Distributionally Robust Warm Starts for a Population of Bandits

Sumantrak Mukherjee[*]
DFKI
Kaiserslautern, Germany

Debabrota Basu[*]
Équipe Scool, Inria
Univ. Lille, CRIStAL, France

Jasmin Brandt
Bielefeld University
Bielefeld, Germany

Viktor Bengs
DFKI
Kaiserslautern, Germany

Eyke Hüllermeier
LMU Munich, Munich
DFKI, Kaiserslautern, Germany

Sebastian J. Vollmer
DFKI, Kaiserslautern
RPTU Kaiserslautern-Landau, Germany

Abstract

Warm-starting a Bayesian multi-armed bandit algorithm for a population of bandits requires an initialization in the form of a prior. To this end, one usually makes use of previously collected data. However, in many real-world applications, the deployment distribution may differ from the distribution from which data was collected. In this work, we propose Distributionally Robust Warm-Starts (DRWS) for warm-starting under distribution shift, i.e., changes of the population or reward distributions. DRWS uses scalable regret proxies and provides guarantees relating proxy error, sample size, and uncertainty-set design to worst-case performance, as well as conditions for robustness gains over nominal warm-starts. Experiments on synthetic and semi-synthetic benchmarks show improved early-round and worst-case regret under realistic shifts.

1 Introduction

In many applications, such as adaptive clinical trials (Varatharajah and Berry, 2022) or recommender systems (Zanker et al., 2019), each new user induces a fresh multi-armed bandit instance with little or no historical data, making effective warm-starting across a population essential. Warm-starting in Bayesian bandits requires selecting a prior before observing task-specific data, and this choice dominates early-round regret. In deployment, however, the task population often shifts relative to historical data, through changes in population composition and through preference drift, so priors optimized for nominal average performance can fail in safety-critical deployments.

We address this by learning a prior with a distributionally robust objective: tasks arise from subpopulation mixtures with uncertain weights and uncertain reward distributions, and we optimize the warm start against the worst-case mixture in a credal set. The objective is defined through within-task Bayes regret of a fixed adaptive learner, so robustness targets the full learning trajectory rather than a static predictor. Figure 1 illustrates the two shift mechanisms and the objective they induce.

To tackle this we introduce a new setting: distributionally robust warm-starting for *adaptive* bandit learning under deployment shift. Prior meta and hierarchical bandit methods typically learn warm

[*]Equal contribution.

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

starts under a fixed meta-distribution (Agrawal and Goyal, 2012; Russo and Van Roy, 2016; Hong et al., 2022; Wan et al., 2021; Kveton et al., 2021; Nabi et al., 2022; Basu et al., 2021), whereas distributionally robust optimization (DRO) and robust-policy methods focus on worst-case guarantees, often for static decisions or non-warm-start settings (Bertsimas and Sim, 2004; Namkoong and Duchi, 2016; Duchi and Namkoong, 2021; Duchi et al., 2021; Kuhn et al., 2025; Hu et al., 2018; Uehara et al., 2020; Faury et al., 2020; Sakhi et al., 2020; Si et al., 2023; Bubeck and Slivkins, 2012; Zimmert and Seldin, 2019; Herbster et al., 2021).

Our setting differs from empirical-Bayes and hierarchical Thompson-sampling approaches because the learned initialization is not judged only under the historical meta-distribution: it is optimized against an uncertainty set of plausible deployment populations. It also differs from standard adversarial or robust bandit work, where the online reward process or policy is robustified directly. Here the online learner is fixed and only its initial belief is chosen robustly, which makes the prior a deployment-time control variable whose effect is path-dependent through posterior updates and future exploration.

A second distinction is methodological. Static DRO and distributionally robust supervised learning often reduce to reweighting losses for a fixed predictor, while DRWS reweights entire task trajectories whose losses depend on future posterior states. Similarly, robust policy-learning methods typically optimize a policy class directly, whereas our optimizer changes the *initial belief* of an otherwise fixed adaptive algorithm. This separation is useful in applications where the online learner is already deployed or certified: DRWS improves robustness by changing the initialization layer, not the per-round decision rule. The closest baseline conceptually is a robust barycenter of task distributions, but such barycenters ignore the asymmetric value of information induced by Thompson sampling, and two priors that are close distributionally can have different early exploration behavior and hence different regret.

Contributions.

- **Robust formulation.** We formulate warm-start prior selection as a minimax DRO problem over composition and preference uncertainty.
- **Estimator and optimisation.** We derive a smooth dual form of that objective, estimate it from historical task samples, and optimize it by projected stochastic gradient steps through a differentiable regret proxy.
- **Guarantees.** We give finite-sample bounds in terms of proxy error, reference sample size, and uncertainty-set size, and show when DRWS improves over nominal warm-starting by reducing within-subgroup regret variability and subgroup-composition sensitivity.
- **Empirical gains under shift.** On synthetic and semi-synthetic benchmarks, DRWS reduces early-round and worst-case regret.

In Section 2, we introduce the problem setting. In Section 3, we model the data-generating process and the uncertainty around it. In Section 4, we introduce DRWS, and in Section 5 we give its finite-sample analysis. In Section 6, we describe the experimental setup and report the results, and Section 7 concludes. Appendix B discusses further related work, Appendices D and E give the algorithms and the proofs, and Appendix F the experimental details.

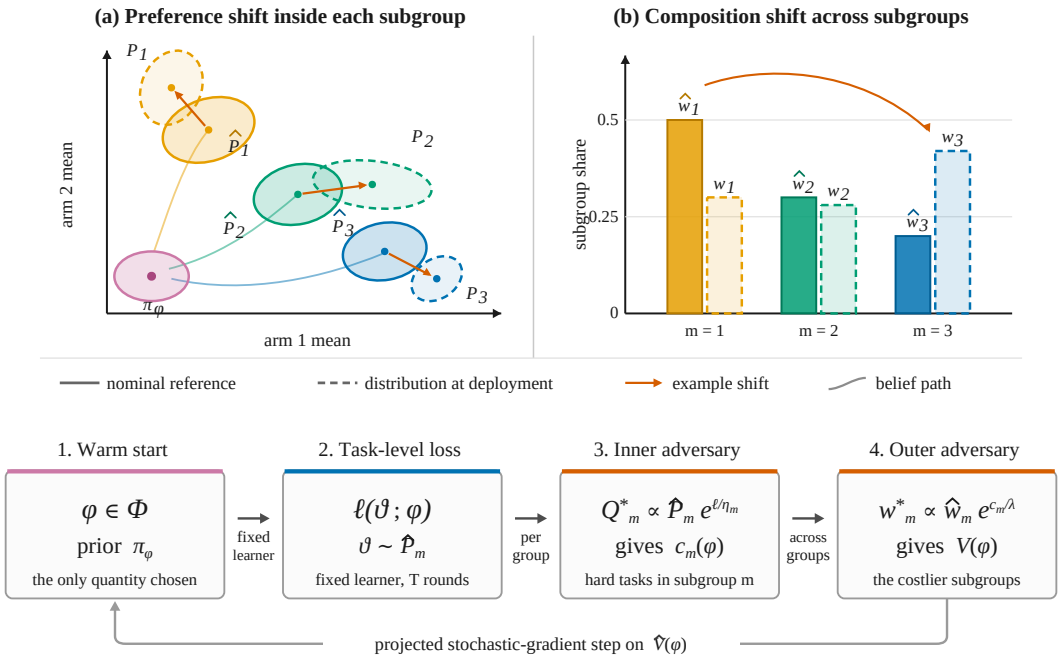
2 Problem Setting

We study regret minimization over a *population of multi-armed bandit instances*. The population has $A \in \mathbb{N}$ arms, and each instance is represented by a mean-reward vector $\vartheta \in \mathbb{R}^A$, which we call the task parameter. To model heterogeneity, instances are drawn from a mixture over M subpopulations,

$$P(\mathbf{w}) := \sum_{m=1}^{M} w_m P_m, \qquad \mathbf{w} \in \Delta_M, \tag{1}$$

where $P_m \in \mathcal{P}(\mathbb{R}^A)$ and Δ_M is the probability simplex on $[M] := \{1, \dots, M\}$.

A deployment consists of K tasks $\vartheta_1, \dots, \vartheta_K$ drawn from $P(\mathbf{w})$, each handled by one run of a fixed base learner `Alg` over a horizon of T rounds. Every run starts from the same warm-start prior π_φ, parameterized by $\varphi \in \Phi \subset \mathbb{R}^d$, and is given no information about the other runs, so the runs may be



(c) The two adversaries, and the objective they define

Figure 1: Deployment shift and the objective DRWS builds from it. **(a)** Preference shift inside each subgroup, in the plane of two arm means: solid ellipses are the nominal references $\widehat{P}_m$, dashed ellipses the subgroup distributions P_m at deployment, and the purple region the warm-start prior π_φ. Each subgroup moves in its own direction and may change shape. **(b)** Composition shift: solid bars are the nominal shares $\widehat{w}_m$, dashed bars the deployment shares w_m. Section 3.2 bounds how far either shift may go. **(c)** The inner adversary tilts each $\widehat{P}_m$ towards its high-regret tasks and returns $c_m(\varphi)$, the outer tilts the shares towards the costlier subgroups and returns $V(\varphi)$, and η_m, λ are the dual variables of the two Kullback–Leibler (KL) constraints.

executed in any order or in parallel. Within the run for task ϑ_k, the learner selects an arm $A_{k,t} \in [A]$ at round t, observes the reward $Y_{k,t} \sim p(\cdot \mid A_{k,t}, \vartheta_k)$ with mean $\mu_a(\vartheta) = \vartheta_a$ bounded by $R_{\max}$, and appends it to the history $\mathcal{H}_{k,t} := \{(A_{k,s}, Y_{k,s})\}_{s=1}^{t}$, so its action is measurable with respect to $\sigma(\mathcal{H}_{k,t-1})$ and its own randomization. It knows M and the nominal weights $\widehat{\mathbf{w}}$, but observes neither the subpopulation identity of ϑ_k nor the deployment weights $\mathbf{w}$. A historical dataset of past interactions is available before the first run, collected under a population that need not coincide with $P(\mathbf{w})$. It is the only source of the nominal quantities $\widehat{\mathbf{w}}$ and $\widehat{P}_{1:M}$ of Section 3.2, and is never updated during a deployment.

For task ϑ_k, the expected episode regret of the warm start π_φ is

$$R_k(\varphi; \vartheta_k) := T\vartheta_k^\star - \mathbb{E}_{\mathtt{Alg}(\pi_\varphi)}\left[\sum_{t=1}^{T} Y_{k,t}\right], \qquad \vartheta_k^\star := \max_{a \in [A]} \vartheta_{k,a}, \tag{2}$$

where the expectation is over the actions and rewards generated by $\mathtt{Alg}$ started at π_φ. Population-level Bayes regret aggregates (2) over the deployment,

$$\mathrm{BR}_{P(\mathbf{w})}(\mathtt{Alg}, \pi_\varphi) := \mathbb{E}_{\vartheta_{1:K} \sim P(\mathbf{w})}\left[\sum_{k=1}^{K} R_k(\varphi; \vartheta_k)\right], \tag{3}$$

and smaller values are better. We write $\mathrm{BR}_T(\vartheta; \pi_\varphi)$ for the single-task quantity (2) and $\mathrm{BR}_T(Q; \pi_\varphi) := \mathbb{E}_{\vartheta \sim Q}[\mathrm{BR}_T(\vartheta; \pi_\varphi)]$ for its average under a task distribution Q. Both are cumulative over the horizon and measure the price of not knowing the task parameter in advance, and both are evaluated at a fixed T, since π_φ acts on the early rounds and its influence decays as the posterior concentrates.

The warm-start objective is to choose a prior that minimizes worst-case population Bayes regret over an uncertainty class $\mathcal{P}$ of plausible deployment distributions,

$$\varphi^\star \in \arg\min_{\varphi\in\Phi} \sup_{P(\mathbf{w})\in\mathcal{P}} \mathrm{BR}_{P(\mathbf{w})}(\texttt{Alg}, \pi_\varphi), \tag{4}$$

where the base learner `Alg` is an input and is never modified. The uncertainty class $\mathcal{P}$ is left unspecified here, and Section 3 builds it from the historical data.

3 Statistical Model of the Task Population

Section 3.1 states the generative model of the task population and instantiates it for Gaussian subgroups. Section 3.2 estimates the two components of that model from the historical data, places a credal set around each estimate, and produces the uncertainty class $\mathcal{P}$ that (4) needs.

