INFORMATION-THEORETIC ANALYSIS OF MULTI-OBJECTIVE BANDITS: PARETO VIOLATIONS & GAP-BASED REGRET

Anonymous authors
Paper under double-blind review

ABSTRACT

We study Bayesian multi-objective bandits and ask a basic information-theoretic question: what aspects of an unknown environment must be resolved to achieve low regret? We show that the answer depends on the regret criterion and on the resolution of information needed to represent the corresponding loss. We first consider Pareto-violation regret, which incurs unit loss whenever the selected action is dominated. Because this loss depends only on Pareto membership, the Pareto-optimal set provides a natural information target. We introduce a separation-based measure of the difficulty of distinguishing among competing candidate Pareto sets, relate it to information gain about the unknown Pareto-optimal set, and use this connection to analyze Thompson Sampling and develop an Information-Directed Sampling strategy. We establish sublinear Bayesian regret under uniform and polynomially decaying separation and prove a lower bound showing that weak separability can impose an unavoidable statistical cost. We then consider gap-based Pareto regret, which measures the magnitude of Pareto suboptimality. Here, Pareto-set information is generally too coarse for a direct magnitude-sensitive analysis: distinct environments can share the same Pareto set while inducing different reward geometries and hence different Pareto gaps, and averaging over environments consistent with the same Pareto set need not preserve the Pareto dominance relations. This motivates retaining finer information about the environment that preserves the relevant reward geometry. We develop parameter-level information-theoretic analyses of Thompson Sampling and Information-Directed Sampling to derive gap-based Pareto regret guarantees. Together, our results reveal a hierarchy in the information needed for multi-objective learning: binary Pareto membership can be controlled through information about the Pareto set, whereas quantitative Pareto-gap suboptimality can require a finer representation of reward geometry.

1 INTRODUCTION

Many real-world decision-making problems require balancing competing objectives, such as engagement, fairness, and diversity in recommendation systems (Abdollahpouri et al., 2020; Ge et al., 2021; Rodriguez et al., 2012), or treatment effectiveness, risks, and adverse outcomes in healthcare (Char et al., 2018; Komorowski et al., 2018; Durand et al., 2018). These settings are naturally modeled as multi-objective bandits, where each action yields a vector of rewards. Unlike single-objective bandits, optimality is defined through Pareto dominance rather than a single best action: an action is Pareto optimal if no other action is at least as good in every objective and strictly better in at least one. Thus, learning requires navigating trade-offs across objectives (Turgay et al., 2018; Roijers et al., 2013).

Most existing work on multi-objective bandits evaluates performance through gap-based or related notions of Pareto regret, which quantify suboptimality relative to the Pareto frontier (Drugan & Nowé, 2013; Xu & Klabjan, 2023; Öner et al., 2018; Davoodi & Maghsudi, 2026). Approaches such as Pareto-UCB (Drugan & Nowé, 2013) have been followed by analyses in stochastic, adversarial, combinatorial, and contextual settings (Xu & Klabjan, 2023; Öner et al., 2018; Park et al., 2025). In contrast, less is understood about the cumulative binary Pareto-violation criterion studied here, which incurs unit loss whenever the selected action is dominated, irrespective of the magnitude of suboptimality. More broadly, existing analyses do not explain how the information required for low regret changes with the regret criterion. Thompson Sampling is well understood in single-objective

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bandits (Thompson, 1933; Agrawal & Goyal, 2012; 2017). Separately, for single-objective bandits, information-theoretic analyses of Thompson Sampling and Information-Directed Sampling relate immediate regret to information that resolves uncertainty about the unknown optimal action, which serves as a natural unknown target (Russo & Van Roy, 2014; 2016). In multi-objective bandits, optimality is instead represented by a Pareto set, and what the learner must resolve remains less clear

This leads to our central question: what information must a learner resolve to achieve low regret in multi-objective bandits? We show that the answer depends on the regret criterion. For Pareto-violation regret, which records whether the selected action is dominated, the Pareto-optimal set provides a natural information target. We identify the separability of competing candidate Pareto sets as a measure of statistical hardness and connect it to information gained about the unknown Pareto-optimal set. Building on this connection, we analyze Thompson Sampling and develop an Information-Directed Sampling (IDS) strategy that balances expected Pareto-violation loss against information gain about the unknown Pareto-optimal set. Under uniform separation, both achieve sublinear Bayesian regret. When separation decays polynomially as $t^{-\beta}$, generic Thompson Sampling incurs sublinear regret for $\beta < 1/2$, while IDS and separation-tuned Thompson Sampling extend this guarantee to $\beta < 1$. We also establish an information-theoretic lower bound showing that weak distinguishability between candidate Pareto sets imposes an unavoidable statistical cost.

For gap-based Pareto regret, where loss measures the magnitude of Pareto suboptimality, Pareto-set identity alone does not determine the reward geometry governing the gaps. Different models can share the same Pareto set yet have different reward geometries and Pareto gaps, and averaging across such models need not preserve Pareto optimality. Thus, gap-based IDS trades expected Pareto gap against information about the underlying model parameter rather than only about the Pareto set. We develop a parameter-level analysis for Thompson Sampling and IDS and obtain gap-based Pareto regret guarantees for standard finite-action Bayesian models. Together, these results show that Pareto regret criteria can require different resolutions of information. The main contributions are as follows

- We study Pareto-violation regret and gap-based Pareto regret in Bayesian multi-objective bandits and show that the information needed for effective learning, as well as the associated statistical difficulty, depends on how multi-objective performance is measured.
- For Pareto-violation regret, we introduce a separability-based hardness measure over competing candidate Pareto-optimal sets and connect it to information gain. We establish sublinear Bayesian regret for Thompson Sampling and IDS under uniform and polynomially decaying separation and prove an information-theoretic lower bound in the weak-separation regime. We further connect these conditions to generic identifiability in finite discrete Bayesian models and provide sufficient conditions based on distinguishable witness actions.
- For gap-based Pareto regret, we show that distinct models with the same Pareto set can induce different reward geometries and Pareto gaps. We thus develop a parameter-level information-theoretic analysis of Thompson Sampling and our proposed Information-Directed Sampling method, yielding general regret guarantees for standard finite-action Bayesian models.

