The reward structure is linear in the full context vector, but only a single observed component is provided to the algorithm.

Context Structure. Each user is assigned a 4-dimensional context vector:

$$\boldsymbol{x} = [x_o,\ x_c,\ x_d,\ 1],$$

where:

- $x_o \sim \mathcal{U}(-1, 1)$ is an **observed** continuous feature,
- $x_c \sim \mathcal{N}(0.5, 0.1^2)$ is a **latent** continuous feature,
- $x_d \sim \text{Bernoulli}(\sigma(x_o))$ is a **latent** binary feature with probability controlled by the observed context,
- 1 is a fixed bias term.

Only the scalar x_o is revealed to the agent; the remaining dimensions of $\boldsymbol{x}$ are unobserved.

Reward Model. The mean reward for each arm $a \in [3]$ is a fixed linear function of the full context:

$$\mu_a = \boldsymbol{w}_a^\top \boldsymbol{x}, \qquad r_{i,t} \sim \mathcal{N}(\mu_a, \sigma^2),$$

where the noise variance is set to $\sigma^2 = 0.01$. The weight matrix is fixed as:

$$W = \begin{bmatrix} 0.6 & 0.1 & 1.0 & -0.9 \\ 0.3 & 0.3 & -1.0 & 0.9 \\ 0.1 & -0.2 & -0.3 & 0.1 \end{bmatrix}.$$

Learning Challenge. Since the agent observes only x_o, but the reward depends nontrivially on latent variables x_c and x_d, a purely linear policy over x_o is insufficient. Effective inference requires marginalizing over plausible latent configurations. This tests the model's ability to represent and reason over latent heterogeneity in the context.

Recruitment Protocol. We adopt the same episodic structure as in Environment F.1, with 30 rounds of recruitment, 5 new users per round, and each user interacting over 5 time steps. This episodic design simulates real-world deployments where users engage in short sessions.

F.3 Environment 3: Bayesian Linear Mixed Model (LMM)

This environment follows the Gaussian bandit setting proposed by Wan et al. [2021], using a Bayesian linear mixed model (LMM) with both fixed and random effects. It is used to evaluate model performance under both aligned and misaligned structural assumptions.

Context and Feature Map. Each user is associated with a scalar metadata feature $\boldsymbol{x}_i \sim \mathcal{U}(-1, 1)$. We define a fixed, known linear mapping $\phi(\boldsymbol{x}, a) \in \mathbb{R}^2$ from the context–arm pair to features:

$$\phi(\boldsymbol{x}, a) = [\rho_a \cdot \boldsymbol{x}, 1],$$

where the slope coefficients are fixed as:

$$\rho = [0.9,\ -1.1,\ 0.2] \quad \text{for arms } a = 1, 2, 3.$$

This yields an input dimension $d = 2$ per arm. For user i, the full arm-level design matrix is:

$$\Phi_i = \begin{bmatrix} 0.9 \cdot \boldsymbol{x}_i & 1 \\ -1.1 \cdot \boldsymbol{x}_i & 1 \\ 0.2 \cdot \boldsymbol{x}_i & 1 \end{bmatrix} \in \mathbb{R}^{3 \times 2}.$$

Fixed and Random Effects. Latent reward means for user i are given by:

$$\boldsymbol{\mu}_i = \Phi_i \theta + \delta_i,$$

where:

$$\theta \sim \mathcal{N}(\mu_\theta, \Sigma_\theta), \quad \mu_\theta = \mathbf{1}_2, \quad \Sigma_\theta = 0.25 \cdot I_2,$$

$$\delta_i \sim \mathcal{N}(0, \Sigma), \quad \Sigma = 0.05^2 \cdot I_3.$$

At the start of each simulation replicate, a single θ is drawn and shared by all users in that replicate. The fixed effect captures shared structure across users and arms, while δ_i models user-specific deviations not explained by the metadata.

Misaligned Setting. To measure the effect of misspecification, we replace the prior over θ by a bimodal Gaussian mixture:

$$\theta \sim \begin{cases} \mathcal{N}(-\mathbf{1}_2, 0.25 \cdot I_2), & \text{with probability } 0.5, \\ \mathcal{N}(+\mathbf{1}_2, 0.25 \cdot I_2), & \text{with probability } 0.5. \end{cases}$$

The population-level distribution of θ is then outside the single-Gaussian family that MTTS assumes, while it stays inside the support of the log-Gaussian-process prior.