3.1 A Generative Model of the Bandit Population

We model task generation hierarchically: first sample a subpopulation, then sample the task parameter from that subpopulation, and keep it fixed within a run. For task k, the learner interacts for T rounds, choosing $A_{k,t} \in [A]$ and observing rewards from the arm-conditional model,

$$z_k \mid \mathbf{w} \sim \mathrm{Categorical}(\mathbf{w}), \tag{5}$$

$$\vartheta_k \mid (z_k = m) \sim P_m, \tag{6}$$

$$Y_{k,t} \mid (A_{k,t} = a, \vartheta_k) \sim p(\,\cdot \mid a, \vartheta_k). \tag{7}$$

The subgroup index z_k is latent and is never revealed to the learner. The first level of the hierarchy is a choice among the finitely many subpopulation labels $[M]$, and a categorical distribution is the general form of a distribution on a finite set, so (5) restricts the population only through the choice of M itself. Its parameter $\mathbf{w}$ is the population composition. The second level places the task parameter in $\mathbb{R}^A$ under the subgroup distribution P_m, and (6) leaves that distribution unrestricted. The two levels of (5)–(6) carry the two shift mechanisms of Figure 1: composition shift moves $\mathbf{w}$, the parameter of the discrete choice, and preference shift moves the subgroup distributions P_m of the continuous one.

Example 1 (Gaussian subgroups)**.** *Writing the task parameter as $\vartheta_k \equiv \mu_k \in \mathbb{R}^A$ and taking Gaussian subgroups makes* (5)–(6) *explicit,*

$$z_k \mid \mathbf{w} \sim \mathrm{Categorical}(\mathbf{w}), \qquad \mu_k \mid (z_k = m) \sim \mathcal{N}\big(\nu_m, \mathrm{diag}(\tau_m^2)\big), \tag{8}$$

with Gaussian rewards $Y_{k,t} \mid (A_{k,t} = a, \mu_k) \sim \mathcal{N}(\mu_{k,a}, \sigma^2)$ in (7)*. Here $\nu_m \in \mathbb{R}^A$ is the arm-mean profile of subgroup m, so its a-th entry is the average reward arm a delivers to a member of that subgroup, and $\tau_m \in \mathbb{R}^A_{>0}$ measures how far individual members depart from that profile. A subgroup whose profile peaks at one arm is a subgroup that prefers that arm, and two subgroups whose profiles peak at different arms disagree about which arm is best. Preference shift moves ν_m or inflates τ_m, and composition shift moves* $\mathbf{w}$.

3.2 Quantifying the Sources of Uncertainty

The reward noise in (7) is irreducible *aleatoric* uncertainty that no choice of prior can eliminate. Deployment shift instead induces *epistemic* uncertainty, which is systemic model uncertainty that we account for explicitly and hedge against with credal sets around the nominal estimates $(\widehat{\mathbf{w}}, \widehat{P}_1, \ldots, \widehat{P}_M)$.

Reference distributions from historical data. For each historical task i assigned to subpopulation m, we use its observed history $H_i = \{(A_{i,t}, Y_{i,t})\}_{t=1}^{T_i}$ to form either a posterior $p(\vartheta \mid H_i)$ or a point estimate $\widehat{\vartheta}_i$, and with D_m indexing the historical tasks in subpopulation m we set

$$\widehat{P}_m = \frac{1}{|D_m|} \sum_{i:\, m_i = m} p(\vartheta \mid H_i) \quad \text{or} \quad \widehat{P}_m = \frac{1}{|D_m|} \sum_{i:\, m_i = m} \delta_{\widehat{\vartheta}_i}, \qquad \widehat{w}_m = \frac{|D_m|}{\sum_j |D_j|}. \tag{9}$$

Expectations are approximated by Monte Carlo samples $\vartheta \sim \widehat{P}_m$, which makes the objective below computable from historical data alone.

Credal sets. We hedge against within-subpopulation and composition shift with the KL balls

$$\mathcal{P}_m := \{Q : D_{\mathrm{KL}}(Q \,\|\, \widehat{P}_m) \le \delta_m\}, \qquad \mathcal{W} := \{\mathbf{w} \in \Delta_M : D_{\mathrm{KL}}(\mathbf{w} \,\|\, \widehat{\mathbf{w}}) \le \delta_w\}, \tag{10}$$

with radii $\delta_m > 0$ and $\delta_w > 0$. KL balls yield smooth dual forms, which keeps the objective tractable and its optimization stable. For a fixed `Alg` and π_φ, define

$$c_m(\varphi) := \sup_{Q_m \in \mathcal{P}_m} \mathrm{BR}_T(Q_m; \pi_\varphi), \qquad V(\varphi) := \sup_{\mathbf{w} \in \mathcal{W}} \sum_{m=1}^{M} w_m \, c_m(\varphi), \tag{11}$$

and optimize

$$\varphi^* \in \arg\min_{\varphi \in \Phi} V(\varphi). \tag{12}$$

This exposes two adversaries: one over subpopulation distributions and one over mixture composition.

Decomposition of deployment shift. The model (10) separates deployment shift into composition shift over subgroup proportions and within-subgroup preference shift, mirroring the KL chain rule,

$$D_{\mathrm{KL}}\big(P(\vartheta, z) \,\|\, Q(\vartheta, z)\big) = D_{\mathrm{KL}}\big(P(z) \,\|\, Q(z)\big) + \mathbb{E}_{z \sim P}\big[D_{\mathrm{KL}}\big(P(\vartheta \mid z) \,\|\, Q(\vartheta \mid z)\big)\big], \tag{13}$$

which motivates the nested structure of (11): the outer adversary controls mixture composition, the inner adversaries control subgroup-specific reward uncertainty, and the two radii may be set from different evidence. Neither supremum can be evaluated as written, and Section 4 removes the two obstacles that stand in the way.

4 Distributionally Robust Evaluation of the Warm Start

In this section we discuss our method for estimating the robust warm-start objective (12) and for optimizing the warm start against it. Two obstacles separate (11) from a quantity a computer can return. The first is the range of each supremum, which is an infinite-dimensional set of distributions. The second is the quantity being maximized, the Bayes regret $\mathrm{BR}_T(\vartheta; \pi_\varphi)$, which is the expected loss of a complete learning trajectory and has no closed form in φ. Section 4.1 removes the first obstacle. Duality turns each supremum into a scalar convex problem over a temperature, and the historical samples turn the remaining expectations into averages. Section 4.2 removes the second. A differentiable proxy $\ell(\vartheta; \varphi)$ takes the place of the Bayes regret, because the optimizer needs a gradient in φ and a single robust evaluation calls the regret once for every sampled task. The proxy is itself estimated from simulated interaction, since the regret of a trajectory is accessible only through the behaviour of the base learner. Section 4.3 combines the two into the optimizer.

4.1 KL-Duality Based Estimators

The suprema in (11) range over infinite-dimensional sets, and the following duality reduces each of them to a one-dimensional convex problem. To separate population quantities from their sample approximations, write P_m^0 and $\mathbf{w}^0$ for the nominal objects that define the KL balls in (10).

Lemma 1 (Dual evaluation of the robust warm start)**.** *Let $\ell(\vartheta; \varphi)$ be measurable with $0 \le \ell \le L_{\max}$, for instance the task regret $\mathrm{BR}_T(\vartheta; \pi_\varphi)$ or a differentiable proxy for it. Then, for every $\varphi \in \Phi$,*

$$c_m(\varphi) = \inf_{\eta_m > 0} \Big\{ \eta_m \delta_m + \eta_m \log \mathbb{E}_{\vartheta \sim P_m^0} \big[\exp\big(\ell(\vartheta; \varphi)/\eta_m\big) \big] \Big\}, \tag{14}$$

$$V(\varphi) = \inf_{\lambda > 0} \Big\{ \lambda \delta_w + \lambda \log \textstyle\sum_{m=1}^{M} w_m^0 \exp\big(c_m(\varphi)/\lambda\big) \Big\}. \tag{15}$$

Appendix E.1 proves Lemma 1 by applying the Donsker–Varadhan variational formula first inside each subgroup and then to the finite mixture, and Remark 1 there records which part of the problem this makes convex, since the outer minimization over φ stays non-convex. A robust evaluation therefore costs one expectation of an exponentiated loss per subgroup, plus a scalar minimization over the dual variables.

The optimal adversaries are exponential tiltings of the references,

$$\frac{dQ_m^\star}{dP_m^0}(\vartheta) \propto \exp\big(\ell(\vartheta; \varphi)/\eta_m^\star\big), \qquad w_m^\star \propto w_m^0 \exp\big(c_m(\varphi)/\lambda^\star\big), \tag{16}$$

where $\eta_m^\star$ and $\lambda^\star$ are the optimal dual variables. The inner tilting raises the weight of high-regret tasks inside a subgroup and the outer tilting raises the weight of high-cost subgroups, so the objective penalizes a warm start whose regret varies a lot within a subgroup, or whose quality varies a lot across subgroups.

Empirical robust objective. In practice P_m^0 and $\mathbf{w}^0$ are the empirical references (9). Given samples $\{\vartheta_{m,j}\}_{j=1}^{K_m} \sim \widehat{P}_m$ and weights $\widehat{w}_m$, replacing the nominal expectation in (14) by a sample average gives

$$\widehat{c}_m(\varphi) := \inf_{\eta_m>0}\Big\{\eta_m\delta_m + \eta_m \log\big(K_m^{-1}\textstyle\sum_{j=1}^{K_m}\exp\big(\ell(\vartheta_{m,j};\varphi)/\eta_m\big)\big)\Big\}, \tag{17}$$

and substituting $\widehat{c}_m(\varphi)$ and $\widehat{w}_m$ into (15) gives the empirical robust objective $\widehat{V}(\varphi)$, which DRWS minimizes over Φ. It preserves the nested structure of (11) and is smooth and directly optimizable in $(\varphi, \eta_{1:M}, \lambda)$ whenever ℓ is differentiable.

4.2 Estimators of Regret

The dual objective requires only a differentiable task-level regret proxy $\ell(\vartheta;\varphi)$. For a base learner initialized at π_φ, the target quantity is the cumulative pseudo-regret on task ϑ,

$$\mathrm{BR}_T(\vartheta;\pi_\varphi) = \mathbb{E}\left[\sum_{t=1}^{T}\big(\mu^\star(\vartheta) - \mu_{A_t}(\vartheta)\big)\right], \qquad \mu^\star(\vartheta) := \max_{a\in[A]}\mu_a(\vartheta), \tag{18}$$

where a Bayesian learner updates its belief by

$$\pi_{k,t}(d\vartheta) = \frac{p(Y_{k,t-1} \mid A_{k,t-1},\vartheta)\,\pi_{k,t-1}(d\vartheta)}{\int p(Y_{k,t-1} \mid A_{k,t-1},\vartheta')\,\pi_{k,t-1}(d\vartheta')}. \tag{19}$$

Exact Monte Carlo evaluation of (18) is possible but too expensive inside repeated robust updates, and Theorem 2 accounts for the error of any replacement through the term $\epsilon_{\mathrm{proxy}}$. We use two interchangeable modules rather than one, because no single estimator supplies both properties the method needs. A module inside the inner loop of the optimizer must be cheap enough to call at every gradient step, and a module that enters the analysis must come with an error bound that can be stated. The amortized neural surrogate meets the first requirement and controls its error only through a validation loss. The belief-ODE estimator meets the second and costs more per call. Running both separates the error of the robust machinery from the error of the regret estimate, since a disagreement between the two modules on the same warm start is attributable to the proxy alone.

The amortized neural surrogate. This module first creates rollout labels $\widehat{\mathrm{BR}}_T(\vartheta;\pi_\varphi) = S^{-1}\sum_{s=1}^{S}\sum_{t=1}^{T}(\mu^\star(\vartheta) - \mu_{A_t^{(s)}}(\vartheta))$ from S runs of the learner, and then trains $f_\phi(\pi_\varphi,\vartheta)$ by squared-error regression on those labels, which makes robust optimization cheap after an offline labelling and training phase.

The belief-ODE estimator. This module instead replaces the stochastic posterior recursion by a deterministic flow. Writing the belief in natural coordinates as $\pi_t = \pi(\xi_t)$ with $\xi_t \in \mathcal{E} \subset \mathbb{R}^d$, and letting $\mathcal{B}(\pi,a,y)$ denote the Bayes operator of (19), the mean drift and the instantaneous regret rate are

$$\dot{\xi}_t = \mathbb{E}_{A_t\sim\psi_t,Y_t}\big[\xi\big(\mathcal{B}(\pi_t,A_t,Y_t)\big) - \xi_t\big], \qquad r(\xi_t,\vartheta) = \mathbb{E}_{A_t\sim\psi_t}\big[\mu^\star(\vartheta) - \mu_{A_t}(\vartheta)\big], \tag{20}$$

where ψ_t is the action distribution induced by π_t, and the proxy is the time integral

$$\widehat{\mathrm{BR}}_T^{\,\mathrm{ODE}}(\vartheta;\pi_\varphi) = \int_0^T r(\xi_t,\vartheta)\,dt. \tag{21}$$

For Beta–Bernoulli Thompson sampling the flow takes the form $\dot{\alpha}_{t,a} = \psi_t(a)\mu_a(\vartheta)$ and $\dot{\beta}_{t,a} = \psi_t(a)(1-\mu_a(\vartheta))$, so it tracks explicitly how the warm start changes exploration and posterior evolution. Proposition 1 and Theorem 5 supply the Lipschitz property and the error bound that Section 5 requires of a proxy. The neural surrogate is therefore our scalable default, and the belief-ODE estimator the structured comparator with theoretical control. Algorithm 1 states both modules side by side.