2 Problem Setup and Preliminaries

We consider a stochastic multi-objective bandit in a Bayesian setting with a finite action set $\mathcal{A}$. Let $\boldsymbol{\theta}$ be an unknown possibly vector-valued parameter drawn from a known prior representing uncertainty about the environment. Conditioned on $\boldsymbol{\theta}$, at each round $t \in \mathbb{N}$ the learner selects an action $A_t \in \mathcal{A}$ and observes only the corresponding reward vector $\mathbf{Y}_{t,A_t} \in \mathbb{R}^d$. The sequence $(\mathbf{Y}_t)_{t\in\mathbb{N}}$, where $\mathbf{Y}_t = (\mathbf{Y}_{t,a})_{a\in\mathcal{A}}$, is i.i.d. conditional on $\boldsymbol{\theta}$. Define $\boldsymbol{\mu}_{\boldsymbol{\theta}}(a) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbf{Y}_{t,a}] \in \mathbb{R}^d$. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, write $\mathbf{x} \succeq \mathbf{y}$ if $x_i \geq y_i$ for all $i \in [d]$. The Pareto-optimal set is $\mathcal{S}^*(\boldsymbol{\theta}) := \{a \in \mathcal{A} \mid \nexists a' \in \mathcal{A}$ such that $\boldsymbol{\mu}_{\boldsymbol{\theta}}(a') \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a)$ and $\boldsymbol{\mu}_{\boldsymbol{\theta}}(a') \neq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a)\}$. We write $\mathcal{S}^* := \mathcal{S}^*(\boldsymbol{\theta})$ for the random Pareto set induced by the prior over $\boldsymbol{\theta}$. This is the multi-objective analogue of the unknown optimal-action target A^* used in single-objective information-theoretic bandit analyses (Russo & Van Roy, 2014).

We study two notions of regret. The first is Pareto-violation regret, with loss $\ell(a, \boldsymbol{\theta}) = \mathbb{1}\{a \notin \mathcal{S}^*(\boldsymbol{\theta})\}$. Its cumulative Bayesian regret over T rounds is $\mathbb{E}\big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*(\boldsymbol{\theta})\}\big]$, i.e., the expected number of times a dominated action is selected. The expectation is over the prior on $\boldsymbol{\theta}$, the observations, and any randomness in the policy. The second is gap-based Pareto regret using standard notion of uniform coordinate-wise Pareto gap (Drugan & Nowé, 2013; Auer et al., 2016; Xu & Klabjan, 2023). For each $a \in \mathcal{A}$, we define $\Delta_{\boldsymbol{\theta}}(a) := \sup\{\epsilon \geq 0 : \exists a' \in \mathcal{A}$ such that $\boldsymbol{\mu}_{\boldsymbol{\theta}}(a') \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \epsilon\mathbf{1}\}$, where $\mathbf{1} \in \mathbb{R}^d$

denotes the all-ones vector. Its cumulative Bayesian regret is $\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right]$. Every $a \in \mathcal{S}^*(\boldsymbol{\theta})$ satisfies $\Delta_{\boldsymbol{\theta}}(a) = 0$, while $\Delta_{\boldsymbol{\theta}}(a) > 0$ implies $a \notin \mathcal{S}^*(\boldsymbol{\theta})$. The converse need not hold because Pareto-violation regret records whether the selected action is dominated, whereas gap-based regret quantifies the magnitude of the largest simultaneous coordinate-wise improvement achievable over that action. The action A_t is chosen based on the history $\mathcal{H}_t = (A_1, \mathbf{Y}_{1,A_1}, \ldots, A_{t-1}, \mathbf{Y}_{t-1,A_{t-1}})$. A randomized policy $\pi = (\pi_t)_{t\in\mathbb{N}}$ maps each history $\mathcal{H}_t$ to a distribution over $\mathcal{A}$. Let $\Pi(\mathcal{A}) := \{p \in \mathbb{R}_+^{|\mathcal{A}|} : \sum_{a\in\mathcal{A}} p(a) = 1\}$ denote the probability simplex over $\mathcal{A}$. With slight abuse of notation, we write $\pi_t(a) := \mathbb{P}(A_t = a \mid \mathcal{H}_t)$. We denote the posterior distribution of $\boldsymbol{\theta}$ given $\mathcal{H}_t$ by p_t, which induces the posterior $p_t(s) := \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t)$ over the random Pareto set for $s \in 2^{\mathcal{A}}$.

3 Learning to Minimize Pareto-Violation Regret

We first study sequential decision-making under Pareto violations, where selecting an action outside the Pareto-optimal set incurs unit loss. Since Pareto-violation loss depends only on Pareto membership, we use the unknown Pareto set $\mathcal{S}^* := \mathcal{S}^*(\boldsymbol{\theta})$ as the information target and track its posterior uncertainty. The learner maintains $p_t(s) := \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t)$ over candidate Pareto sets $s \in 2^{\mathcal{A}}$. Define the disagreement probability $G_t := \mathbb{P}(S \neq S' \mid \mathcal{H}_t) = 1 - \sum_{s\in 2^{\mathcal{A}}} p_t(s)^2$, where $S, S' \overset{\text{i.i.d.}}{\sim} p_t$ conditional on $\mathcal{H}_t$; hence $G_t > 0$ precisely when multiple Pareto-set candidates remain plausible. For $a \in \mathcal{A}$, define $I_t(a) := I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$, and let $I_t := \sum_{a\in\mathcal{A}} \pi_t(a) I_t(a)$ denote the one-step information gain under the sampling distribution π_t. For $p_t(s) > 0$, let $P^a_{t,s} := \mathcal{L}(\mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a, \mathcal{S}^* = s)$ denote the posterior-predictive reward law under action a and Pareto-set candidate s, where $\mathcal{L}(X \mid \mathcal{G})$ denotes the conditional law of X given $\mathcal{G}$. Lemma 3.1 connects predictive distinguishability of competing candidate Pareto sets to information gain about $\mathcal{S}^*$.

Lemma 3.1. ***(Lower bound on information gain via average pairwise distinguishability)*** *Fix a round t, any realizable history $\mathcal{H}_t$, and an action $a \in \mathcal{A}$. Let $S, S' \overset{\text{i.i.d.}}{\sim} p_t$. Then*

$$I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \;\geq\; \frac{1}{2}\,\mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right]. \tag{1}$$

Lemma 3.1 shows that actions whose predictive reward laws better distinguish plausible Pareto-set candidates provide more information about the unknown Pareto set. To summarize this distinguishability across actions using a policy-independent reference measure, whenever $G_t > 0$ define $\kappa_t(\mathcal{H}_t) := \mathbb{E}\big[\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}\big[\mathrm{TV}(P^a_{t,S}, P^a_{t,S'})^2\big] \mid S \neq S', \mathcal{H}_t\big]$. Thus, $\kappa_t(\mathcal{H}_t)$ captures the average squared predictive separation between competing Pareto-set candidates at history $\mathcal{H}_t$ and provides the separation measure used for the analyses below. Averaging Lemma 3.1 uniformly over actions yields $\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}[I_t(a)] \geq \frac{1}{2} G_t \kappa_t(\mathcal{H}_t)$, directly linking predictive separation to information gain. We next prove that $\kappa_t(\mathcal{H}_t)$ is generically nonzero in the finite discrete model.