Recruitment Schemes. We adopt distinct user recruitment protocols depending on model alignment:

- **Correctly Specified Setting:** All users are recruited simultaneously to match the batch training regime assumed by MTTS. We simulate 50 users, each interacting for 5 rounds. We report cumulative Bayes regret after one interaction round per user (i.e., 50 global time steps).
- **Misspecified Setting:** Users arrive sequentially in batches of 5 over 20 recruitment rounds (100 users total). Each user interacts for 5 rounds post-arrival. This tests adaptation as later users reveal a population that is not Gaussian under the MTTS random-effect specification.

F.4 Compute Resources and Runtime

All experiments were executed on a MacBook Air (2020) equipped with an Apple M1 chip and 8GB of unified memory. Code is implemented in Python using PyTorch and NumPy, with GPU acceleration. Each complete experiment for each random seed runs in under 60 minutes.

Random Seeds. Each experiment is repeated across fixed seeds for seed $\in \{0, 1, \ldots, 9\}$. These seeds are used to initialize environment generation, user recruitment, and baseline algorithms. For particle-based inference, additional random seeds are sampled during each round of Sequential Monte Carlo (SMC) updates to drive variability in resampling and rejuvenation steps.

F.5 Hyperparameters (All Models)

- Number of SMC particles: $S = 200$
- KL expansion rank: $J = 80$
- Nyström inducing-point count: $M_{\text{ind}} \ll G$, selected with the grid resolution to reduce the GP kernel cost
- Grid size: constructed via discretization
  - For each arm mean $\mu_a \in [-2, 2]$, 20 uniformly spaced grid points
  - For context $x \in [-1, 1]$, 10 uniformly spaced points
- GP kernel: squared exponential with lengthscale = 0.7 (tuned)
- GP mean function: constant zero
- Observation noise: $\sigma = 0.1$
- Langevin MCMC step size: $\eta = 0.05$
- Resampling schedule: triggered when $\text{ESS} < S/2$
- SMC update frequency: one incremental likelihood update per observed reward; optional rejuvenation follows resampling
- Annealing: adaptive inverse temperature schedule $\beta_t \in [0, 1]$ selected to maintain target effective sample size

KL Basis. The Karhunen–Loève basis vectors used in the truncated GP expansion are computed once at initialization and fixed throughout all experiments. No per-task or per-user recomputation is performed.

Baselines. All parametric baselines use closed-form posterior updates where applicable.

G Additional Experiments

G.1 Ablation Studies

To better understand which modelling and computational choices drive performance, we conduct a set of ablation studies centred on the context-dependent mixture-of-Gaussians environment from Appendix Section F.1. Unless otherwise stated, each ablation varies a single factor relative to the

default configuration in Appendix Section F.5, while keeping the reward-generation process and recruitment schedule fixed. We report both cumulative multi-task regret (CMTR) and cumulative Bayes regret (CBR), averaged over the same random seeds as in the main experiments.

Default ablation data-generating process. For the ablation studies only, we instantiate the Gaussian-mixture environment as a two-mode, three-arm contextual population bandit with one-dimensional observed context $x \in [-1, 1]$. Each arriving user has an unobserved reward-mean vector

$$\mu = (\mu_1, \mu_2, \mu_3) \in \mathbb{R}^3,$$

drawn from the context-dependent mixture

$$p(\mu \mid x) = \sigma(x)\mathcal{N}(\mu; m_1, 0.15I_3) + (1 - \sigma(x))\mathcal{N}(\mu; m_2, 0.15I_3),$$

where $\sigma(x) = (1 + \exp(-x))^{-1}$, and

$$m_1 = (1.5,\ 0.8,\ -1.0), \qquad m_2 = (0.8,\ 1.5,\ 0.2).$$

In the first mode arm 1 is optimal, while in the second mode arm 2 is optimal. Arm 3 is suboptimal in both modes but has different relative quality across the two populations. The observed context affects only the mixture weights over the latent modes, not the reward means directly, and since $x \in [-1, 1]$ both modes remain present throughout the context range, with the probability of mode 1 varying from approximately 0.27 to 0.73. For a selected arm a_t, rewards are generated as

$$r_t \mid \mu, a_t \sim \mathcal{N}(\mu_{a_t}, \sigma_r^2),$$

with default observation noise $\sigma_r = 0.1$. Unless explicitly stated otherwise, each ablation changes only the named parameter, while all remaining modelling, inference, and experimental settings are kept at their baseline values from Appendix Section F.5.