Algorithm 1 The two regret-estimation modules. The belief-ODE estimator integrates the deterministic flow (20) along one trajectory, and is structured but slower. The neural surrogate amortizes rollout labels offline, after which one evaluation is a forward pass.

(a) Belief-ODE regret estimator

Require: prior π_φ, task ϑ, horizon T, step Δt

set $\xi \leftarrow \xi(\pi_\varphi)$, $\widehat{R} \leftarrow 0$
for $u = 0, \Delta t, 2\Delta t, \dots, T$ **do**
  compute $\psi(a \mid \pi(\xi))$ for all arms $a \in [A]$
  $\widehat{R} \leftarrow \widehat{R} + \Delta t \sum_a \psi(a \mid \pi(\xi))\big(\mu^\star(\vartheta) - \mu_a(\vartheta)\big)$
  $\dot{\xi} \leftarrow \mathbb{E}_{A,Y}\big[\xi(\mathcal{B}(\pi(\xi), A, Y)) - \xi\big]$
  $\xi \leftarrow \xi + \Delta t\, \dot{\xi}$
end for
return $\widehat{R}$

(b) Neural regret surrogate

Require: sampler for (π_φ, ϑ), rollouts S, horizon T, network f_ϕ, dataset size J

for $j = 1$ to J **do**
  sample $(\pi^{(j)}, \vartheta^{(j)})$
  run S rollouts of `Alg` from $\pi^{(j)}$ on $\vartheta^{(j)}$
  store $\widehat{r}_j = \frac{1}{S} \sum_s \sum_{t=1}^{T} \big(\mu^\star - \mu_{A_t^{(s)}}\big)$
end for
train f_ϕ by squared-error regression on $\{\widehat{r}_j\}$
return trained proxy $f_\phi(\pi_\varphi, \vartheta)$

Algorithm 2 Distributionally Robust Warm-Start optimization

Require: nominal references $\widehat{P}_{1:M}$ and weights $\widehat{w}_{1:M}$; radii $\delta_{1:M}$ and δ_w; regret proxy ℓ; admissible set Φ; initial parameter $\varphi_0 \in \Phi$; step sizes $\{\rho_s\}$; minibatch sizes $\{K_m\}$; iteration budget S

initialize $\varphi^{\mathrm{best}} \leftarrow \varphi_0$ and $v^{\mathrm{best}} \leftarrow +\infty$
for $s = 0$ to $S - 1$ **do**
  draw a minibatch $\{\vartheta_{m,j}\}_{j=1}^{K_m} \sim \widehat{P}_m$ and evaluate $\ell(\vartheta_{m,j}; \varphi_s)$ for every $m \in [M]$
  solve the inner dual (17) for η_m and obtain $\widehat{c}_m(\varphi_s)$ for every $m \in [M]$
  solve the outer dual (15) for λ and obtain $\widehat{V}(\varphi_s)$
  update $\varphi_{s+1} \leftarrow \mathrm{Proj}_\Phi\big(\varphi_s - \rho_s \nabla_\varphi \widehat{V}(\varphi_s)\big)$ as in (22)
  if $\widehat{V}(\varphi_{s+1}) < v^{\mathrm{best}}$ **then**
    set $\varphi^{\mathrm{best}} \leftarrow \varphi_{s+1}$ and $v^{\mathrm{best}} \leftarrow \widehat{V}(\varphi_{s+1})$
  end if
end for
return warm-start prior $\pi_{\varphi^{\mathrm{best}}}$

4.3 Optimizing the Warm Start

The empirical objective of Section 4.1 serves as the training loss for the prior. At a candidate φ_s we draw minibatches from the empirical subgroup references, compute the inner KL-dual subgroup costs $\widehat{c}_m(\varphi_s)$ from (17), and combine them through the outer KL dual to obtain $\widehat{V}(\varphi_s)$. Since the regret proxy and both dual objectives are differentiable, gradients are backpropagated through both levels of the robust objective, and the resulting update is the projected stochastic-gradient step

$$\varphi_{s+1} = \mathrm{Proj}_\Phi\big(\varphi_s - \rho_s \nabla_\varphi \widehat{V}(\varphi_s)\big), \qquad \nabla_\varphi \widehat{V}(\varphi_s) = \sum_{m=1}^{M} w_m^\star \, \nabla_\varphi \widehat{c}_m(\varphi_s), \tag{22}$$

with the scalar dual variables $(\eta_{1:M}, \lambda)$ optimized inside each objective evaluation and $w^\star$ the tilted weights of (16) evaluated at φ_s. Since the objective is generally non-convex, we use multiple random initializations and keep the iterate with the lowest validation robust regret. Algorithm 2 summarizes the training loop.

5 Finite-Sample Analysis

Theorem 1 controls the distance between a robust evaluation built from finitely many historical tasks and its population counterpart, and Theorem 2 turns that into a guarantee for the learned warm start.

Assumption 1 (Bounded and Lipschitz regret proxy)**.** *The proxy satisfies $0 \le \ell(\vartheta; \varphi) \le L_{\max}$ and $|\ell(\vartheta; \varphi) - \ell(\vartheta; \varphi')| \le L_\Phi \|\varphi - \varphi'\|$ for all ϑ and all $\varphi, \varphi' \in \Phi$. Bounded rewards over a finite horizon give the first condition, and Proposition 1 gives the second for the belief-ODE proxy.*

Assumption 2 (Compact parameter set and non-degenerate duals)**.** *The set $\Phi \subset \mathbb{R}^d$ is compact with ε-covering number $\mathcal{N}(\varepsilon, \Phi)$, and the optimal dual variables satisfy $\eta_m^\star(\varphi) \in [\eta_{\min}, \eta_{\max}]$ and*

$\lambda^\star(\varphi) \in [\lambda_{\min}, \lambda_{\max}]$ *with* $\eta_{\min}, \lambda_{\min} > 0$*, uniformly in* m *and* φ*. Appendix E.2 shows that this holds whenever each radius stays below the value at which the adversary could concentrate all mass on a maximizer of the proxy.*

Assumption 3 (Uniform proxy error)**.** *With* V^{proxy} *and* V^{true} *the objective* (11) *built from the proxy and from the exact task regret* (18)*,* $\sup_{\varphi\in\Phi} |V^{\mathrm{proxy}}(\varphi) - V^{\mathrm{true}}(\varphi)| \le \epsilon_{\mathrm{proxy}}$.

Theorem 1 (Uniform convergence of the robust evaluation)**.** *Let* $\widehat{V}$ *be the empirical objective of Section 4.1 formed from* K_m *independent samples of* $\widehat{P}_m$*, let* V *be its population counterpart* (11) *under the same proxy, set* $K_{\min} := \min_m K_m$*, and suppose Assumptions 1 and 2 hold. Then, for every* $\varepsilon > 0$ *and* $\delta \in (0,1)$*, with probability at least* $1-\delta$ *over the task samples,*

$$\sup_{\varphi\in\Phi} \big|\widehat{V}(\varphi) - V(\varphi)\big| \le C\sqrt{\frac{\log \mathcal{N}(\varepsilon,\Phi) + \log(M/\delta)}{K_{\min}}} + O(\varepsilon), \tag{23}$$

where $C = C(L_{\max}, L_\Phi, \eta_{\min}, \eta_{\max}, \lambda_{\min}, \lambda_{\max})$.

Appendix E.4 proves Theorem 1, where the constant C is built from the problem parameters of Assumptions 1 and 2 and does not depend on $K_{\min}$, M or δ. The first term is sampling error and decays at the parametric rate $K_{\min}^{-1/2}$, with only a logarithmic price for the size of Φ and the number of subgroups, and the second term is the resolution of the covering argument. If n historical tasks come from the nominal mixture and $\widehat{w}_m \ge w_{\min} > 0$ for every m, then $K_{\min} \gtrsim n w_{\min}$ with high probability, so the rate in the total sample size is $\widetilde{O}((n w_{\min})^{-1/2})$: the smallest subgroup sets the rate, which is the statistical price of worst-case performance across all subgroups.

Theorem 2 (Excess robust regret of the learned warm start)**.** *Let* $\widehat{\varphi} \in \arg\min_{\varphi\in\Phi} \widehat{V}(\varphi)$ *be the output of DRWS and* $\varphi^\star \in \arg\min_{\varphi\in\Phi} V^{\mathrm{true}}(\varphi)$ *optimal under the true robust objective, and suppose Assumptions 1–3 hold. Then, for every* $\varepsilon > 0$ *and* $\delta \in (0,1)$*, with probability at least* $1-\delta$ *over the task samples,*

$$V^{\mathrm{true}}(\widehat{\varphi}) - V^{\mathrm{true}}(\varphi^\star) \le 2\epsilon_{\mathrm{proxy}} + 2C\sqrt{\frac{\log \mathcal{N}(\varepsilon,\Phi) + \log(M/\delta)}{K_{\min}}} + O(\varepsilon), \tag{24}$$

with C*,* $K_{\min}$*, and* $\mathcal{N}(\varepsilon,\Phi)$ *as in Theorem 1.*

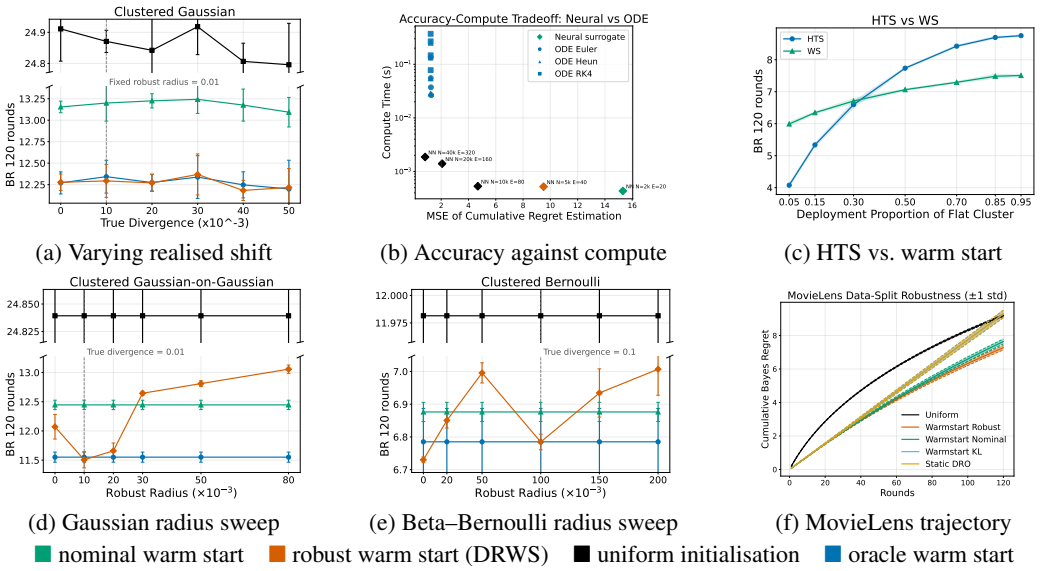
Appendix E.6 proves Theorem 2. The estimation term shrinks as more historical tasks arrive, whereas the proxy term does not, and is the price of replacing rollouts by a differentiable approximation.

Error contributed by the belief-ODE proxy. Under Assumption 4 in Appendix E.8, Proposition 1 makes the belief-ODE proxy Lipschitz in its initialization with constant $2R_{\max}L_\psi(e^{L_F T}-1)/L_F$, which supplies the Lipschitz half of Assumption 1, and Theorem 5 bounds its error by $L_r T\big((e^{L_F T}-1)/L_F\,\epsilon_{\mathrm{disc}} + \epsilon_{\mathrm{num}}\big)$, an admissible value of $\epsilon_{\mathrm{proxy}}$ that grows with the horizon.

When robustness pays. Let φ_{nom} minimize the nominal average regret and set $\Gamma := V(\varphi_{\mathrm{nom}}) - \inf_{\varphi\in\Phi} V(\varphi) \ge 0$. Corollary 1 in Appendix E.7 gives $V(\widehat{\varphi}) \le V(\varphi_{\mathrm{nom}}) - \Gamma + \epsilon_{\mathrm{stat}}$ whenever the excess robust risk of $\widehat{\varphi}$ is at most the ϵ_{stat} that Theorem 2 bounds, and Theorem 4 lower-bounds Γ at small radii, where V equals the nominal average regret plus a within-subgroup dispersion penalty $\sum_m \widehat{w}_m\sqrt{2\delta_m}\,\sigma_m(\varphi)$ and a composition-sensitivity penalty $\sqrt{2\delta_w \mathrm{Var}_{m\sim\widehat{w}}(\mu_m(\varphi))}$. A warm start gains when it trades a little average regret for a reduction in either, which Figures 2(d) and 2(e) test.