Theorem 3.1. *Consider a multi-objective bandit model with finite discrete reward distributions and finite action set $\mathcal{A}$, leading to Dirichlet count updates in the posterior over $\boldsymbol{\theta}$. Let $S, S' \overset{\text{i.i.d.}}{\sim} p_t$, where $p_t(s) := \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t)$, and let $q \in \Pi(\mathcal{A})$ satisfy $q(a) > 0$ for all $a \in \mathcal{A}$. Then, for all but a Lebesgue measure-zero set of Dirichlet prior parameters, for all t and all realized histories $\mathcal{H}_t$,*

$$\mathbb{E}\left[\mathbb{E}_{a\sim q}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right] \;\middle|\; S \neq S'\right] > 0,$$

whenever $\mathbb{P}(S \neq S' \mid \mathcal{H}_t) > 0$, under mild regularity and non-degeneracy conditions.

Theorem 3.1 establishes generic predictive separability for finite discrete rewards with Dirichlet count updates: outside a measure-zero set of prior parameters, whenever multiple Pareto-set candidates remain plausible, their posterior-predictive reward laws cannot be indistinguishable under every action. In particular, taking $q = \mathrm{Unif}(\mathcal{A})$ gives $\kappa_t(\mathcal{H}_t) > 0$, so Lemma 3.1 implies that some action has strictly positive information gain. For bounded rewards, a related finite-observation construction follows by Bernoulliization: for $R_{\max} > 0$ and $R \in [0, R_{\max}]^d$, generate $Z_j \mid R_j \sim \mathrm{Bernoulli}(R_j/R_{\max})$ independently across coordinates conditional on R. Since $\mathbb{E}[Z] = \mathbb{E}[R]/R_{\max}$, Pareto ordering of mean rewards is preserved up to positive scaling. Moreover, Bernoulliization is a fixed Markov kernel from R to Z, so distinct posterior-predictive laws for Z imply distinct laws for R. Hence, if the induced finite-output Bayesian model satisfies Theorem 3.1, its predictive-separability conclusion also certifies separation of the underlying reward laws. The

formal regularity and non-degeneracy conditions and proof are deferred to the supplementary material, where Appendix A.4 verifies them explicitly for a two-action, two-objective Beta–Bernoulli model. Theorem 3.1 guarantees only pointwise positivity of $\kappa_t(\mathcal{H}_t)$. Define the finite-horizon separation $\underline{\kappa}_T := \inf_{1\le t\le T,\ \mathcal{H}_t \text{ realizable},\ G_t>0} \kappa_t(\mathcal{H}_t)$. In the finite-action discrete model, only finitely many posterior count states are realizable up to any fixed T, so outside the Lebesgue measure-zero exceptional set, $\underline{\kappa}_T > 0$ for every fixed T. The assumptions below quantify how this separation behaves over time.

Assumption 3.1. ***(Uniform separation)*** *There exists a constant $\kappa > 0$ such that $\kappa_t(\mathcal{H}_t) \ge \kappa$ for every realizable history $\mathcal{H}_t$ satisfying $G_t > 0$ and every $t \in [T]$.*

Assumption 3.2. ***(Decaying separation)*** *There exist constants $c > 0$ and $\beta \ge 0$ such that $\kappa_t(\mathcal{H}_t) \ge c\,t^{-\beta}$ for every realizable history $\mathcal{H}_t$ satisfying $G_t > 0$ and every $t \in [T]$.*

Although $\kappa_t(\mathcal{H}_t)$ is history-dependent, Assumptions 3.1–3.2 require their bounds to hold over every realizable history $\mathcal{H}_t$ at every $t \in [T] := \{1, \dots, T\}$ for a finite horizon T. They are therefore model-level assumptions, independent of the particular algorithm. The quantities $\kappa_t(\mathcal{H}_t)$, κ, c, and β are instance-dependent and need not be known by Algorithms 1/2. The assumptions admit interpretable sufficient conditions. For bounded rewards, separation in one posterior-predictive mean coordinate yields an explicit lower bound on $\kappa_t(\mathcal{H}_t)$; see Proposition A.1 in Appendix A.5. More generally, if every competing posterior-supported Pareto-set pair admits, at each realizable $\mathcal{H}_t$ with $G_t > 0$, a witness action whose predictive laws have TV distance atleast $\delta_t > 0$, then $\kappa_t(\mathcal{H}_t) \ge \delta_t^2/|\mathcal{A}|$. Thus, $\delta_t \ge \delta > 0$ implies Assumption 3.1, while $\delta_t \ge \delta t^{-\gamma}$ implies Assumption 3.2 with $c = \delta^2/|\mathcal{A}|$ and $\beta = 2\gamma$. For example, consider a two-point prior over two environments such that action a_0 has deterministic reward $(1/2, 1/2)$ in both, whereas a_1 yields (Z, Z) with $Z \sim \mathrm{Bernoulli}(1/2 - \delta)$ in one and $Z \sim \mathrm{Bernoulli}(1/2 + \delta)$ in the other. The two environments induce different Pareto sets. For a_0, the predictive laws coincide and have TV distance 0, while for a_1 their TV distance is 2δ. Uniformly averaging over two actions gives $\kappa_t(\mathcal{H}_t) = 2\delta^2$ whenever both candidates remain plausible. Since $0 < 1/2 \pm \delta < 1$, neither environment is ruled out after any finite realizable history. Thus, Assumption 3.1 holds with $\kappa = 2\delta^2$, and Assumption 3.2 holds with $c = 2\delta^2$ & $\beta = 0$.

Uniform separation keeps competing Pareto-set candidates statistically separated, whereas decaying separation allows their predictive laws to approach one another at a controlled rate. These separation conditions characterize the model, but Algorithm 1 must still sample informative actions sufficiently often. At round t, Algorithm 1 draws $B_t \sim \mathrm{Bernoulli}(1 - \varepsilon_t)$. If $B_t = 1$, it samples $\boldsymbol{\theta}_t \sim p_t$, sets $S_t = \mathcal{S}^*(\boldsymbol{\theta}_t)$, and selects $A_t \sim \mathrm{Unif}(S_t)$; otherwise, it sets $S_t = \mathcal{A}$ and explores uniformly. Thus, with probability $1 - \varepsilon_t$ it follows a posterior Pareto-set draw, and with probability ε_t it explores uniformly. We analyze Bayesian cumulative Pareto-violation regret under Assumptions 3.1 and 3.2.