Ablation overview. The ablations show one recurring trade-off between representational capacity and statistical efficiency. A finer grid, a higher Karhunen–Loève rank, or a longer kernel lengthscale reduced regret when the corresponding modelling assumption matched the data-generating mechanism. Each of these also raised either the computational cost or the number of degrees of freedom that the learner had to resolve in the early rounds. Higher observation noise reduced performance through a different channel, by making each interaction less informative about the latent means. The two effects together indicate that the useful operating point has enough capacity to separate heterogeneous users and to normalise the induced densities, and no more.

G.1.1 Grid Resolution

We vary the discretisation used for the joint reward–context grid to study the bias–variance and accuracy–runtime trade-off induced by numerical approximation. Concretely, we compare coarse, medium, and fine grids by varying the number of points per arm-mean dimension and the number of observed-context grid points. Finer grids should reduce discretisation error and improve conditional posterior quality, but they also increase the cost of GP evaluation, KL projection, and particle updates.

Observations. Figures 4 and 9 show that regret decreased as the grid was refined. A finer discretisation resolves the context-conditional latent density more accurately, and therefore separates heterogeneous users more accurately. The grid also enters the estimate of the normalising constant and of the derived posterior quantities, so a coarse grid distorts the inference itself rather than adding an independent approximation error. Table 5 shows the cost: refining the grid from 10 to 25 points per dimension raised the wall-clock runtime from 15 minutes to 2 hours 13 minutes. The operating point is therefore the coarsest grid that still separates the heterogeneity present in the population.

Computational impact. Let G denote the total number of grid points, with $G \asymp n_\mu^K n_x^{d_{\mathrm{obs}}}$ when using n_μ points per reward dimension and n_x points per observed-context dimension. Increasing grid resolution therefore enlarges G exponentially in the number of modelled dimensions at fixed per-dimension refinement. Without Nyström compression, kernel construction and eigendecomposition scale as $\mathcal{O}(G^3)$ time and $\mathcal{O}(G^2)$ memory; with M_{ind} inducing points they scale as approximately $\mathcal{O}(GM_{\mathrm{ind}}^2 + M_{\mathrm{ind}}^3)$ time and $\mathcal{O}(GM_{\mathrm{ind}} + M_{\mathrm{ind}}^2)$ memory. The online KL/SMC stage then contributes about $\mathcal{O}(SJG)$ time per update and $\mathcal{O}(SG + SJ)$ memory. Thus this ablation is the most computationally expensive one: runtime is cubic in G without low-rank structure and linear in G in the particle-update term, while memory is quadratic in G without Nyström and linear in G with Nyström.

Ablation	Setting	Runtime
Grid	10	15m
Grid	20	1h 13m
Grid	25 (baseline)	2h 13m
KL terms	40	1h 50m
KL terms	80	1h 45m
KL terms	120 (baseline)	2h 13m
Noise	0.05	1h 28m
Noise	0.2	1h 30m
Noise	0.4	1h 30m
Noise	0.1 (baseline)	1h 29m
Lengthscale	0.3	1h 30m
Lengthscale	0.5	1h 30m
Lengthscale	1.0	1h 31m
Lengthscale	2.0	1h 44m
Lengthscale	0.7 (baseline)	1h 33m
Users/round	2	1h 57m
Users/round	10	2h 06m
Users/round	5 (baseline)	2h 13m

Table 5: Wall-clock runtimes for the Gaussian-mixture ablation studies. Baseline settings match the default ablation configuration used elsewhere in this section.

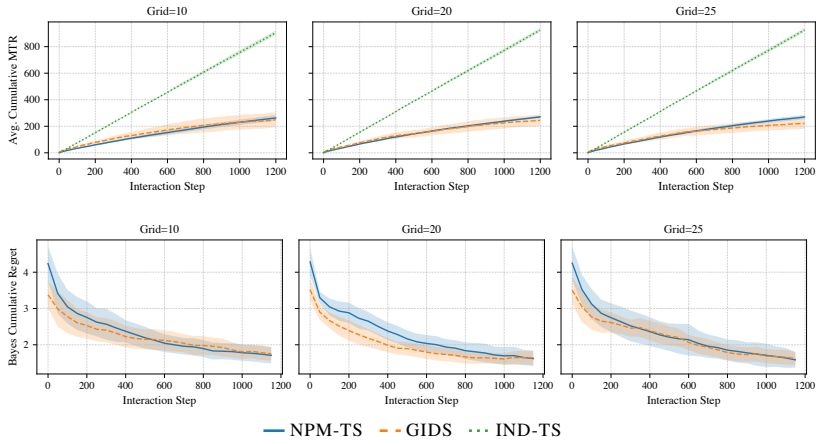


Figure 4: Grid-resolution ablation in the three-arm Gaussian-mixture environment. Top: trajectories of average cumulative multi-task regret. Bottom: trajectories of evaluation regret.