6 Experiments

Three questions drive the evaluation: whether hedging pays when the shift is real, how right the assumed radius has to be, and what the proxy costs in accuracy for what it saves in compute. We evaluate DRWS on Beta–Bernoulli bandits, Gaussian bandits, and a semi-synthetic MovieLens recommendation benchmark (Harper and Konstan, 2015), using Thompson sampling as the within-task learner. The two synthetic populations use three latent subgroups, five arms, and nominal weights $(0.5, 0.3, 0.2)$, and they set the type and magnitude of the deployment shift exactly, so the realized divergence can be varied while the assumed radius is held fixed, and the reverse. MovieLens supplies a shift we did not impose, from a temporal split of real interaction logs with clustered user and item feedback.



(a) Varying realised shift (b) Accuracy against compute (c) HTS vs. warm start

(d) Gaussian radius sweep (e) Beta–Bernoulli radius sweep (f) MovieLens trajectory

Figure 2: Main results, lower is better everywhere. Figures 2(a), (c), (d), (e) plot Bayes regret (BR) at horizon 120, (f) plots cumulative Bayes regret against the round index, and (b) plots inference compute against the mean squared error (MSE) of cumulative-regret prediction. Figures 2(a), (d), (e) use the shared colour key above and a broken vertical axis, with uniform initialisation in the upper segment; (b), (c), (f) carry their own keys, adding three belief-ODE solvers, hierarchical Thompson sampling (HTS) against the neural warm start (WS), and a KL-barycentre warm start with a static distributionally robust policy. Bands show one standard deviation across deployment draws.

Baselines. We compare six warm starts: a uniform initialization, a nominal warm start (WS) fit to $(\widehat{P}_m, \widehat{\mathbf{w}})$, DRWS, a KL-barycenter warm start that ignores learning dynamics, hierarchical Thompson sampling (HTS) as a meta-learning comparator, and an oracle warm start that knows the deployment distribution and gives a reference lower bound. All neural variants share the surrogate architecture, rollout-label budget, optimizer family, and restart and validation-selection protocol.

Protocol and metrics. Every method runs with horizon $T = 120$ over repeated deployment draws and Monte Carlo Thompson-sampling rollouts. We report cumulative Bayes regret at the horizon, area-under-curve regret, and worst-case regret across draws, with uncertainty reported as a 95% confidence interval or one standard deviation across draws, and we sweep the assumed KL radius at a fixed realized deployment divergence. Appendix F gives the sample sizes, seeds, and surrogate training details.

6.1 Results

Figure 2 reports six settings: whether the robust warm start helps as the realized preference shift grows (a), what accuracy each proxy gives per unit of inference compute (b), when warm-start optimization beats hierarchical Thompson sampling (c), how sensitive DRWS is to the assumed robustness radius (d, e), and how its advantage develops over the deployment horizon (f).

Fixed robustness radius, varying preference shift. Figure 2(a) fixes the training robustness radius and keeps population-composition shift fixed, while varying within-group preference shift at deployment, which isolates preference-shift sensitivity from composition effects. The robust prior tracks the oracle warm start across the swept range, to within the plotted error bars, while the nominal prior sits above it by a roughly constant margin and uniform initialization is worse by an order of magnitude. Neither adaptive prior degrades visibly over this range, so the evidence here is a level shift rather than a difference in slope, and Figures 2(d) and 2(e) carry the sensitivity claim.

This matches (12): robust training optimizes against an adversarial neighbourhood and improves tail behaviour under shift. More broadly, warm-start optimization is more nonlinear than standard

static off-policy evaluation. In static off-policy evaluation one often exploits convex or near-convex structure, whereas here the prior affects posterior updates, action probabilities, and long-horizon regret accumulation, which creates a path-dependent non-convex objective.

Inference compute versus regret-prediction error. Figure 2(b) compares the surrogates on inference-time compute against MSE in predicting cumulative regret. ODE surrogates are accurate because they preserve more of the bandit dynamics, but they are substantially more expensive at inference. Neural surrogates require one-time training yet give inexpensive inference, and at a large enough labelling budget they reach the accuracy of the ODE surrogates. Their downside is potential overfitting, weaker out-of-distribution generalization, and end-to-end guarantees that are harder to obtain than for structured ODE proxies.

Hierarchical Thompson sampling versus neural warm-starting. Figure 2(c) compares hierarchical Thompson sampling with warm-starting via a neural surrogate in a controlled two-subgroup setting where the deployment distribution is assumed known. One subgroup has a strong treatment preference while the other is ambiguous, and we vary the subgroup proportions. In this asymmetric setting, hierarchical Thompson sampling, like many meta methods, averages across tasks and under-exploits the subgroup asymmetry, so its regret grows steeply as the ambiguous subgroup takes over the population. Neural warm-starting optimizes the deployment-relevant initialization directly, so its curve is the flatter of the two and it overtakes hierarchical Thompson sampling once the ambiguous subgroup is common enough.

Radius sweeps and deployment-time regret evolution. Figures 2(d) and 2(e) report Gaussian and Beta–Bernoulli sweeps of the assumed robustness radius at fixed deployment divergence. The expected shape is U-like: underestimating the radius under-hedges the shift, matching the realized shift is near-optimal, and overestimating it becomes overly conservative through worst-case subgroup emphasis. In both settings the minimum is attained at the matched radius, where the robust prior coincides with the oracle warm start. Figure 2(f) is a cumulative-regret curve on the deployed semi-synthetic population, which shows the time evolution and the early-versus-late behaviour of the different warm-start choices, and complements the sweeps by visualizing deployment trajectory quality directly. The warm starts are indistinguishable while every prior still dominates the posterior, after which the robust prior separates from the nominal one and keeps the lowest cumulative regret to the horizon. The U-shape can be softened by non-convex surrogate optimization, since different radii may converge to different local minima, so we interpret the sweeps qualitatively, focusing on under-hedging at small radii and conservatism at large radii rather than on exact monotonicity.

7 Conclusion and Limitations

We formulated warm-start selection for a population of bandits as a minimax problem over a nested Kullback–Leibler uncertainty set. Its dual form is smooth, computable from historical task samples, and exposes both adversaries as exponential tiltings of the nominal references. Our analysis separates estimation error at rate $\widetilde{O}((nw_{\min})^{-1/2})$ from proxy error ϵ_{proxy}, and the learned warm start matched an oracle whenever the assumed radius equalled the realized divergence.

Subgroup assignments and radii are inputs from pilot data: too few subgroups hide vulnerable populations, too many increase estimation error through a smaller $K_{\min}$. KL balls yield smooth duals but cannot represent a shift that places mass outside the support of the reference, and neural proxies scale at the price of an out-of-distribution risk that Theorem 5 does not cover. A Bayes regret bound for an arbitrary prior relative to a reference prior is a natural next step.

Acknowledgments

Debabrota Basu is partially supported by the ANR JCJC project REPUBLIC (ANR-22-CE23-0003-01). 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

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings.

Avin, C., Lotker, Z., Mannor, S., Shabat, G., Shteingart, H., and Yadgar, R. (2025). Adversarial bandit over bandits: Hierarchical bandits for online configuration management. *arXiv preprint arXiv:2505.19061*.

Basu, S., Kveton, B., Zaheer, M., and Szepesvári, C. (2021). No regrets for learning the prior in bandits. *Advances in neural information processing systems*, 34:28029–28041.

Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations research*, 52(1):35–53.

Bubeck, S. and Slivkins, A. (2012). The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings.

Duchi, J. C., Glynn, P. W., and Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969.

Duchi, J. C. and Namkoong, H. (2021). Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406.

Faury, L., Tanielian, U., Dohmatob, E., Smirnova, E., and Vasile, F. (2020). Distributionally robust counterfactual risk minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3850–3857.

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Herbster, M., Pasteris, S., Vitale, F., and Pontil, M. (2021). A gang of adversarial bandits. *Advances in Neural Information Processing Systems*, 34:2265–2279.

Hong, J., Kveton, B., Zaheer, M., and Ghavamzadeh, M. (2022). Hierarchical bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR.

Hu, Z. and Hong, L. J. (2013). Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9.

Husain, H., Nguyen, V., and van den Hengel, A. (2023). Distributionally robust bayesian optimization with φ-divergences. *Advances in Neural Information Processing Systems*, 36:20133–20145.

Kuhn, D., Shafiee, S., and Wiesemann, W. (2025). Distributionally robust optimization. *Acta Numerica*, 34:579–804.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-w., Mladenov, M., Boutilier, C., and Szepesvari, C. (2021). Meta-thompson sampling. In *International Conference on Machine Learning*, pages 5884–5893. PMLR.

Mu, T., Chandak, Y., Hashimoto, T. B., and Brunskill, E. (2022). Factored dro: Factored distributionally robust policies for contextual bandits. *Advances in Neural Information Processing Systems*, 35:8318–8331.

Nabi, S., Nassif, H., Hong, J., Mamani, H., and Imbens, G. (2022). Bayesian meta-prior learning using empirical bayes. *Management Science*, 68(3):1737–1755.

Namkoong, H. and Duchi, J. C. (2016). Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Sakhi, O., Faury, L., and Vasile, F. (2020). Improving offline contextual bandits with distributional robustness. *arXiv preprint arXiv:2011.06835*.

Si, N., Zhang, F., Zhou, Z., and Blanchet, J. (2023). Distributionally robust batch contextual bandits. *Management Science*, 69(10):5772–5793.

Uehara, M., Kato, M., and Yasui, S. (2020). Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems*, 33:49–61.

Varatharajah, Y. and Berry, B. (2022). A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8):1277.

Wan, R., Ge, L., and Song, R. (2021). Metadata-based multi-task bandits with bayesian hierarchical models. *Advances in Neural Information Processing Systems*, 34:29655–29668.

Zanker, M., Rook, L., and Jannach, D. (2019). Measuring the impact of online personalisation: Past, present and future. *International Journal of Human-Computer Studies*, 131:160–168.

Zimmert, J. and Seldin, Y. (2019). An optimal algorithm for stochastic and adversarial bandits. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 467–475. PMLR.

A Notation

Table 1 lists every symbol used in the paper. The task parameter ϑ and the warm-start parameter φ are distinct objects and use distinct symbols throughout.

Symbol	Meaning
A	number of arms of each bandit task
M	number of subgroups of the population
K	number of deployment episodes, one task per episode
T	within-episode horizon
$\vartheta \in \mathbb{R}^A$	task parameter, the vector of arm means
$z_k \in [M]$	latent subgroup index of episode k
P_m	task distribution inside subgroup m
$\mathbf{w} \in \Delta_M$	mixture weights over subgroups
$\widehat{P}_m, \widehat{\mathbf{w}}$	nominal references estimated from historical data
$\varphi \in \Phi$	warm-start parameter, the decision variable of DRWS
π_φ	warm-start prior induced by φ
$\Phi \subset \mathbb{R}^d$	admissible set of warm-start parameters
`Alg`	fixed base learner, Thompson sampling in our experiments
$\mathrm{BR}_T(\vartheta; \pi_\varphi)$	Bayes regret of one task over horizon T
$\ell(\vartheta; \varphi)$	differentiable regret proxy
$\mathcal{P}_m, \mathcal{W}$	KL uncertainty sets around $\widehat{P}_m$ and $\widehat{\mathbf{w}}$
δ_m, δ_w	KL radii of $\mathcal{P}_m$ and $\mathcal{W}$
$c_m(\varphi)$	worst-case regret inside subgroup m
$V(\varphi)$	robust objective, worst case over $\mathcal{W}$
$\widehat{c}_m, \widehat{V}$	empirical counterparts of c_m and V
η_m, λ	dual variables of the inner and the outer KL constraint
$K_m, K_{\min}$	samples per subgroup and their minimum
$n, w_{\min}$	total historical tasks and smallest nominal weight
$\xi \in \mathcal{E}$	coordinates of the belief state of the base learner
$\psi_{k,t}$	action distribution at round t of episode k
$\mathcal{H}_{k,t}$	history up to round t of episode k
ϵ_{proxy}	uniform error of the regret proxy
$\epsilon_{\text{disc}}, \epsilon_{\text{num}}$	discretisation and solver error of the belief-ODE proxy

Table 1: Notation for the DRWS formulation, optimisation, and analysis.

B Extended Related Work

Analyses of Bayesian bandit algorithms fix a prior that agrees with the task distribution and bound regret under that prior. Agrawal and Goyal (2012) give such a bound for Thompson sampling, and Russo and Van Roy (2016) give an information-theoretic version. Prior selection is an input to both analyses and not an object of optimisation. Hierarchical and empirical Bayes methods take the next step and learn a shared prior across tasks. Hong et al. (2022) place a hierarchical model over task parameters, Kveton et al. (2021) and Wan et al. (2021) learn a meta-prior from previous tasks, Nabi et al. (2022) study Bayesian meta-learning for bandits, and Basu et al. (2021) analyse meta-learning of exploration. All of these methods target average performance under a fixed meta-distribution, and none of them controls the worst case over a set of plausible deployment mixtures.