Theorem 3.2. *The Bayesian cumulative Pareto-violation regret of Algorithm 1 satisfies:*

- *Under Assumption 3.1 (Uniform separation)* $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le \left(2 + \frac{2H(\mathcal{S}^*)}{\kappa}\right)\sqrt{T}$.
- *Under Assumption 3.2 (Decaying separation)* $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le 2\sqrt{T} + \frac{2H(\mathcal{S}^*)}{c}\, T^{\beta+\frac{1}{2}}$. *In particular, under Assumption 3.2, the regret is sublinear whenever $\beta < \frac{1}{2}$.*

Theorem 3.2 shows that the schedule $\varepsilon_t = t^{-1/2}$ yields $O(\sqrt{T})$ expected Pareto-violation regret under uniform separation and sublinear regret under decaying separation whenever $\beta < \frac{1}{2}$. The proof bounds the one-step expected violation loss using the posterior disagreement $G_t = \mathbb{P}(S \ne S' \mid \mathcal{H}_t)$ and relates it to information gain using Lemma 3.1, the separation condition, and explicit exploration. By the mutual-information chain rule, the cumulative expected information gain equals $I(\mathcal{S}^*; \mathcal{H}_{T+1}) \le H(\mathcal{S}^*)$, where $H(\mathcal{S}^*)$ is the entropy of the random Pareto set induced by the prior through $\boldsymbol{\theta} \mapsto \mathcal{S}^*(\boldsymbol{\theta})$; since $\mathcal{S}^* \in 2^{\mathcal{A}}$, $H(\mathcal{S}^*) \le |\mathcal{A}| \log 2$. This information budget is independent of T. Proposition A.2 further shows that, when the separation profile is known, a tuned exploration schedule improves the rate to $T^{(1+\beta)/2}$, which is sublinear for every $\beta < 1$; thus, the $\beta < \frac{1}{2}$ threshold is a consequence of the fixed exploration schedule. Proposition 3.1 strengthens the regret guarantee by showing that Algorithm 1 also asymptotically tracks the true Pareto set, i.e., $\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}\{S_t \ne \mathcal{S}^*\}] = O(\sqrt{T})$ under Assumption 3.1 and $O(T^{\beta+\frac{1}{2}})$ under Assumption 3.2, which is sublinear for $\beta < \frac{1}{2}$. On TS rounds, S_t is a posterior draw of the Pareto set, while exploration rounds set $S_t = \mathcal{A}$ and can themselves cause mismatch. Overall, however, the sampled set is incorrect on only a vanishing fraction of rounds. Full proofs are deferred to the supplementary material.

Proposition 3.1. *For Algorithm 1, the expected cumulative Pareto-set mismatch, defined as* $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\Big]$, *satisfies* $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\Big] \leq \Big(2 + \frac{2H(\mathcal{S}^*)}{\kappa}\Big)\sqrt{T}$ *under Assumption 3.1, and* $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\Big] \leq 2\sqrt{T} + \frac{2H(\mathcal{S}^*)}{c}\, T^{\beta+\frac{1}{2}}$ *under Assumption 3.2.*

Algorithm 1: TS-Pareto with Uniform Exploration

Input: Prior p_1 over $\boldsymbol{\theta}$ and exploration schedule $\{\varepsilon_t\}_{t\geq 1}$ with $\varepsilon_t = t^{-1/2}$

for $t = 1, 2, \ldots, T$ **do**
- Draw $B_t \sim \text{Bernoulli}(1 - \varepsilon_t)$
- **if** $B_t = 1$ **then**
  - Sample $\boldsymbol{\theta}_t \sim p_t(\boldsymbol{\theta})$ and construct $S_t \leftarrow \mathcal{S}^*(\boldsymbol{\theta}_t)$; `// TS-Pareto branch`
- **else**
  - Set $S_t \leftarrow \mathcal{A}$; `// Uniform exploration branch`
- Play $A_t \sim \text{Unif}(S_t)$, observe $\mathbf{Y}_{t,A_t}$ & Update posterior: $p_{t+1}(\boldsymbol{\theta}) \propto p_t(\boldsymbol{\theta})\, \mathcal{L}(\mathbf{Y}_{t,A_t} \mid A_t, \boldsymbol{\theta})$.

We also develop an information-directed sampling (IDS) strategy for the Pareto-violation setting (Algorithm 2). For each action a, define $\Delta_t(a) = 1 - \mathbb{P}(a \in \mathcal{S}^* \mid \mathcal{H}_t) = \mathbb{P}(a \notin \mathcal{S}^* \mid \mathcal{H}_t)$ as its posterior expected one-step Pareto-violation loss, and $I_t(a) = I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$ as the information gain about the unknown Pareto set. Algorithm 2 selects a policy π_t^{IDS} minimizing the information ratio $\psi_t(\pi) := \Delta_t(\pi)^2 / I_t(\pi)$, where $\Delta_t(\pi) = \mathbb{E}_{a\sim\pi}[\Delta_t(a)]$ and $I_t(\pi) = \mathbb{E}_{a\sim\pi}[I_t(a)]$. IDS then samples $A_t \sim \pi_t^{\text{IDS}}$. Thus, IDS favors sampling distributions that balance low expected Pareto-violation loss against information gain about the unknown Pareto set. Unlike Algorithm 1, IDS incorporates exploration directly through its objective and requires no separate exploration schedule. Its Bayesian cumulative Pareto-violation regret satisfies $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_t(\pi_t^{\text{IDS}})\right]$, since $\Delta_t(\pi_t^{\text{IDS}})$ is exactly the posterior expected one-step violation probability under policy π_t^{IDS}.

Theorem 3.3. *Assume that, for some* $\Psi > 0$, $\psi_t\big(\pi_t^{\text{IDS}}\big) \leq \Psi$ *for all realizable* $\mathcal{H}_t$ *and* $\forall t \in [T]$. *Then, for Algorithm 2, the Bayesian cumulative Pareto-violation regret satisfies* $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\Big] = \mathbb{E}\Big[\sum_{t=1}^{T} \Delta_t\big(\pi_t^{\text{IDS}}\big)\Big] \leq \sqrt{\Psi\, H(\mathcal{S}^*)\, T}$. *Furthermore, under Assumption 3.1,* $\psi_t\big(\pi_t^{\text{IDS}}\big) \leq \frac{8}{\kappa}$ *for all* $t \in [T]$ *and* $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\Big] \leq \sqrt{\frac{8}{\kappa} H(\mathcal{S}^*)\, T}$. *Under Assumption 3.2,* $\psi_t\big(\pi_t^{\text{IDS}}\big) \leq \frac{8}{c} t^{\beta}$ *for all* $t \in [T]$, *and* $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\Big] \leq \sqrt{\frac{8}{c} H(\mathcal{S}^*)\, T^{\beta+1}}$, *which is sublinear whenever* $\beta < 1$.