G.1.2 Number of KL Expansion Terms

We vary the truncation level of the Karhunen–Loève expansion to measure how sensitive performance is to the rank of the GP approximation. Small truncation levels may under-represent multimodal structure, while very large truncation levels may increase variance and runtime without materially improving regret. This study isolates whether the main gains come from coarse structural modelling or from a richer functional basis.

Observations. Figures 5 and 9 show that regret generally decreased as the number of retained Karhunen–Loève terms increased, because the retained basis represents the function space more accurately. The same figures show the opposite effect on GIDS in the early rounds. A larger truncation level raises the number of effective degrees of freedom that must be resolved, and GIDS allocates early pulls to resolving them. When the environment does not require that capacity, those pulls are spent for no later gain. The lengthscale ablation below shows the same mechanism from the prior side.

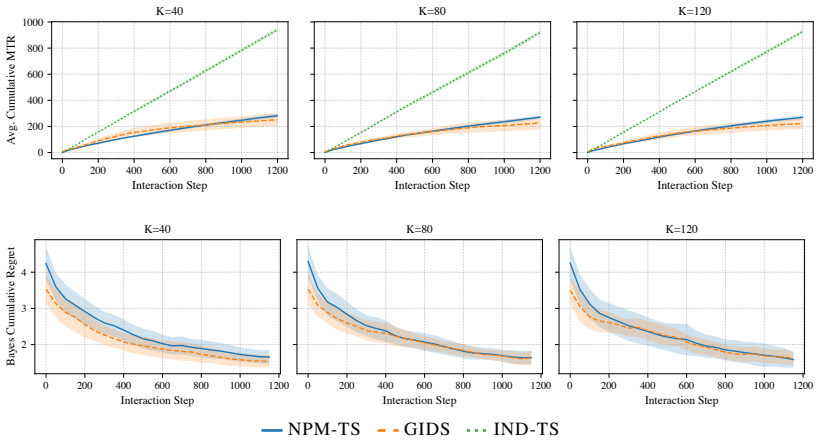


Figure 5: KL-expansion ablation in the three-arm Gaussian-mixture environment. Top: trajectories of average cumulative multi-task regret. Bottom: trajectories of evaluation regret.

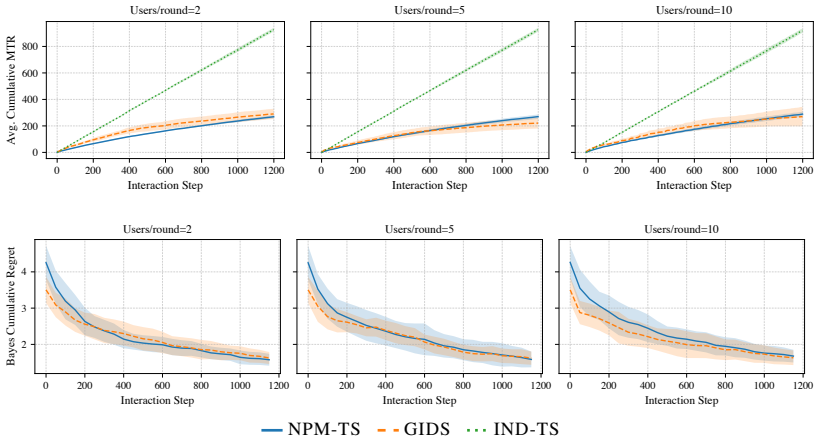


Figure 6: Number-of-tasks ablation in the three-arm Gaussian-mixture environment. Top: trajectories of average cumulative multi-task regret. Bottom: trajectories of evaluation regret.

Computational impact. Holding the grid fixed, increasing the truncation level J leaves the kernel construction cost unchanged but increases the cost of storing and propagating the retained basis functions. The dominant online cost grows approximately linearly as $\mathcal{O}(SJG)$ per sequential update, because each of the S particles must evaluate a J-term KL expansion over G grid points. Memory likewise increases roughly linearly in J, through storing KL coefficients and basis values, giving an additional $\mathcal{O}(JG + SJ)$ footprint. So this ablation is linear in J for both runtime and memory, conditional on the same grid and inducing-point budget.

G.1.3 Number of Tasks

To examine scaling with data availability, we vary the total number of recruited users while keeping the interaction horizon per user fixed. This ablation quantifies how quickly the learned meta-prior improves as more tasks are observed, and whether the benefit of global structure learning is concentrated in the early, medium, or late stages of recruitment.