Distributionally robust optimisation supplies that worst-case machinery. Bertsimas and Sim (2004) formalise the price of robustness in optimisation, Namkoong and Duchi (2016) give stochastic first-order methods for f-divergence balls, Duchi and Namkoong (2021) and Duchi et al. (2021) develop the statistics of the resulting estimators, and Kuhn et al. (2025) survey the field. Hu et al. (2018) ask when robustness helps, and Sagawa et al. (2019) and Husain et al. (2023) study group-structured uncertainty sets. Kullback–Leibler balls in particular admit the dual form we use, and Hu and Hong (2013) give the corresponding analysis for stochastic programs. The decision object in all of this work is a static predictor or a static decision, so the loss of a candidate decision is evaluated on one data point at a time.

Robust and adversarial bandit work robustifies the online process itself. Bubeck and Slivkins (2012) analyse algorithms that are simultaneously good in the stochastic and the adversarial regime, Zimmert and Seldin (2019) give an optimal best-of-both-worlds algorithm, Herbster et al. (2021) study robustness on graphs, and Avin et al. (2025) treat adversarial corruption. Robust off-policy work optimises a policy under an uncertainty set instead, as in Uehara et al. (2020), Faury et al. (2020), Sakhi et al. (2020), Si et al. (2023), and Mu et al. (2022).

DRWS differs from both groups in the object it robustifies. The adversary acts on the task distribution and not on the reward sequence, so the online algorithm is untouched and needs no recertification. The loss of a candidate warm start is the regret of a whole trajectory of the fixed learner, so the objective reweights trajectories rather than pointwise losses. The nearest conceptual baseline is a barycentre of the subgroup task distributions, which our experiments include. A barycentre ignores the asymmetric value of information under Thompson sampling, and two priors that are close in divergence can therefore induce different exploration and different regret.

C Worked Examples

Example 1 in the main text instantiates the generative model. This section instantiates the remaining objects of Sections 2 and 4.

Example 2 (Thompson-sampling action probabilities under a warm start)**.** *The warm-start prior π_φ initialises the belief trajectory $\pi_{k,t}$, which in turn induces the action distribution $\psi_{k,t}$. Thompson sampling draws $\widetilde{\vartheta}_{k,t} \sim \pi_{k,t}$ and plays a maximising arm, so*

$$\psi_{k,t}(a) = \mathbb{P}_{\vartheta\sim\pi_{k,t}}\big(\vartheta_a \geq \vartheta_b \text{ for all } b \neq a\big).$$

For Beta–Bernoulli bandits with independent Beta marginals this becomes

$$\psi_a(\pi) = \int_0^1 f_a(x;\alpha_a,\beta_a) \prod_{b\neq a} F_b(x;\alpha_b,\beta_b)\,dx, \tag{25}$$

where f_a and F_a are the Beta density and distribution functions. The example shows how a change of π_φ changes exploration, which is the mechanism by which the choice of φ in Section 2 affects cumulative regret.

Example 3 (Belief-ODE proxy, step by step)**.** *Fix a task parameter ϑ, a warm-start prior π_φ, and a horizon T. The proxy of Section 4.2 computes the following steps.*

1. *Set $\pi(0) = \pi_\varphi$ and map it to belief coordinates $\xi_0 = \xi(\pi_\varphi)$.*

2. *Compute the action distribution $\psi_t(a) = \mathbb{P}(A_t = a \mid \pi_t)$ of the base learner at the current belief.*

3. *For each arm a, form the reward distribution $p(\cdot \mid a, \vartheta)$ at the given task parameter.*

4. *For each pair (a, y), apply the Bayes operator $\mathcal{B}(\pi_t, a, y) = p(y \mid a, \vartheta)\pi_t(\vartheta) / \int p(y \mid a, \vartheta')\pi_t(\vartheta')\, d\vartheta'$.*

5. *Average the one-step coordinate change over the action distribution and the reward noise, which gives the drift $\dot{\xi}_t$ of* (20)*.*

6. *Integrate the drift from 0 to T with a numerical solver, which gives the path ξ_t and hence $\pi_t = \pi(\xi_t)$.*

7. *Compute the instantaneous regret rate $r(\xi_t, \vartheta)$ of* (20) *along that path.*

8. *Integrate the regret rate over the horizon, which gives $\widehat{\mathrm{BR}}_T^{\mathrm{ODE}}(\vartheta; \pi_\varphi)$.*

The resulting number is the loss that enters the subgroup costs $\widehat{c}_m(\varphi)$ in (17)*, which connects trajectory-level estimation to the minimax selection of φ.*

D Algorithms

Algorithm 3 is the meta-level template of DRWS, and it leaves two components open on purpose: the design of the uncertainty set, and the regret proxy. Algorithm 2 in the main text instantiates the template with KL balls and the projected stochastic-gradient step (22). Algorithm 1 in the main text states the two regret-estimation modules.

Line 1 of Algorithm 3 turns the historical logs into the nominal objects of (9), one reference per subgroup plus the composition, and nothing in the rest of the template touches the raw logs again. Line 2 fixes how far the deployment population is allowed to move from those nominal objects, which is the only place where the radii enter. Line 3 fixes what a candidate warm start costs on one task, and any estimator of the episode regret may be used here, including exact Monte Carlo rollouts. Line 4 solves the resulting minimax problem, and Line 5 uses the returned prior unchanged for every task of the deployment, which is what keeps the online decision rule untouched.

The belief-ODE estimator. The state Algorithm 1(a) carries through its loop is the belief coordinate ξ and the accumulated regret $\widehat{R}$. Line 1 maps the warm-start prior into those coordinates. Inside the loop, Line 3 reads the action distribution off the current belief, which for Thompson sampling is the posterior probability of each arm being optimal. Line 4 adds the instantaneous regret rate $r(\xi, \vartheta)$ of (20) over one step of length Δt, Line 5 averages the one-step Bayes update over the action and reward randomness, which is the drift of (20), and Line 6 takes an explicit Euler step on that drift. The loop therefore integrates (21) along a single deterministic trajectory rather than averaging over sampled trajectories, which is where the cost saving over rollouts comes from, and the price is the discretisation error ϵ_{disc} and the solver error ϵ_{num} that Theorem 5 accounts for.

The neural surrogate. The first loop of Algorithm 1(b) builds a labelled dataset. Line 2 samples a prior and a task independently, so the surrogate is trained on the whole product space rather than on the priors that one particular optimisation run happens to visit, and Lines 3–4 estimate the label by S independent rollouts of the base learner, so the label is an unbiased estimate of (18) with variance of order $1/S$. The second stage fits f_ϕ by stochastic gradient descent on the squared error, after which one proxy evaluation is a single forward pass. Both stages run once, offline, and their cost is amortised over every subsequent robust optimisation run.

Algorithm 3 DRWS: meta-level template

Require: subgroup data or simulators $\{\mathcal{D}_m\}_{m=1}^M$; radii $\{\delta_m\}_{m=1}^M$ and δ_w; horizon T; admissible set Φ

1: fit a nominal reference $\widehat{P}_m$ from $\mathcal{D}_m$ for every $m \in [M]$, and a nominal composition $\widehat{\mathbf{w}}$
2: define the uncertainty sets $\mathcal{P}_m$ and $\mathcal{W}$ as in (10)
3: choose a regret proxy $\ell(\vartheta; \varphi)$ for the episode regret of the base learner started at π_φ
4: solve $\varphi^\star \in \arg\min_{\varphi \in \Phi} \sup_{\mathbf{w} \in \mathcal{W}} \sum_{m=1}^M w_m \sup_{Q \in \mathcal{P}_m} \mathbb{E}_Q\big[\ell(\vartheta; \varphi)\big]$
5: **at deployment:** for every new task, run the base learner for T rounds started at $\pi_{\varphi^\star}$

E Proofs

Throughout this appendix, $\varphi \in \Phi$ is a warm-start parameter, $\ell(\vartheta; \varphi)$ is the task-level regret proxy, and $\widehat{P}_m$ and $\widehat{w}_m$ are the nominal references. We write

$$c_m(\varphi) := \sup_{Q: D_{\mathrm{KL}}(Q \| \widehat{P}_m) \le \delta_m} \mathbb{E}_Q\big[\ell(\vartheta; \varphi)\big], \qquad V(\varphi) := \sup_{\mathbf{w}: D_{\mathrm{KL}}(\mathbf{w} \| \widehat{\mathbf{w}}) \le \delta_w} \sum_{m=1}^M w_m c_m(\varphi),$$

and $\widehat{V}$ replaces each subgroup expectation by an average over K_m samples.

E.1 Kullback–Leibler Duality and Exponential Tilting

Proof of Lemma 1. The Donsker–Varadhan variational formula states that, for a reference distribution P, a bounded measurable h, and a radius $\delta > 0$,

$$\sup_{Q: D_{\mathrm{KL}}(Q \| P) \le \delta} \mathbb{E}_Q[h] = \inf_{\eta > 0} \big\{ \eta \delta + \eta \log \mathbb{E}_P\big[\exp(h/\eta)\big] \big\}. \tag{26}$$

We apply (26) twice, once inside each subgroup and once across subgroups.

Step 1: the inner supremum. Fix $m \in [M]$ and $\varphi \in \Phi$. Assumption 1 gives $0 \le \ell(\cdot; \varphi) \le L_{\max}$, so $\ell(\cdot; \varphi)$ is bounded and measurable and (26) applies with $P = P_m^0$, $h = \ell(\cdot; \varphi)$, and $\delta = \delta_m$. Hence

$$\begin{aligned} c_m(\varphi) &= \sup_{Q: D_{\mathrm{KL}}(Q \| P_m^0) \le \delta_m} \mathbb{E}_Q\big[\ell(\vartheta; \varphi)\big] && \text{by the definition of } c_m \\ &= \inf_{\eta_m > 0} \Big\{ \eta_m \delta_m + \eta_m \log \mathbb{E}_{\vartheta \sim P_m^0}\big[\exp\big(\ell(\vartheta; \varphi)/\eta_m\big)\big] \Big\} && \text{by (26)}, \end{aligned} \tag{27}$$

which is (14). Monotonicity of the expectation in the last display also gives $0 \le c_m(\varphi) \le L_{\max}$, since $\mathbb{E}_Q[\ell]$ lies in $[0, L_{\max}]$ for every feasible Q.

Step 2: the outer supremum. Let $\mathbf{c}(\varphi) := (c_1(\varphi), \dots, c_M(\varphi))$. The map $\mathbf{w} \mapsto \sum_{m=1}^M w_m c_m(\varphi)$ is the expectation of the function $m \mapsto c_m(\varphi)$ on the finite space $[M]$ under the probability vector $\mathbf{w}$, and Step 1 showed that this function is bounded by $L_{\max}$. Applying (26) with $P = \mathbf{w}^0$, $h = c_{(\cdot)}(\varphi)$, and $\delta = \delta_w$ therefore gives

$$\begin{aligned} V(\varphi) &= \sup_{\mathbf{w}: D_{\mathrm{KL}}(\mathbf{w} \| \mathbf{w}^0) \le \delta_w} \sum_{m=1}^M w_m c_m(\varphi) && \text{by the definition of } V \\ &= \inf_{\lambda > 0} \Big\{ \lambda \delta_w + \lambda \log \sum_{m=1}^M w_m^0 \exp\big(c_m(\varphi)/\lambda\big) \Big\} && \text{by (26)}, \end{aligned} \tag{28}$$

which is (15). Displays (27) and (28) are the two claims of the lemma. □

The optimiser of (26) is an exponential tilting of the reference. Differentiating the inner objective and using the first-order condition gives

$$\frac{dQ_m^\star}{d\widehat{P}_m}(\vartheta) = \frac{\exp\big(\ell(\vartheta; \varphi)/\eta_m^\star\big)}{\mathbb{E}_{\widehat{P}_m}\big[\exp\big(\ell(\vartheta; \varphi)/\eta_m^\star\big)\big]}, \qquad w_m^\star = \frac{\widehat{w}_m \exp\big(c_m(\varphi)/\lambda^\star\big)}{\sum_{j=1}^M \widehat{w}_j \exp\big(c_j(\varphi)/\lambda^\star\big)},$$

which are the tilted worst cases quoted in Section 4.1.

Remark 1 (Convexity of the dual problem)**.** *For every fixed φ, the dual objectives* (14) *and* (15) *are convex in η_m and in λ, because both are perspective transforms of a log-sum-exp function. The duality applies to the inner maximisation over distributions and not to the outer minimisation over warm starts. Non-convexity enters only through the dependence of $\ell(\vartheta;\varphi)$ on the belief trajectory that π_φ induces, so the outer problem is a smooth non-convex problem and Algorithm 2 uses several restarts.*

The nested construction agrees with the KL chain rule (13), so a budget on the composition and a budget on the conditional distributions are the two terms of one divergence budget.

E.2 Properties of the Dual Variables

We assume each radius stays below the value at which the adversary could place all of its mass on a maximiser of the proxy. Equivalently, the optimal temperatures stay bounded away from zero, so that $\eta_m^\star(\varphi) \geq \underline{\eta}_m > 0$ and $\lambda^\star(\varphi) \geq \underline{\lambda} > 0$ for all m and all φ. This condition holds automatically for small radii when the reference has continuous support and the proxy is bounded and non-constant. For a reference with finite support it holds whenever the radius is smaller than the KL cost of concentrating on the support point of maximal proxy value.