Theorem 3.3 gives the Pareto-violation regret guarantee for Algorithm 2. The proof bounds the IDS information ratio via a two-action comparison distribution combining a low-loss action with an action informative about competing Pareto-set candidates, and then uses mutual-information chain rule and entropy budget. Under Assumption 3.2, IDS achieves the $T^{(1+\beta)/2}$ rate and is therefore sublinear for every $\beta < 1$, without requiring a prescribed exploration schedule. By Proposition A.2, a separation-tuned TS schedule attains the same horizon exponent when the separation profile is known. IDS instead adapts its sampling distribution directly to the current loss-information tradeoff, at the cost of an additional optimization at each round (Russo & Van Roy, 2014). We next show that the separation parameter captures an intrinsic statistical difficulty in the weak-separation regime.

Proposition 3.2. *For every* $\kappa \in (0, 1/8]$, *there exists a two-armed, two-objective Bayesian bandit instance for which, at every finite realizable history* $\mathcal{H}_t$ *with* $G_t > 0$, *the average squared-TV separation satisfies* $\kappa_t(\mathcal{H}_t) = \kappa$. *Furthermore, for every horizon* $T \geq 1$, *every possibly randomized, adaptive algorithm satisfies* $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*(\boldsymbol{\theta})\}\Big] \geq \frac{1}{32} \min\{T, \frac{1}{\kappa}\}$.

Proposition 3.2 shows that weak separation creates an intrinsic statistical cost: for every $\kappa \in (0, 1/8]$, some instance forces every adaptive randomized algorithm to incur Bayesian Pareto-violation regret at least $\frac{1}{32}\min\{T, 1/\kappa\}$. Hence, for $T \leq 1/\kappa$ the lower bound is linear in T, reaching order $1/\kappa$ at the crossover $T \asymp 1/\kappa$. The restriction $\kappa \leq 1/8$ isolates the weak-separation regime; when $\kappa > 1/8$, $1/\kappa$ is bounded by a universal constant. Theorem 3.3 gives Pareto-IDS regret $O(\sqrt{H(\mathcal{S}^*)T/\kappa})$, while Proposition A.2 gives the same T–κ dependence for tuned Thompson Sampling, using $\varepsilon_t = \min\{\alpha/\sqrt{t}, 1\}$ with $\alpha = \sqrt{H(\mathcal{S}^*)/\kappa}$. At the crossover $T \asymp 1/\kappa$, both upper

bounds scale as $O(\sqrt{H(\mathcal{S}^*)}/\kappa)$ versus the lower bound $\Omega(1/\kappa)$, matching the κ^{-1} dependence up to the factor $\sqrt{H(\mathcal{S}^*)}$ and universal constants; see Proposition A.2 in Section A.10.

Algorithm 2: Pareto-IDS for Minimizing Pareto Violations

Input: Prior p_1 over $\boldsymbol{\theta}$

for $t = 1, 2, \ldots, T$ **do**

- Observe history $\mathcal{H}_t$ and posterior $p_t(\boldsymbol{\theta})$, inducing $p_t(s) = \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t)$ over $\mathcal{S}^*$
- Compute $\forall a \in \mathcal{A}$, $\Delta_t(a) = 1 - \mathbb{P}(a \in \mathcal{S}^* \mid \mathcal{H}_t)$ and $I_t(a) = I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$
- Define $\Delta_t(\pi) = \mathbb{E}_{a\sim\pi}[\Delta_t(a)]$ and $I_t(\pi) = \mathbb{E}_{a\sim\pi}[I_t(a)]$ for $\pi \in \Pi(\mathcal{A})$.
- Choose $\pi_t^{\mathrm{IDS}} \in \arg\min_{\pi\in\Pi(\mathcal{A})} \frac{\Delta_t(\pi)^2}{I_t(\pi)}$; `// See Section A.1`
- Play $A_t \sim \pi_t^{\mathrm{IDS}}$, observe $\mathbf{Y}_{t,A_t}$, and update the posterior $p_{t+1}(\boldsymbol{\theta}) \propto p_t(\boldsymbol{\theta})\, \mathcal{L}(\mathbf{Y}_{t,A_t} \mid A_t, \boldsymbol{\theta})$.

4 Learning to Minimize Gap-Based Pareto Regret

We now turn to learning to minimize gap-based Pareto regret. Unlike the binary Pareto-violation objective in Section 3, $\mathbb{1}\{a \notin \mathcal{S}^*(\boldsymbol{\theta})\}$, we consider the Pareto gap $\Delta_{\boldsymbol{\theta}}(a) := \sup\{\epsilon \geq 0 : \exists a' \in \mathcal{A}$ such that $\boldsymbol{\mu}_{\boldsymbol{\theta}}(a') \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \epsilon\mathbf{1}\}$, which measures the magnitude of suboptimality and is standard in prior work on multi-objective bandits (Drugan & Nowé, 2013; Auer et al., 2016; Xu & Klabjan, 2023). Equivalently, $\Delta_{\boldsymbol{\theta}}(a) = \big(\sup_{b\in\mathcal{A}} \min_{j\in[d]} \big(\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a)\big)\big)_+$, where $(x)_+ := \max\{x, 0\}$ and $\mu_{\boldsymbol{\theta},j}(a)$ is the j-th coordinate of $\boldsymbol{\mu}_{\boldsymbol{\theta}}(a)$; see Proposition A.3. This creates a key conceptual distinction from Section 3. There, the relevant unknown object is the Pareto-optimal set $\mathcal{S}^*(\boldsymbol{\theta})$, because conditional on $\mathcal{S}^*(\boldsymbol{\theta}) = s$, the violation loss of action a is exactly $\mathbb{1}\{a \notin s\}$, regardless of which parameter inducing s is the true environment. By contrast, gap-based regret depends on the reward geometry induced by $\boldsymbol{\theta}$ and is not determined by the Pareto set alone. Thus Pareto-set information can be too coarse for magnitude-sensitive regret analysis, and our analysis retains finer parameter-level information about the reward geometry. Importantly, the issue is not merely that the Pareto set fails to determine every numerical Pareto gap: in a single-objective bandit, the identity of the optimal action A^* likewise does not determine every scalar gap. The key distinction is that averaging over parameter values consistent with the same Pareto set can alter the reward geometry and may even change the induced Pareto ordering required to represent gap-based regret itself.