Observations. Figures 6 and 9 vary how a fixed total number of users is recruited over time. Recruiting more users per round produced a faster decrease in evaluation regret for GIDS, because data from the first users enters the meta-prior that later users receive. The same mechanism acted on NPM-TS: the earliest users draw from a weak prior and incur high regret, and a larger early cohort

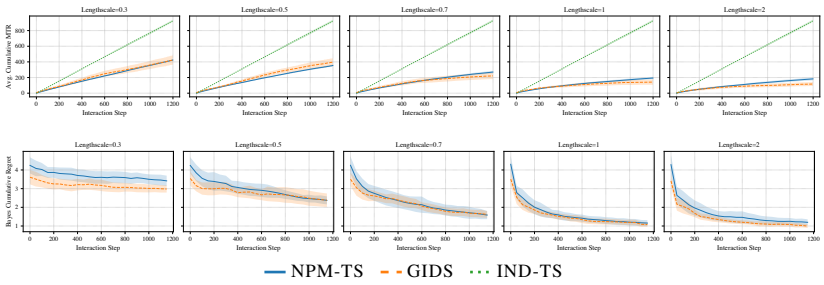


Figure 7: Kernel-lengthscale ablation in the three-arm Gaussian-mixture environment. Top: trajectories of average cumulative multi-task regret. Bottom: trajectories of evaluation regret.

replaces that prior sooner. Front-loading recruitment therefore improved the performance of later users and worsened the performance of the earliest ones.

Computational impact. If each task is served for horizon T_{user}, then increasing the number of tasks N scales the total experimental workload approximately linearly in the total number of interactions NT_{user}. For fixed G, J, S, and M_{ind}, the per-interaction cost is unchanged, so total runtime grows roughly as $\mathcal{O}(NT_{\text{user}}(SJG + \text{conditioning cost}))$ after initialization. Peak memory need not grow linearly with N if past rewards are retained through the posterior state and only active user histories are materialised; the dominant model-state cost is then $\mathcal{O}(GM_{\text{ind}} + M_{\text{ind}}^2 + JG + SJ + SG)$. Retaining all histories explicitly adds storage linear in the number of observed interactions.

G.1.4 Kernel Lengthscale

We vary the squared-exponential kernel lengthscale used in the meta-prior over latent reward densities to test how strongly performance depends on the smoothness bias imposed across the joint reward–context grid. Smaller lengthscales allow sharper local variation and may better capture abrupt changes in the context-conditional mixture weights, whereas larger lengthscales impose stronger sharing across neighbouring contexts and can oversmooth multimodal structure.

Observations. Figures 7 and 9 show that the kernel lengthscale had the largest effect of the modelling parameters varied here. Regret decreased as the lengthscale increased over the range 0.3 to 2.0. The true density in this environment is smooth, and a longer lengthscale concentrates prior mass on smoother functions, so the prior excludes explanations that the data would otherwise have to rule out, and shares information across neighbouring contexts. This result is specific to a smooth data-generating density; populations with sharp context-dependent transitions are beyond the scope of this ablation.

Computational impact. Changing the kernel lengthscale does not alter the asymptotic state size, since the grid, particle count, truncation level, and inducing-point budget remain fixed. The main effect is numerical rather than dimensional: extremely small lengthscales can lead to less correlated kernel matrices and therefore weaker low-rank compression, whereas very large lengthscales can make the representation overly smooth. For fixed G, J, S, and M_{ind}, the nominal runtime and memory orders remain the same as in the default configuration, with only constant-factor changes from conditioning and approximation quality.

G.1.5 Observation Noise

We vary the reward-noise standard deviation in the Gaussian observation model. Higher observation noise makes within-task inference less precise and reduces the information carried by each pull. Lower noise should sharpen posterior concentration and make cross-task transfer visible earlier in the recruitment process.