Boundedness of the proxy also bounds the dual optimisers from above.

Lemma 2 (Upper bounds on the dual optimisers)**.** *Under Assumption 1, every minimiser of the inner dual objective satisfies $\eta_m^\star(\varphi) \leq 2L_{\max}/\delta_m$, and every minimiser of the outer dual objective satisfies $\lambda^\star(\varphi) \leq 2L_{\max}/\delta_w$.*

Proof. Write $g_m(\eta) := \eta\delta_m + \eta \log \mathbb{E}_{\widehat{P}_m}[\exp(\ell(\vartheta;\varphi)/\eta)]$ for the inner dual objective, and evaluate it at the particular point $\eta_0 := L_{\max}/\delta_m$. Since $\ell \leq L_{\max}$ pointwise,

$$\log \mathbb{E}_{\widehat{P}_m}\big[\exp(\ell/\eta_0)\big] \leq \log \exp\big(L_{\max}/\eta_0\big) = \frac{L_{\max}}{\eta_0}, \tag{29}$$

and therefore

$$g_m(\eta_0) \leq \eta_0\delta_m + \eta_0 \cdot \frac{L_{\max}}{\eta_0} = L_{\max} + L_{\max} = 2L_{\max}. \tag{30}$$

On the other hand, $\log \mathbb{E}[\exp(\ell/\eta)] \geq 0$ because $\ell \geq 0$, so $g_m(\eta) \geq \eta\delta_m$ for every $\eta > 0$. Any $\eta > 2L_{\max}/\delta_m$ hence satisfies $g_m(\eta) > 2L_{\max} \geq g_m(\eta_0)$ by (30), so such an η cannot be a minimiser, and every minimiser lies in $(0, 2L_{\max}/\delta_m]$. Step 1 of the proof of Lemma 1 gives $0 \leq c_m(\varphi) \leq L_{\max}$, so the identical argument applied to the outer dual objective with δ_w in place of δ_m gives $\lambda^\star(\varphi) \in (0, 2L_{\max}/\delta_w]$. □

Together with the lower bounds above, Lemma 2 places both dual optimisers in compact intervals bounded away from zero, which is what Assumption 2 requires.

E.3 Regularity of the Robust Objective

Lemma 3 (Lipschitz continuity of the robust objective)**.** *Under Assumption 1, both the subgroup costs and the robust objective are L_Φ-Lipschitz on Φ, so that $|c_m(\varphi) - c_m(\varphi')| \leq L_\Phi\|\varphi - \varphi'\|$ for every m and $|V(\varphi) - V(\varphi')| \leq L_\Phi\|\varphi - \varphi'\|$.*

Proof. Fix $m \in [M]$ and $\varphi, \varphi' \in \Phi$.

Step 1: the subgroup costs. For any feasible $Q \in \mathcal{P}_m$,

$$\begin{aligned} \big|\mathbb{E}_Q\big[\ell(\vartheta;\varphi)\big] - \mathbb{E}_Q\big[\ell(\vartheta;\varphi')\big]\big| &\leq \mathbb{E}_Q\big|\ell(\vartheta;\varphi) - \ell(\vartheta;\varphi')\big| && \text{by Jensen's inequality} \\ &\leq L_\Phi\|\varphi - \varphi'\| && \text{by Assumption 1.} \end{aligned} \tag{31}$$

Bound (31) holds uniformly over $Q \in \mathcal{P}_m$, so

$$c_m(\varphi) = \sup_{Q\in\mathcal{P}_m} \mathbb{E}_Q\big[\ell(\vartheta;\varphi)\big] \leq \sup_{Q\in\mathcal{P}_m} \Big\{\mathbb{E}_Q\big[\ell(\vartheta;\varphi')\big] + L_\Phi\|\varphi - \varphi'\|\Big\} = c_m(\varphi') + L_\Phi\|\varphi - \varphi'\|. \tag{32}$$

Exchanging the roles of φ and φ' gives $|c_m(\varphi) - c_m(\varphi')| \leq L_\Phi\|\varphi - \varphi'\|$, which is the first claim.

Step 2: the robust objective. For any feasible $\mathbf{w} \in \mathcal{W}$, Step 1 gives

$$\sum_{m=1}^{M} w_m c_m(\varphi) \le \sum_{m=1}^{M} w_m c_m(\varphi') + L_\Phi \|\varphi - \varphi'\| \sum_{m=1}^{M} w_m = \sum_{m=1}^{M} w_m c_m(\varphi') + L_\Phi \|\varphi - \varphi'\|, \tag{33}$$

where the last equality uses $\mathbf{w} \in \Delta_M$. Taking the supremum over $\mathbf{w} \in \mathcal{W}$ on both sides yields $V(\varphi) \le V(\varphi') + L_\Phi \|\varphi - \varphi'\|$, and exchanging φ and φ' gives the second claim. □

Lemma 3 converts control on a finite net into control on all of Φ. If Φ_ε is an ε-net of Φ, then a uniform bound on Φ_ε extends to Φ at an additional cost of $2L_\Phi\varepsilon$, which is the $O(\varepsilon)$ term of Theorem 1.

E.4 Uniform Convergence

Theorem 3 (Uniform convergence in the total sample size)**.** *Suppose Assumptions 1 and 2 hold. Then, for every $\varepsilon > 0$ and every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{\varphi \in \Phi} \left| \widehat{V}(\varphi) - V(\varphi) \right| \le C \sqrt{\frac{\log \mathcal{N}(\varepsilon, \Phi) + \log(M/\delta)}{K_{\min}}} + O(\varepsilon).$$

If in addition n tasks are drawn from the nominal mixture and $\widehat{w}_m \ge w_{\min} > 0$ for every m, then $K_{\min} \gtrsim n w_{\min}$ with high probability and the same bound holds with $K_{\min}$ replaced by $n w_{\min}$.

Proof. We prove the claim in four steps.

Step 1: concentration of the exponential moments. Fix $m \in [M]$, $\varphi \in \Phi$, and $\eta_m \in [\eta_{\min}, \eta_{\max}]$, and set $Z_m(\vartheta) := \exp(\ell(\vartheta;\varphi)/\eta_m)$. Assumption 1 and $\eta_m \ge \eta_{\min}$ give $1 \le Z_m \le \exp(L_{\max}/\eta_{\min}) =: B_Z$. Hoeffding's inequality for bounded variables, a union bound over the M subgroups, and a union bound over an ε-net Φ_ε of Φ therefore give, with probability at least $1-\delta$,

$$\max_{m \in [M]} \sup_{\varphi \in \Phi_\varepsilon} \left| \frac{1}{K_m} \sum_{j=1}^{K_m} Z_m(\vartheta_{m,j}) - \mathbb{E}_{\widehat{P}_m}[Z_m] \right| \le B_Z \sqrt{\frac{\log \mathcal{N}(\varepsilon, \Phi) + \log(M/\delta)}{K_{\min}}}.$$

Write $\varrho := B_Z \sqrt{(\log \mathcal{N}(\varepsilon, \Phi) + \log(M/\delta))/K_{\min}}$ for that right-hand side, and let $\mathcal{E}$ denote the event on which it holds.

Step 2: transfer to the inner dual value. Work on $\mathcal{E}$ and fix m, $\varphi \in \Phi_\varepsilon$ and $\eta_m \in [\eta_{\min}, \eta_{\max}]$. Write $\widehat{Z} := K_m^{-1} \sum_j Z_m(\vartheta_{m,j})$ and $Z := \mathbb{E}_{\widehat{P}_m}[Z_m]$, both of which lie in $[1, B_Z]$. On that interval $z \mapsto \eta_m \log z$ has derivative $\eta_m / z \le \eta_{\max}$, so

$$\begin{aligned} \left| \eta_m \log \widehat{Z} - \eta_m \log Z \right| &\le \eta_{\max} \left| \widehat{Z} - Z \right| && \text{by the mean value theorem} \\ &\le \eta_{\max}\, \varrho && \text{by Step 1.} \end{aligned} \tag{34}$$

Bound (34) is uniform in η_m, and for any two functions f, g on a common domain one has $|\inf f - \inf g| \le \sup |f - g|$. Applying that inequality to the two dual objectives, whose difference is exactly the left-hand side of (34), gives

$$\max_{m \in [M]} \sup_{\varphi \in \Phi_\varepsilon} \left| \widehat{c}_m(\varphi) - c_m(\varphi) \right| \le \eta_{\max}\, \varrho. \tag{35}$$

Step 3: transfer to the outer dual value. Fix $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ and define $G_\lambda(\mathbf{c}) := \lambda \delta_w + \lambda \log \sum_m \widehat{w}_m \exp(c_m/\lambda)$ on $[0, L_{\max}]^M$. Its partial derivatives are

$$\frac{\partial G_\lambda}{\partial c_m}(\mathbf{c}) = \frac{\widehat{w}_m \exp(c_m/\lambda)}{\sum_j \widehat{w}_j \exp(c_j/\lambda)} = w_m^\star, \tag{36}$$

which are non-negative and sum to one, so $\|\nabla G_\lambda\|_1 = 1$ and G_λ is 1-Lipschitz in $\|\cdot\|_\infty$. Combining this with (35) and again using $|\inf f - \inf g| \le \sup |f - g|$, now for the infimum over λ, gives

$$\sup_{\varphi \in \Phi_\varepsilon} \left| \widehat{V}(\varphi) - V(\varphi) \right| \le \left\| \widehat{\mathbf{c}}(\varphi) - \mathbf{c}(\varphi) \right\|_\infty \le \eta_{\max}\, \varrho. \tag{37}$$

Step 4: from the net to Φ. Let $\varphi \in \Phi$ be arbitrary and let $\varphi_\varepsilon \in \Phi_\varepsilon$ satisfy $\|\varphi - \varphi_\varepsilon\| \le \varepsilon$. Lemma 3 applies to V and, with the same argument applied to the empirical references, to $\widehat{V}$, so

$$\begin{aligned}\big|\widehat{V}(\varphi) - V(\varphi)\big| &\le \big|\widehat{V}(\varphi) - \widehat{V}(\varphi_\varepsilon)\big| + \big|\widehat{V}(\varphi_\varepsilon) - V(\varphi_\varepsilon)\big| + \big|V(\varphi_\varepsilon) - V(\varphi)\big| \\ &\le L_\Phi \varepsilon + \eta_{\max}\, \varrho + L_\Phi \varepsilon = \eta_{\max}\, \varrho + 2L_\Phi \varepsilon, \end{aligned} \tag{38}$$

where the middle term uses (37). Taking the supremum over $\varphi \in \Phi$ and collecting the constants of Steps 1 to 3 into $C = C(L_{\max}, L_\Phi, \eta_{\min}, \eta_{\max}, \lambda_{\min}, \lambda_{\max})$ gives the first display of the theorem.

Step 5: from $K_{\min}$ to the total sample size. Suppose the n reference tasks are drawn from the nominal mixture, so that the subgroup count K_m is binomial with n trials and success probability $\widehat{w}_m \ge w_{\min}$. A multiplicative Chernoff bound gives, for each m,

$$\mathbb{P}\big(K_m < n w_{\min}/2\big) \le \exp\big(-n w_{\min}/8\big), \tag{39}$$

so a union bound over $m \in [M]$ gives $K_{\min} \ge n w_{\min}/2$ with probability at least $1 - M\exp(-n w_{\min}/8)$. Substituting this bound for $K_{\min}$ in the first display, and absorbing the factor $\sqrt{2}$ into C, gives the second display. □

Theorem 1 in the main text is the first display of Theorem 3.

E.5 Sample Complexity

Theorem 3 converts directly into a sample requirement. To reach $\sup_{\varphi\in\Phi} |\widehat{V}(\varphi) - V(\varphi)| \le \alpha$ with probability at least $1 - \delta$, up to the covering term $O(\varepsilon)$, it suffices that

$$K_{\min} \gtrsim \frac{C^2}{\alpha^2}\left(\log \mathcal{N}(\varepsilon, \Phi) + \log\frac{M}{\delta}\right).$$

If $\Phi \subset \mathbb{R}^d$ has diameter B, then $\log \mathcal{N}(\varepsilon, \Phi) \le d \log(3B/\varepsilon)$, so the requirement becomes

$$K_{\min} \gtrsim \frac{C^2}{\alpha^2}\left(d \log\frac{B}{\varepsilon} + \log\frac{M}{\delta}\right).$$

With n tasks drawn from the nominal mixture and smallest nominal weight $w_{\min}$, replace $K_{\min}$ by $n w_{\min}$ up to constants. The dependence on d is the usual parametric price for the size of the warm-start class, and the dependence on $w_{\min}$ is the price for insisting on worst-case performance across every subgroup.