This raises a natural question: why is the single-objective case different, even though A^* likewise does not determine every scalar gap? In information-theoretic analyses of Thompson Sampling (Russo & Van Roy, 2014; 2016), the relevant object is the optimal action $A^* := A^*(\boldsymbol{\theta})$ (with fixed tie-breaking). There, the instantaneous regret of action a is $\Delta_t(a) := \mathbb{E}[R_{t,A^*} - R_{t,a} \mid \mathcal{H}_t]$, which decomposes as $\Delta_t(a) = \sum_{a^*\in\mathcal{A}} \mathbb{P}_t(A^* = a^*)\, \mathbb{E}[\mu_{\boldsymbol{\theta}}(a^*) - \mu_{\boldsymbol{\theta}}(a) \mid A^* = a^*, \mathcal{H}_t]$. For each $a^* \in \mathcal{A}$, let $\Theta_{a^*} := \{\boldsymbol{\theta} : A^*(\boldsymbol{\theta}) = a^*\}$. Since $\mu_{\boldsymbol{\theta}}(a^*) \geq \mu_{\boldsymbol{\theta}}(a)$ for every $\boldsymbol{\theta} \in \Theta_{a^*}$ and all $a \in \mathcal{A}$, conditioning on $A^* = a^*$ preserves these inequalities: $\mathbb{E}[\mu_{\boldsymbol{\theta}}(a^*) \mid A^* = a^*, \mathcal{H}_t] \geq \mathbb{E}[\mu_{\boldsymbol{\theta}}(a) \mid A^* = a^*, \mathcal{H}_t]$. Thus, although scalar gap magnitudes may vary across parameters compatible with $A^* = a^*$, averaging over them preserves the optimality of the fixed comparator a^*, and conditional regret remains linear in mean rewards. Hence the A^*-level reduction is compatible with scalar information-theoretic analysis: conditioning on $A^* = a^*$ preserves both the comparator and the sign of every regret difference relative to it, while linearity keeps conditional regret aligned with the A^*-conditioned predictive model. This is precisely what can fail in the multi-objective setting. The Pareto set $\mathcal{S}^*(\boldsymbol{\theta})$ records which actions are non-dominated, but not the objective-wise tradeoffs responsible for their non-domination. Different parameter values can therefore induce the same Pareto set with different reward geometries, and averaging over them can alter the Pareto relations themselves. Thus conditioning on $\mathcal{S}^*$ does not generally preserve the structure needed for gap-based regret: unlike scalar regret, the Pareto gap is a nonlinear function of the reward geometry, $\Delta_{\boldsymbol{\theta}}(a) = \big(\sup_{b\in\mathcal{A}} \min_{j\in[d]} (\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a))\big)_+$.

A simple example makes this instability concrete. Consider three actions a_1, a_2, a_3, two objectives, and two parameter values $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$, with $\boldsymbol{\mu}_{\boldsymbol{\theta}_1}(a_1) = (1, 3)$, $\boldsymbol{\mu}_{\boldsymbol{\theta}_1}(a_2) = (3, 1)$, $\boldsymbol{\mu}_{\boldsymbol{\theta}_1}(a_3) = (2.2, 2.2)$, while $\boldsymbol{\mu}_{\boldsymbol{\theta}_2}(a_1) = (3, 1)$, $\boldsymbol{\mu}_{\boldsymbol{\theta}_2}(a_2) = (1, 3)$, and $\boldsymbol{\mu}_{\boldsymbol{\theta}_2}(a_3) = (2.2, 2.2)$. In each environment no action dominates another, so $\mathcal{S}^*(\boldsymbol{\theta}_1) = \mathcal{S}^*(\boldsymbol{\theta}_2) = \{a_1, a_2, a_3\}$. Under the equal-weight average $\bar{\boldsymbol{\mu}}(a) := \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\theta}_1}(a) + \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\theta}_2}(a)$, however, $\bar{\boldsymbol{\mu}}(a_1) = \bar{\boldsymbol{\mu}}(a_2) = (2, 2)$ whereas $\bar{\boldsymbol{\mu}}(a_3) = (2.2, 2.2)$, so the Pareto set collapses to $\{a_3\}$. Indeed, in the averaged geometry, a_1 and a_2 have gap 0.2, although both have zero gap in each original environment. Thus actions that are Pareto-optimal in every

constituent environment can become dominated after averaging. Nor is this an isolated equal-weight phenomenon: mixing the environments with weights λ and $1-\lambda$ gives $\bar{\boldsymbol{\mu}}_\lambda(a_1) = (3-2\lambda, 1+2\lambda)$, $\bar{\boldsymbol{\mu}}_\lambda(a_2) = (1+2\lambda, 3-2\lambda)$, and $\bar{\boldsymbol{\mu}}_\lambda(a_3) = (2.2, 2.2)$; for every $\lambda \in (0.4, 0.6)$, a_3 strictly dominates both a_1 and a_2. Hence an entire interval of mixtures changes the Pareto relations even though both underlying parameters induce exactly the same Pareto set. This example also explains why conditioning on a coarser information target can be problematic. Such coarse conditioning forms a predictive law whose mean averages reward geometries over parameters consistent with that event. In the scalar case, conditioning on $A^* = a^*$ preserves the optimality of the fixed comparator a^* under this averaging. In contrast, conditioning on $\mathcal{S}^* = s$ can average multi-objective reward geometries in a way that changes their Pareto relations. Thus the problem is not merely that $\mathcal{S}^*$ does not determine the numerical values of $\Delta_{\boldsymbol{\theta}}(a)$: the Pareto gap of the averaged geometry need not equal the average of the parameter-specific Pareto gaps, because the gap functional is nonlinear in the reward geometry. This blocks a direct A^*-level reduction for gap-based Pareto regret and motivates retaining parameter-level information. Accordingly, for each $a \in \mathcal{A}$, define $I_t(a) := I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$, and under the actual sampling distribution $\pi_t(a) := \mathbb{P}(A_t = a \mid \mathcal{H}_t)$, define $I_t := \sum_{a\in\mathcal{A}} \pi_t(a) I_t(a)$. Also let $P^a_{\boldsymbol{\theta}} := \mathcal{L}(\mathbf{Y}_{t,a} \mid \boldsymbol{\theta})$ denote the action-specific reward law. By retaining $\boldsymbol{\theta}$ rather than collapsing parameters that share the same Pareto set, the analysis preserves the reward geometry needed to control the actual Pareto gap, while Lemma 4.1 links greater distinguishability among the parameter-conditioned reward laws for action a to greater information gain from playing it.

Lemma 4.1. *Fix a round t, a realizable history $\mathcal{H}_t$, and an action $a \in \mathcal{A}$. Let p_t denote the posterior distribution of $\boldsymbol{\theta}$ given $\mathcal{H}_t$, and let $\boldsymbol{\theta}, \boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t$. Under the fixed-action interpretation,*

$$I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \,|\, \mathcal{H}_t, A_t = a) \geq \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2\right]. \tag{2}$$

Algorithm 3: Thompson Sampling for Gap-Based Pareto Regret

Input: Prior p_1 over $\boldsymbol{\theta}$

for $t = 1, 2, \ldots, T$ **do**

  Sample $\boldsymbol{\theta}_t \sim p_t$ and construct $S_t \leftarrow \mathcal{S}^*(\boldsymbol{\theta}_t)$.

  Play $A_t \sim \mathrm{Unif}(S_t)$, observe $\mathbf{Y}_{t,A_t}$ & Update posterior: $p_{t+1}(\boldsymbol{\theta}) \propto p_t(\boldsymbol{\theta})\, \mathcal{L}(\mathbf{Y}_{t,A_t} \mid A_t, \boldsymbol{\theta})$.