Observations. Figures 8 and 9 show that increasing the reward-noise standard deviation from 0.05 to 0.4 increased regret in every case. A noisier reward carries less information about the latent user-specific means, so more pulls are required before the posterior concentrates. Higher noise

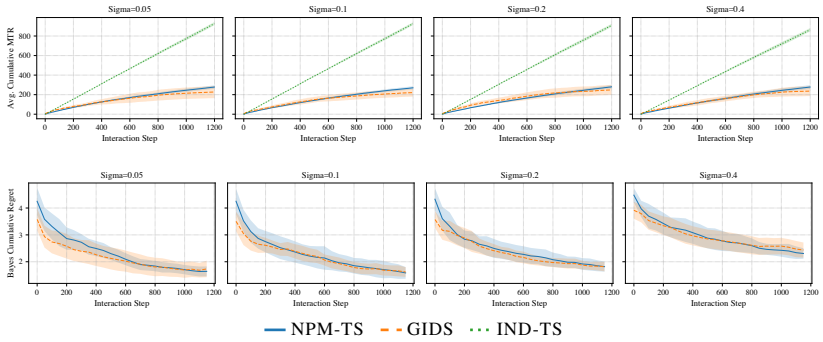


Figure 8: Observation-noise ablation in the three-arm Gaussian-mixture environment. Top: trajectories of average cumulative multi-task regret. Bottom: trajectories of evaluation regret.

therefore slows within-task learning and cross-task transfer at the same time, because the population update in (4) consumes the same observations.

Computational impact. There is no additional computation for different noise scales.

G.2 MovieLens Semi-Synthetic Environment

To complement the fully synthetic environments, we construct a semi-synthetic population bandit from the MovieLens 1M dataset [Harper and Konstan, 2015], available at https://grouplens.org/datasets/movielens/1m/. These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. The goal is to retain realistic user-level preference heterogeneity while preserving a controlled sequential recruitment process compatible with our population-bandit setting.

G.2.1 Construction

We restrict attention to users with sufficiently rich rating histories, keeping only users with at least 300 total ratings and at least 20 ratings in each selected arm. In the experiment reported, the three arms correspond to the genres COMEDY, ADVENTURE, and SCI-FI. We deliberately restrict the number of arms to three for computational reasons: because our nonparametric model operates over a discretised joint reward–context grid, adding more arms increases the dimensionality of that grid and the associated GP approximation cost very quickly. After filtering, the resulting pool contains 906 eligible users. Each user is treated as a task, and the observed context consists only of gender, encoded as a one-dimensional feature.

G.2.2 Reward Model

For each user i and genre arm a, we collect the set of historical ratings associated with that user–genre pair. The user-specific mean reward for arm a is the empirical mean of this pool, while the realized online reward is obtained by sampling a rating from the same pool with replacement. Thus, the stochasticity in the environment comes from the empirical within-user variability of MovieLens ratings rather than from an externally imposed noise model.

G.2.3 Population-Bandit Protocol

The full pool is split into a held-out evaluation set of 50 users and a recruitment pool of 856 users. The simulator then samples users from this recruitment pool over a global horizon of 60 steps. We begin with 5 users, recruit 5 new users at each subsequent step, and keep each recruited user active for 5 interactions before that user exits. In the reported run, this yields a total of 305 recruited users. The resulting sequential population process lets early users inform the meta-prior used for later users.

Evaluation is performed every 5 global steps on the fixed held-out set, using short Monte Carlo rollouts to estimate evaluation regret. This setup separates online training interactions from out-of-sample population evaluation and is intended to mimic a realistic recommendation setting in which