E.6 Excess Robust Regret

Proof of Theorem 2. Write V^{proxy} for the population robust objective built from the proxy and $\widehat{V}$ for its empirical version. Assumption 3 gives $\sup_\varphi |V^{\mathrm{proxy}}(\varphi) - V^{\mathrm{true}}(\varphi)| \le \epsilon_{\mathrm{proxy}}$, and Theorem 1 gives $\sup_\varphi |\widehat{V}(\varphi) - V^{\mathrm{proxy}}(\varphi)| \le \epsilon_{\mathrm{est}}$ on an event of probability at least $1 - \delta$, where ϵ_{est} denotes the right-hand side of (23). On that event,

$$\begin{aligned} V^{\mathrm{true}}(\widehat{\varphi}) &\le V^{\mathrm{proxy}}(\widehat{\varphi}) + \epsilon_{\mathrm{proxy}} && \text{by Assumption 3} && (40)\\ &\le \widehat{V}(\widehat{\varphi}) + \epsilon_{\mathrm{proxy}} + \epsilon_{\mathrm{est}} && \text{by Theorem 1} && (41)\\ &\le \widehat{V}(\varphi^\star) + \epsilon_{\mathrm{proxy}} + \epsilon_{\mathrm{est}} && \text{by optimality of } \widehat{\varphi} \text{ for } \widehat{V} && (42)\\ &\le V^{\mathrm{proxy}}(\varphi^\star) + \epsilon_{\mathrm{proxy}} + 2\epsilon_{\mathrm{est}} && \text{by Theorem 1} && (43)\\ &\le V^{\mathrm{true}}(\varphi^\star) + 2\epsilon_{\mathrm{proxy}} + 2\epsilon_{\mathrm{est}} && \text{by Assumption 3.} && (44) \end{aligned}$$

Each inequality holds on the same event of probability at least $1 - \delta$, so the chain gives

$$V^{\mathrm{true}}(\widehat{\varphi}) - V^{\mathrm{true}}(\varphi^\star) \le 2\epsilon_{\mathrm{proxy}} + 2\epsilon_{\mathrm{est}}, \tag{45}$$

and substituting the expression for ϵ_{est} from (23) gives the stated bound. □

The two error sources enter with the same factor of two, and neither can be traded for the other. The estimation term shrinks with the number of historical tasks, and the proxy term does not.

E.7 Gain over a Nominal Warm Start

For the gain statements, write $L_\varphi(\vartheta) := \ell(\vartheta;\varphi)$ and define the subgroup mean and the subgroup standard deviation of the loss under the nominal reference,

$$\mu_m(\varphi) := \mathbb{E}_{\widehat{P}_m}\big[L_\varphi(\vartheta)\big], \qquad \sigma_m^2(\varphi) := \mathrm{Var}_{\widehat{P}_m}\big(L_\varphi(\vartheta)\big).$$

Theorem 4 (Small-radius expansion of the robust objective)**.** *Suppose L_φ is bounded and non-constant under every $\widehat{P}_m$. Then, as* $\max\{\delta_w, \delta_1, \ldots, \delta_M\} \downarrow 0$*,*

$$V(\varphi) = \underbrace{\sum_{m=1}^M \widehat{w}_m \mu_m(\varphi)}_{\textit{nominal average regret}} + \underbrace{\sum_{m=1}^M \widehat{w}_m \sqrt{2\delta_m}\, \sigma_m(\varphi)}_{\textit{within-subgroup dispersion}} + \underbrace{\sqrt{2\delta_w\, \mathrm{Var}_{m\sim\widehat{w}}\big(\mu_m(\varphi)\big)}}_{\textit{composition sensitivity}} + O\Big(\delta_w + \sum_{m=1}^M \widehat{w}_m \delta_m\Big).$$

Proof. We expand the inner dual first and then the outer dual.

Step 1: the subgroup cost at a small radius. Fix $m \in [M]$. Since L_φ is bounded, its cumulant generating function admits the expansion

$$\log \mathbb{E}_{\widehat{P}_m}\big[\exp(L_\varphi/\eta_m)\big] = \frac{\mu_m(\varphi)}{\eta_m} + \frac{\sigma_m^2(\varphi)}{2\eta_m^2} + O(\eta_m^{-3}) \qquad \text{as } \eta_m \to \infty. \tag{46}$$

Substituting (46) into the inner dual (14) gives

$$\begin{aligned} c_m(\varphi) &= \inf_{\eta_m>0}\Big\{\eta_m\delta_m + \mu_m(\varphi) + \frac{\sigma_m^2(\varphi)}{2\eta_m} + O(\eta_m^{-2})\Big\} \\ &= \mu_m(\varphi) + \inf_{\eta_m>0}\Big\{\eta_m\delta_m + \frac{\sigma_m^2(\varphi)}{2\eta_m}\Big\} + O(\delta_m), \end{aligned} \tag{47}$$

where the remainder is $O(\delta_m)$ because the minimiser below scales as $\delta_m^{-1/2}$. The inner minimisation in (47) is a scalar problem of the form $\inf_{\eta>0}\{a\eta + b/\eta\}$ with $a = \delta_m$ and $b = \sigma_m^2(\varphi)/2$, whose minimiser is $\eta_m^\star = \sqrt{b/a} = \sigma_m(\varphi)/\sqrt{2\delta_m}$ and whose value is $2\sqrt{ab} = \sqrt{2\delta_m}\,\sigma_m(\varphi)$. Hence

$$c_m(\varphi) = \mu_m(\varphi) + \sqrt{2\delta_m}\,\sigma_m(\varphi) + O(\delta_m). \tag{48}$$

Step 2: the outer objective at a small radius. The outer dual (15) has the same form, with the finite reference $\widehat{\mathbf{w}}$ in place of $\widehat{P}_m$ and the bounded loss $m \mapsto c_m(\varphi)$ in place of L_φ. Repeating Step 1 verbatim therefore gives

$$V(\varphi) = \sum_{m=1}^M \widehat{w}_m c_m(\varphi) + \sqrt{2\delta_w\, \mathrm{Var}_{m\sim\widehat{w}}\big(c_m(\varphi)\big)} + O(\delta_w). \tag{49}$$

Step 3: combining the two expansions. Substituting (48) into the first term of (49) gives

$$\sum_{m=1}^M \widehat{w}_m c_m(\varphi) = \sum_{m=1}^M \widehat{w}_m \mu_m(\varphi) + \sum_{m=1}^M \widehat{w}_m \sqrt{2\delta_m}\,\sigma_m(\varphi) + O\Big(\sum_m \widehat{w}_m \delta_m\Big). \tag{50}$$

For the second term of (49), (48) gives $c_m(\varphi) = \mu_m(\varphi) + O(\max_m \sqrt{\delta_m})$, and the variance is Lipschitz in the sup norm of its argument on a bounded range, so

$$\mathrm{Var}_{m\sim\widehat{w}}\big(c_m(\varphi)\big) = \mathrm{Var}_{m\sim\widehat{w}}\big(\mu_m(\varphi)\big) + O\big(\max_m \sqrt{\delta_m}\big). \tag{51}$$

Substituting both displays into (49) and retaining the terms of first order in the radii proves the claim. □

Corollary 1 (Gain over a nominal warm start)**.** *Let φ_{nom} minimise the nominal average regret $\sum_m \widehat{w}_m \mu_m(\varphi)$, let φ_{rob} minimise the robust objective V of* (11)*, and define the robust suboptimality gap of the nominal warm start by $\Gamma := V(\varphi_{\mathrm{nom}}) - \inf_{\varphi\in\Phi} V(\varphi) \geq 0$. If an empirical robust warm start $\widehat{\varphi}_{\mathrm{rob}}$ satisfies $V(\widehat{\varphi}_{\mathrm{rob}}) - \inf_{\varphi\in\Phi} V(\varphi) \leq \epsilon_{\mathrm{stat}}$, then $V(\widehat{\varphi}_{\mathrm{rob}}) \leq V(\varphi_{\mathrm{nom}}) - \Gamma + \epsilon_{\mathrm{stat}}$. In particular, $V(\varphi_{\mathrm{rob}}) = V(\varphi_{\mathrm{nom}}) - \Gamma$.*

Proof of Corollary 1. By the definition of the robust suboptimality gap,

$$\inf_{\varphi\in\Phi} V(\varphi) = V(\varphi_{\mathrm{nom}}) - \Gamma. \tag{52}$$

Since φ_{rob} attains that infimum, (52) gives $V(\varphi_{\mathrm{rob}}) = V(\varphi_{\mathrm{nom}}) - \Gamma$, which is the second claim. For the first claim, the assumed excess robust risk of $\widehat{\varphi}_{\mathrm{rob}}$ and (52) give

$$\begin{aligned} V(\widehat{\varphi}_{\mathrm{rob}}) &\leq \inf_{\varphi\in\Phi} V(\varphi) + \epsilon_{\mathrm{stat}} && \text{by hypothesis} \\ &= V(\varphi_{\mathrm{nom}}) - \Gamma + \epsilon_{\mathrm{stat}} && \text{by (52).} \end{aligned} \tag{53}$$

Whenever $\Gamma > \epsilon_{\mathrm{stat}}$, the right-hand side is strictly smaller than $V(\varphi_{\mathrm{nom}})$. □

Theorem 4 identifies when Γ is strictly positive. For small radii the robust objective adds two penalties to the nominal average regret. A warm start with a slightly larger nominal average regret can therefore have a strictly smaller robust objective, provided it reduces the within-subgroup dispersion $\sigma_m(\varphi)$ or the across-subgroup variance of $\mu_m(\varphi)$. Corollary 1 then gives a strict improvement over the nominal warm start whenever $\Gamma > \epsilon_{\mathrm{stat}}$, and Theorem 2 bounds ϵ_{stat}.

E.8 Belief-ODE Approximation

This subsection proves the two results on the belief-ODE proxy of Section 4.2 that Section 5 quotes. Throughout, ξ_t is the belief trajectory of the flow (20), $F(\cdot, \vartheta)$ is its drift, and $r(\cdot, \vartheta)$ is the instantaneous regret rate.

Assumption 4 (Belief-ODE regularity)**.** *The belief state of the base learner is represented by coordinates $\xi \in \mathcal{E}$, where $\mathcal{E} \subset \mathbb{R}^d$ is compact. The action map $\psi : \mathcal{E} \to \Delta_A$ is L_ψ-Lipschitz, the belief drift $F(\cdot, \vartheta) : \mathcal{E} \to \mathbb{R}^d$ of* (20) *is L_F-Lipschitz for every task parameter ϑ, the instantaneous regret rate $r(\cdot, \vartheta) : \mathcal{E} \to \mathbb{R}$ is L_r-Lipschitz, and rewards are bounded by $R_{\max}$.*

Proposition 1 (Stability of the belief-ODE proxy in its initialisation)**.** *Suppose Assumption 4 holds. Then, for any two initial belief coordinates $\xi_0, \xi_0' \in \mathcal{E}$ and any task parameter ϑ,*

$$\left|\widehat{\ell}_{\mathrm{ODE}}(\vartheta; \xi_0) - \widehat{\ell}_{\mathrm{ODE}}(\vartheta; \xi_0')\right| \leq 2R_{\max} L_\psi \, \frac{e^{L_F T} - 1}{L_F} \, \|\xi_0 - \xi_0'\|, \tag{54}$$

with the convention $(e^{L_F T} - 1)/L_F = T$ when $L_F = 0$.

Theorem 5 (Approximation error of the belief-ODE proxy)**.** *Suppose Assumption 4 holds, that the discrete Bayesian update has local ODE approximation error at most ϵ_{disc}, and that the numerical solver has global error at most ϵ_{num}. Then, for every task parameter ϑ and every warm start π_φ,*

$$\left|R_T^{\mathrm{ODE}}(\vartheta; \pi_\varphi) - R_T^{\mathrm{true}}(\vartheta; \pi_\varphi)\right| \leq L_r T \left(\frac{e^{L_F T} - 1}{L_F} \, \epsilon_{\mathrm{disc}} + \epsilon_{\mathrm{num}} \right). \tag{55}$$

Proof of Proposition 1. We first bound how far two belief trajectories can separate, and then convert that into a bound on the integrated regret rate.