We propose Thompson Sampling for gap-based Pareto regret in Algorithm 3. At each round t, the algorithm samples $\boldsymbol{\theta}_t \sim p_t$, forms $S_t := \mathcal{S}^*(\boldsymbol{\theta}_t)$, and selects $A_t \sim \mathrm{Unif}(S_t)$. Unlike in the Pareto-violation setting, the analysis operates at the parameter level and requires no explicit exploration schedule. Regret is controlled through the information budget $I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$, where $\mathcal{H}_{T+1} := (A_1, \mathbf{Y}_{1,A_1}, \ldots, A_T, \mathbf{Y}_{T,A_T})$ denotes the observed interaction history; this quantity measures how much the interaction reveals about the underlying environment. Unlike $I(\mathcal{S}^*; \mathcal{H}_{T+1})$ in Theorem 3.2, it need not admit a finite-entropy bound because $\boldsymbol{\theta}$ may be continuous. Nevertheless, for many standard Bayesian models, $I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$ grows only logarithmically in T, yielding near-$\sqrt{T}$ regret; model-specific bounds are given later in this section. Dependence on the Pareto-set size is captured by $M := \sup_{\boldsymbol{\theta}\in\mathrm{supp}(p_1)} |\mathcal{S}^*(\boldsymbol{\theta})|$, the largest Pareto-set size over the prior support. Since $\mathcal{S}^*(\boldsymbol{\theta}) \subseteq \mathcal{A}$, we have $M \leq |\mathcal{A}|$, with equality possible in the worst case.

Theorem 4.1. *Under bounded rewards, for Algorithm 3, the Bayesian cumulative gap-based Pareto regret satisfies $\mathbb{E}\left[\sum_{t=1}^T \Delta_{\boldsymbol{\theta}}(A_t)\right] = O\left(\sqrt{|\mathcal{A}|\, M\, \mathrm{IB}_T\, T}\right)$, where $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$, $\mathcal{H}_{T+1} := (A_1, \mathbf{Y}_{1,A_1}, \ldots, A_T, \mathbf{Y}_{T,A_T})$ denotes the observed interaction history up to horizon T, and $M := \sup_{\boldsymbol{\theta}\in\mathrm{supp}(p_1)} |\mathcal{S}^*(\boldsymbol{\theta})| \leq |\mathcal{A}|$. Since $M \leq |\mathcal{A}|$, the worst-case bound is $O\big(|\mathcal{A}|\sqrt{\mathrm{IB}_T\, T}\big)$.*

The proof of Theorem 4.1, in the supplementary material, uses Lemmas A.1 and A.2. Lemma A.1 relates the Pareto gap of an action $a \in S_t = \mathcal{S}^*(\boldsymbol{\theta}_t)$ to total-variation separation between relevant observation laws by constructing a Pareto-optimal witness under the true parameter $\boldsymbol{\theta}$. Lemma A.2 then combines this relation with the sampling probabilities and Lemma 4.1 to relate one-step expected regret to information gain. The main difficulty is that A_t is sampled from the random set S_t, and is therefore coupled with the posterior sample $\boldsymbol{\theta}_t$. Let $q_t(a) := \mathbb{P}(a \in S_t \mid \mathcal{H}_t)$ and $\pi_t(a) := \mathbb{P}(A_t = a \mid \mathcal{H}_t)$. Since $|S_t| \leq M$ and A_t is drawn uniformly from S_t, $\pi_t(a) \geq q_t(a)/M$, so $q_t(a)/\pi_t(a) \leq M$. Combining Lemma A.1, this relation, Cauchy–Schwarz, and Lemma 4.1 yields $\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t]^2 \leq 32 R^2_{\max} |\mathcal{A}| M \sum_{a\in\mathcal{A}} \pi_t(a) I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$. Summing over t and applying the mutual-information chain rule gives the stated regret bound. The additional $\sqrt{M}$ factor

arises from the relation $\pi_t(a) \geq q_t(a)/M$; when $M = 1$, it disappears, leaving the familiar $\sqrt{|\mathcal{A}|}$ dependence of information-theoretic Thompson Sampling bounds (Russo & Van Roy, 2014; 2016).

Corollary 4.1. *Under bounded rewards, for Algorithm 3, let $S_t := \mathcal{S}^*(\boldsymbol{\theta}_t)$ denote the sampled Pareto set at round t. Then $\mathbb{E}\Big[\sum_{t=1}^{T} \sum_{a \in S_t} \Delta_{\boldsymbol{\theta}}(a)\Big] = O\Big(\sqrt{|\mathcal{A}|\, M^3\, \mathrm{IB}_T\, T}\Big)$, where $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$. In particular, since $M \leq |\mathcal{A}|$, the worst-case bound is $O\big(|\mathcal{A}|^2 \sqrt{\mathrm{IB}_T T}\big)$.*

Corollary 4.1 further controls the aggregate true Pareto gap of the sampled sets, complementing the action guarantee. Algorithm 4 is the IDS analogue for gap-based Pareto regret, using the same structure as Algorithm 2 but different quantities in the information ratio. In the Pareto-violation setting, the loss depends only on whether an action belongs to $\mathcal{S}^*$, so IDS trades off the posterior violation probability $1 - \mathbb{P}(a \in \mathcal{S}^* \mid \mathcal{H}_t)$ against information about $\mathcal{S}^*$. In the gap-based setting, the loss is instead $\Delta_{\boldsymbol{\theta}}(a)$, which depends on reward geometry encoded by $\boldsymbol{\theta}$ and is not determined by $\mathcal{S}^*$ alone. Accordingly, IDS uses the posterior expected gap $\Delta_t(a) := \mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t]$ and parameter-level information gain $I_t(a) := I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$. For $\pi \in \Pi(\mathcal{A})$, define $\Delta_t(\pi) := \mathbb{E}_{a\sim\pi}[\Delta_t(a)]$, $I_t(\pi) := \mathbb{E}_{a\sim\pi}[I_t(a)]$, and $\psi_t(\pi) := \Delta_t(\pi)^2 / I_t(\pi)$. Algorithm 4 selects π_t^{IDS} minimizing $\psi_t(\pi)$.