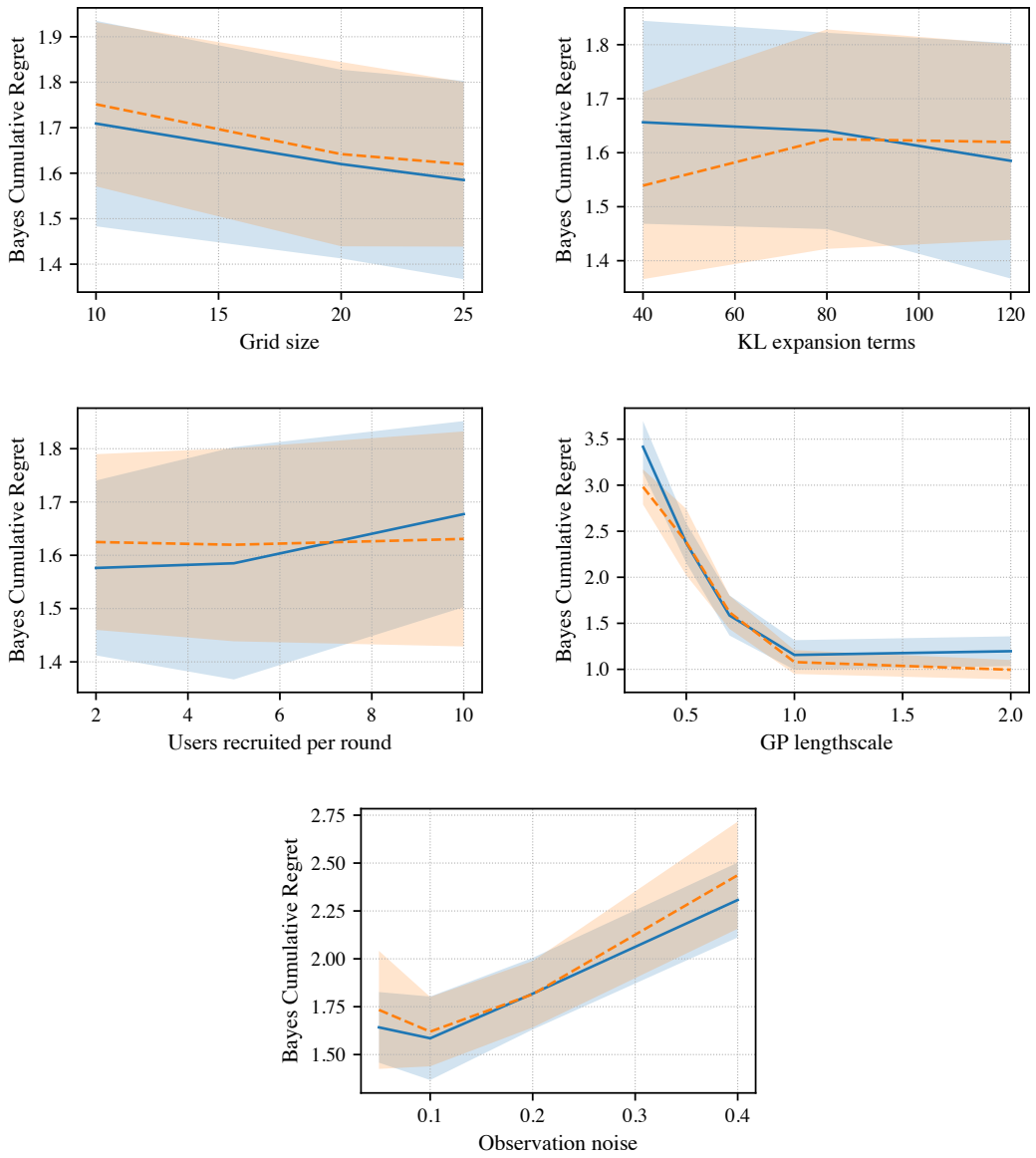


Figure 9: Final evaluation-regret comparisons for all Gaussian-mixture ablations. Top left: grid resolution. Top right: KL truncation level. Middle left: number of recruited tasks. Middle right: kernel lengthscale. Bottom: observation noise.

the algorithm must generalize from previously observed users to newly arriving users drawn from the same filtered MovieLens population.

Observations. In Figure 10, NPM-TS attained the lowest cumulative and held-out evaluation regret of the compared rules. The two parametric information-sharing baselines, Linear TS and MTB, attained the highest regret, which is consistent with a sharing structure too restrictive for the latent heterogeneity of the MovieLens population. The margin over independent Thompson sampling was small. The realised rewards in this environment are resampled from each user's own historical ratings, so the within-user variance is large and limits how much any transfer across users can gain. NPM-TS also attained lower regret than GIDS here, whereas the ordering was reversed in the MoG condition of the main text; the reward signal in MovieLens is noisier, which reduces the return on pulls made for global information. These are single-environment observations over the reported seeds; a significance test for the ordering is beyond the scope of this paper.