Step 1: separation of the belief trajectories. Let ξ_t and ξ_t' solve the drift equation in (20) for the same task parameter ϑ, from initial values ξ_0 and ξ_0'. Writing F for the drift, their difference satisfies

$$\frac{d}{dt}\big(\xi_t - \xi_t'\big) = F(\xi_t, \vartheta) - F(\xi_t', \vartheta), \tag{56}$$

so that, in integral form and by the triangle inequality,

$$\begin{aligned} \big\|\xi_t - \xi_t'\big\| &\leq \big\|\xi_0 - \xi_0'\big\| + \int_0^t \big\|F(\xi_u, \vartheta) - F(\xi_u', \vartheta)\big\| \, du \\ &\leq \big\|\xi_0 - \xi_0'\big\| + L_F \int_0^t \big\|\xi_u - \xi_u'\big\| \, du && \text{by Assumption 4.} \end{aligned} \tag{57}$$

Grönwall's inequality applied to the last display yields

$$\big\|\xi_t - \xi'_t\big\| \le e^{L_F t}\big\|\xi_0 - \xi'_0\big\| \qquad \text{for all } t \in [0,T]. \tag{58}$$

Step 2: from belief separation to regret separation. The instantaneous regret rate is $r(\xi,\vartheta) = \sum_a \psi_a(\xi)\big(\mu^\star(\vartheta) - \mu_a(\vartheta)\big)$. The gaps satisfy $|\mu^\star(\vartheta) - \mu_a(\vartheta)| \le 2R_{\max}$ and ψ is L_ψ-Lipschitz by Assumption 4, so

$$\big|r(\xi_t,\vartheta) - r(\xi'_t,\vartheta)\big| \le 2R_{\max}\big\|\psi(\xi_t) - \psi(\xi'_t)\big\|_1 \le 2R_{\max}L_\psi\big\|\xi_t - \xi'_t\big\|. \tag{59}$$

Step 3: integrating over the horizon. Combining (59) with (58) and integrating the proxy (21) over $[0,T]$ gives

$$\begin{aligned}\big|\widehat{\ell}_{\mathrm{ODE}}(\vartheta;\xi_0) - \widehat{\ell}_{\mathrm{ODE}}(\vartheta;\xi'_0)\big| &\le \int_0^T \big|r(\xi_t,\vartheta) - r(\xi'_t,\vartheta)\big|\,dt \\ &\le 2R_{\max}L_\psi\big\|\xi_0 - \xi'_0\big\| \int_0^T e^{L_F t}\,dt \\ &= 2R_{\max}L_\psi\,\frac{e^{L_F T}-1}{L_F}\,\big\|\xi_0 - \xi'_0\big\|, \end{aligned} \tag{60}$$

which is the claimed bound. When $L_F = 0$ the integral equals T, which gives the stated limiting value. □

Proof of Theorem 5. Three trajectories appear in the argument: $\bar{\xi}_t$, the expected belief trajectory of the discrete Bayesian update; ξ_t^{ODE}, the exact solution of the limiting flow (20); and ξ_t^{num}, the numerical solution the proxy actually computes. We bound the two gaps in turn and then integrate.

Step 1: discretisation gap. By assumption the discrete update deviates from the flow by at most ϵ_{disc} per unit time, so the difference $e_t := \bar{\xi}_t - \xi_t^{\mathrm{ODE}}$ satisfies

$$\|e_t\| \le \int_0^t \Big(L_F\|e_u\| + \epsilon_{\mathrm{disc}}\Big)\,du, \tag{61}$$

since both trajectories start at ξ_0 and F is L_F-Lipschitz. Grönwall's inequality in its integral form therefore gives

$$\sup_{t\le T}\big\|\bar{\xi}_t - \xi_t^{\mathrm{ODE}}\big\| \le \frac{e^{L_F T}-1}{L_F}\,\epsilon_{\mathrm{disc}}. \tag{62}$$

Step 2: solver gap and the combined bound. The solver assumption gives $\sup_{t\le T}\|\xi_t^{\mathrm{ODE}} - \xi_t^{\mathrm{num}}\| \le \epsilon_{\mathrm{num}}$, so the triangle inequality with (62) yields

$$\sup_{t\le T}\big\|\bar{\xi}_t - \xi_t^{\mathrm{num}}\big\| \le \frac{e^{L_F T}-1}{L_F}\,\epsilon_{\mathrm{disc}} + \epsilon_{\mathrm{num}}. \tag{63}$$

Step 3: from belief error to regret error. The instantaneous regret rate is L_r-Lipschitz in the belief coordinate by Assumption 4, and both cumulative regrets integrate their rate over a horizon of length T, so

$$\begin{aligned}\big|R_T^{\mathrm{ODE}}(\vartheta;\pi_\varphi) - R_T^{\mathrm{true}}(\vartheta;\pi_\varphi)\big| &\le \int_0^T \big|r(\xi_t^{\mathrm{num}},\vartheta) - r(\bar{\xi}_t,\vartheta)\big|\,dt \\ &\le L_r\int_0^T \big\|\xi_t^{\mathrm{num}} - \bar{\xi}_t\big\|\,dt \\ &\le L_r T\left(\frac{e^{L_F T}-1}{L_F}\,\epsilon_{\mathrm{disc}} + \epsilon_{\mathrm{num}}\right), \end{aligned} \tag{64}$$

where the last step uses (63). This is the claimed bound, and taking the supremum over (ϑ,φ) makes it uniform, which is the form Assumption 3 requires. □

Table 2: Data-generating process of the two synthetic environments. Both use three subgroups, five arms, and nominal weights $(0.5, 0.3, 0.2)$.

Component	Gaussian environment	Beta–Bernoulli environment
Task parameter	Arm means are real valued and drawn from subgroup-specific Gaussians.	Arm means are probabilities in $(0, 1)$, with Beta-distributed reward parameters.
Subgroup centres	$(0.90, 0.15, 0.05, -0.10, -0.15)$, $(0.10, 0.85, 0.00, -0.20, -0.25)$, $(0.00, 0.10, 0.80, -0.15, -0.20)$.	$(0.75, 0.40, 0.35, 0.30, 0.25)$, $(0.30, 0.70, 0.35, 0.25, 0.20)$, $(0.25, 0.30, 0.65, 0.35, 0.20)$.
	Subgroup standard deviations $(0.18, 0.20, 0.20, 0.20, 0.20)$, $(0.20, 0.18, 0.20, 0.20, 0.20)$, $(0.20, 0.20, 0.18, 0.20, 0.20)$.	Subgroup standard deviations $(0.05, 0.07, 0.07, 0.07, 0.07)$, $(0.07, 0.05, 0.07, 0.07, 0.07)$, $(0.07, 0.07, 0.05, 0.07, 0.07)$.
Sampling details	Draw arm means from the subgroup Gaussian and clip each mean to $[-1.5, 1.5]$. The reward noise standard deviation is 1.0.	Draw arm means from the subgroup Gaussian and clip to $[0.02, 0.98]$. Draw a concentration parameter log-uniformly on $[15, 80]$ and convert each mean to Beta parameters.
Dataset scale	12 000 source tasks, of which 8 000 train the prior and surrogate components.	12 000 source tasks, of which 8 000 train the prior and surrogate components.

For conjugate Beta–Bernoulli bandits, Assumption 4 holds on compact interior domains. If the arm means are bounded away from 0 and 1, then the posterior parameters stay in a compact subset of $(0, \infty)^{2A}$. If in addition a positive margin separates the best arm from the others, then the Thompson-sampling action-probability map (25), the instantaneous regret rate, and the drift are all locally Lipschitz. Under these conditions the proxy error is controlled by the deviation between the stochastic belief trajectory and its deterministic mean-field path, plus the solver error.

F Experimental Details

F.1 Environments

We use three environments: a clustered Gaussian bandit population, a clustered Beta–Bernoulli bandit population, and a semi-synthetic population built from MovieLens (Harper and Konstan, 2015). The two synthetic environments give exact control over the type and the magnitude of the deployment shift, which the radius sweeps require. The MovieLens environment supplies a deployment shift that we did not impose, because it arises from a temporal split of real interaction logs.

Both synthetic environments use three latent subgroups, nominal mixture weights $(0.5, 0.3, 0.2)$, and five arms per task. A task is generated by sampling a subgroup, sampling the arm parameters from that subgroup, and then running Thompson sampling from the shared warm-start prior. Table 2 states the data-generating process of both environments.

Semi-synthetic MovieLens environment. We build the third environment from MovieLens-2k interactions (Harper and Konstan, 2015). We split the interactions in time, using the earliest 80% for training and the latest 20% for deployment and evaluation, and we binarise ratings into Bernoulli feedback with $y = 1$ for a rating of at least 4.0 and $y = 0$ otherwise. From the training split we compute low-rank collaborative-filtering embeddings, cluster users into $C_u = 8$ segments, and cluster items into $K = 5$ groups, where each item group acts as one arm. For every pair of a user segment and an arm we fit a smoothed Beta model with $(\alpha, \beta) = (\text{likes} + 1, \text{dislikes} + 1)$ and a minimum-count guard. The nominal composition is the segment distribution of the training split, and the deployment composition is the segment distribution of the evaluation split. Each evaluation trajectory fixes one deployment user, draws rewards from that user's arm-wise Bernoulli model with replacement, and runs for $T = 120$ rounds.

F.2 Shift Models and Radius Sweeps

The primary shift in both synthetic environments is population-composition shift. We perturb the deployment mixture weights around the nominal mixture while controlling the KL divergence from

the nominal composition. For the Gaussian environment the target divergence is 0.05 and the swept radii are $\{0.00, 0.01, 0.02, 0.03, 0.05, 0.08\}$. For the Beta–Bernoulli environment the primary run uses target divergence 0.02 with radii $\{0.00, 0.002, 0.005, 0.01, 0.015, 0.02\}$. We also report a reward-shift configuration for the Beta–Bernoulli environment, in which arm means are perturbed inside an ℓ_∞ ball of radius 0.02 and the swept radii are $\{0.00, 0.01, 0.02, 0.03, 0.05, 0.08\}$. Every radius is evaluated on five deployment draws.

F.3 Surrogate Models and Warm-Start Families

The regret surrogate is a multilayer perceptron trained on $30\,000$ training pairs and $6\,000$ validation pairs for 60 epochs, with AdamW, batch size 512, and gradient clipping at 5. Labels are Monte Carlo Thompson-sampling estimates of cumulative pseudo-regret at horizon 120, with 24 rollouts per label. The Gaussian surrogate uses hidden size 512 and the Beta–Bernoulli surrogate uses hidden size 768. The input features summarise the candidate prior and the sampled task statistics, including the mean, the scale, and the relative offsets.

We compare five warm-start families: a uniform prior, a nominal neural warm start, a robust neural warm start at a fixed radius, a KL-barycentre warm start, and a static distributionally robust non-adaptive policy. All neural variants share the surrogate architecture, the rollout-label budget, the optimiser settings, and the restart and validation-selection rules.

F.4 Evaluation Protocol and Reporting

Every adaptive warm-start method runs Thompson sampling at horizon $T = 120$, with repeated sampled deployment users and trajectories and 8 Monte Carlo rollouts per setting, and each deployment draw uses $1\,500$ sampled trajectories. We report final regret at the horizon and area-under-curve regret, computed as the mean cumulative regret over the horizon points. For the Gaussian experiments we summarise the mean and the standard deviation across draws. For the main Beta–Bernoulli composition-shift experiments we also report the worst case across the five draws.

F.5 Reproducibility

Unless stated otherwise, the experiments use global random seed 17, deployment-seed base 3100, and shift-generation seed 191. We release the raw and aggregated result tables, the regret plots, and the resolved configuration files. The experiments ran on the central processing unit of an Apple M2 laptop with 24 GB of memory, under Python 3.12.11 and PyTorch 2.5.1, and the full suite completes in about one hour on that machine.

G Computational Complexity

We separate offline training from the computation of a warm start. Let M be the number of subgroups, K_m the number of reference tasks of subgroup m, $K_{\mathrm{tot}} := \sum_m K_m$, B the minibatch size, S the number of robust-optimisation iterations, and C_{proxy} the cost of one proxy evaluation.

A nominal or KL-barycentre warm start reads the subgroup summaries once, so its cost is $\mathcal{O}(K_{\mathrm{tot}})$ plus solver overhead, which is small next to rollout-based regret estimation. Each DRWS iteration evaluates the proxy on the minibatches and then solves $M + 1$ scalar dual problems, so its cost is $\mathcal{O}(K_{\mathrm{tot}} C_{\mathrm{proxy}})$ and the total optimisation cost is $\mathcal{O}(S\, K_{\mathrm{tot}}\, C_{\mathrm{proxy}})$. The scalar dual updates in M are not the dominant term, and repeated proxy evaluation is.

The two proxy modules differ in C_{proxy}. A belief-ODE or Monte Carlo proxy costs $C_{\mathrm{proxy}}^{\mathrm{ODE}} = \mathcal{O}(T\, N_{\mathrm{int}}\, N_{\mathrm{mc}})$, where N_{int} is the number of integration steps and N_{mc} the number of Monte Carlo samples. A trained surrogate costs one forward pass, so $C_{\mathrm{proxy}}^{\mathrm{NN}} = \mathcal{O}(P)$ with P the network size. Training the surrogate costs $\mathcal{O}(J\, R\, T\, C_{\mathrm{TS}})$ for label generation, with J pairs, R rollout replicates, and C_{TS} the cost of one Thompson-sampling step, plus $\mathcal{O}(U\, B\, P)$ for U gradient steps. That cost is paid once and is amortised over many robust runs and deployment distributions.

At deployment, the per-round action selection of the base learner is unchanged, because DRWS modifies the initialisation and not the decision rule. The added online latency is therefore zero, and the whole cost of robustness sits in the offline optimisation. Direct robust optimisation with

simulation-based proxies is the most expensive regime, and the amortised surrogate trades a higher one-time training cost for a much lower cost per robust run.