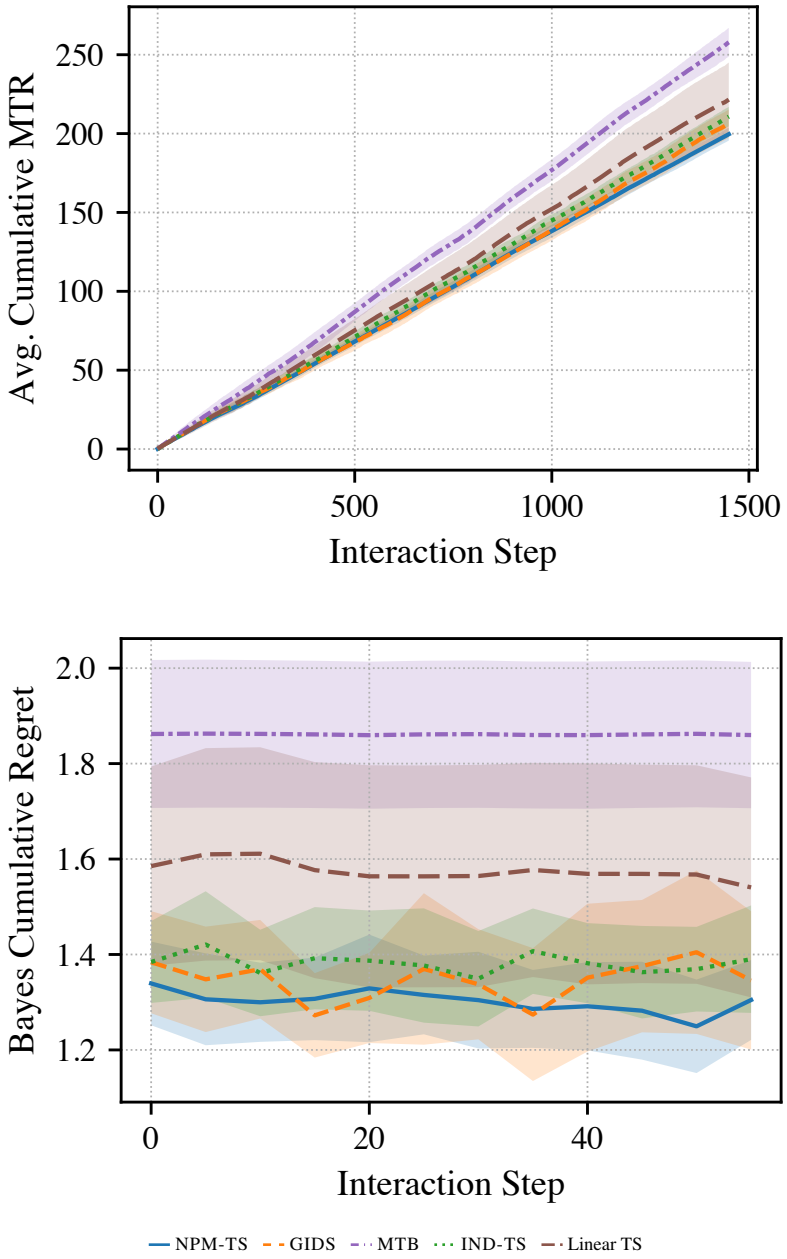


Figure 10: MovieLens semi-synthetic population-bandit results. Top: cumulative regret over online recruitment and interaction. Bottom: held-out evaluation regret measured on the fixed evaluation user set.

H Societal Impact Statement

This paper studies when learning a metadata-conditional population prior improves sequential decision-making across tasks. The work introduces no deployed system. It may nevertheless have downstream effects if it is used in recommendation, education, resource allocation, or clinical decision support. In those settings faster adaptation can help users, in particular in cold-start regimes. The same construction can also propagate historical bias present in the population data, produce uneven performance across subgroups, and raise privacy concerns through its use of metadata. Deployments therefore require careful feature selection, restricted use of sensitive attributes, subgroup auditing, and human oversight in high-stakes decisions.

I Extended Discussion and Future Work

Theorem 1 covers NPM-TS in the sequential aligned construction of Section 5, under posterior contraction and the local conditioning-stability condition of Assumption 6. Extending it to GIDS, to concurrent deployment, and to the grid, Karhunen–Loève, SMC, and Nyström approximation error is the main open theoretical question raised here, and is beyond the scope of this paper. The empirical results are most relevant when the tasks come from related distributions, so that sharing across tasks can pay, and when the cold-start uncertainty of a new user must be quantified rather than assumed away. Removing the parametric specification of the population prior raises its own questions: distribution shift, nonstationarity, missing or uninformative metadata, and populations whose tasks are less related than the model presumes. The hierarchy stays useful under all of these because it keeps uncertainty about a user's latent reward vector, uncertainty about the shared population, and observation noise separate. Calibrated diagnostics and model checks for these three sources are the corresponding practical next step.

A second direction is scalability. The present construction keeps the statistical and uncertainty-quantification questions visible, at the price of a joint reward–context grid whose size grows as $G \asymp n_\mu^K n_x^{d_{\mathrm{obs}}}$. Larger action spaces and higher-dimensional metadata therefore require a different representation: neural encoders for the metadata, amortised posterior approximations, deep-kernel or representation learning, or generative models that return a posterior rather than a point prediction. The constraint on any such replacement is that it must retain calibrated uncertainty for cold-start users and sparse-data tasks, which is the property the hierarchy exists to supply. Establishing when nonparametric population-prior learning measurably improves on parametric hierarchical baselines in deployed decision systems remains open.