Algorithm 4: IDS for Gap-Based Pareto Regret

Input: Prior p_1 over $\boldsymbol{\theta}$

for $t = 1, 2, \ldots, T$ **do**

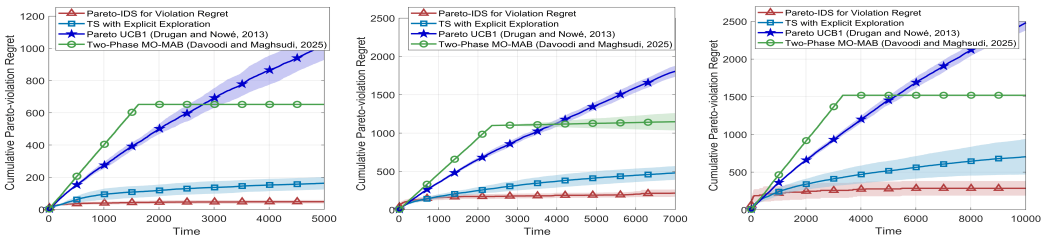
- Observe history $\mathcal{H}_t$ and posterior $p_t(\boldsymbol{\theta})$.
- Compute, for each $a \in \mathcal{A}$, $\Delta_t(a) = \mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t]$ and $I_t(a) = I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$.
- Define $\Delta_t(\pi) = \mathbb{E}_{a\sim\pi}[\Delta_t(a)]$ and $I_t(\pi) = \mathbb{E}_{a\sim\pi}[I_t(a)]$ for $\pi \in \Pi(\mathcal{A})$.
- Choose $\pi_t^{\mathrm{IDS}} \in \arg\min_{\pi \in \Pi(\mathcal{A})} \Delta_t(\pi)^2 / I_t(\pi)$; `// See Section A.1`
- Play $A_t \sim \pi_t^{\mathrm{IDS}}$, observe $\mathbf{Y}_{t,A_t}$, and update $p_{t+1}(\boldsymbol{\theta}) \propto p_t(\boldsymbol{\theta})\, \mathcal{L}(\mathbf{Y}_{t,A_t} \mid A_t, \boldsymbol{\theta})$.

Theorem 4.2. *Assume that rewards are bounded and that, for some $\Psi > 0$, $\psi_t(\pi_t^{\mathrm{IDS}}) \leq \Psi$ for every realizable $\mathcal{H}_t$, $\forall t \in [T]$. Then Algorithm 4 satisfies $\mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\Big] \leq \sqrt{\Psi\, \mathrm{IB}_T\, T}$, where $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$. Moreover, there exists a constant $C > 0$, depending only on the fixed reward bound, such that $\psi_t(\pi_t^{\mathrm{IDS}}) \leq C|\mathcal{A}|M$ for all $t \in [T]$. Consequently, $\mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\Big] = O\Big(\sqrt{|\mathcal{A}|\, M\, \mathrm{IB}_T\, T}\Big)$. Since $M \leq |\mathcal{A}|$, the worst-case bound is $O\big(|\mathcal{A}| \sqrt{\mathrm{IB}_T T}\big)$.*

The proof of Theorem 4.2, given in the supplementary material, converts a uniform information-ratio bound into cumulative regret via Cauchy–Schwarz and the mutual-information chain rule. For an explicit bound, we compare IDS with the one-step Thompson-sampling distribution from Algorithm 3; since π_t^{IDS} minimizes the information ratio, its ratio is no larger than that of this feasible comparator, whose ratio is $O(|\mathcal{A}|M)$ by the one-step analysis underlying Theorem 4.1.

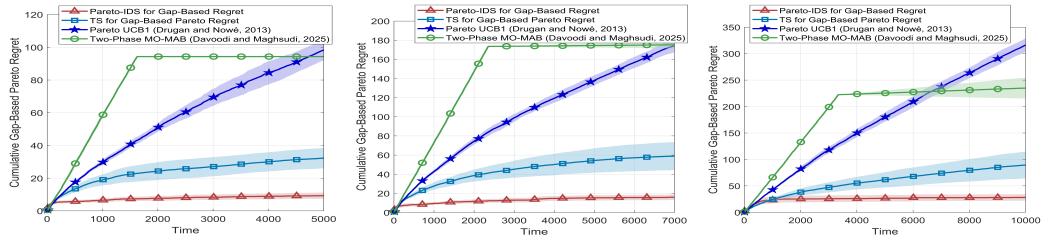
Proposition 4.1. *For finite-action Bayesian models, the following bounds hold for $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$. For finite parameter space Θ: $\mathrm{IB}_T \leq H(\boldsymbol{\theta}) \leq \log|\Theta|$. In the multi-objective categorical reward model with d objectives and K categories per objective, suppose that $\boldsymbol{\theta}$ consists of the categorical probability vectors and that each action–objective pair has an independent Dirichlet prior with parameters $(\alpha_{a,j,1}, \ldots, \alpha_{a,j,K})$ satisfying $\alpha_{a,j,k} \geq 1$ for all $a \in \mathcal{A}$, $j \in [d]$, and $k \in [K]$. Then $\mathrm{IB}_T = O(dK|\mathcal{A}| \log T)$. In the finite-action Gaussian reward model, suppose that $\boldsymbol{\theta} = (\{\mu_a\}_{a\in\mathcal{A}}, \Sigma)$, $\mathbf{Y}_{t,A_t} \mid (A_t = a, \boldsymbol{\theta}) \sim \mathcal{N}(\mu_a, \Sigma)$, Σ has an inverse-Wishart prior, and $\{\mu_a\}_{a\in\mathcal{A}}$ have independent Gaussian priors conditional on Σ. Then $\mathrm{IB}_T = O(d|\mathcal{A}| \log T + d^2 \log T)$.*

Proposition 4.1 shows that the parameter-level information budget $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$ grows at most logarithmically in T for several standard finite-action Bayesian models. Substituting these bounds into Theorems 4.1 and 4.2 yields near-$\sqrt{T}$ gap-based Pareto regret, up to logarithmic factors. For bounded rewards, Bernoulliization preserves the relevant Pareto geometry: for $Y_{t,j} \in [0, R_{\max}]$, let $Z_{t,j} \mid Y_{t,j} \sim \mathrm{Bernoulli}(Y_{t,j}/R_{\max})$. Then $\mathbb{E}[Z_{t,j} \mid \boldsymbol{\theta}, A_t = a] = \mathbb{E}[Y_{t,j} \mid \boldsymbol{\theta}, A_t = a]/R_{\max}$, so Pareto dominance is preserved and Pareto gaps are scaled by $1/R_{\max}$. If the induced Bernoulli model satisfies Proposition 4.1, it therefore inherits a logarithmic budget. For unbounded Gaussian rewards, Theorems 4.1 and 4.2 do not apply directly. Under bounded-mean and sub-Gaussian tail conditions, clipping at order $\sqrt{\log T}$ yields a bounded model whose parameter-information budget is no larger than the original by data processing. Thus, the logarithmic information-budget bound of Proposition 4.1 transfers to the clipped process with only an $o(1)$ cumulative gap-bias term; see Appendix A.17. The same argument extends to finite-action multi-objective linear-Gaussian bandits.



(a) $|\mathcal{A}| = 10,\ d = 4$ (b) $|\mathcal{A}| = 15,\ d = 5$ (c) $|\mathcal{A}| = 20,\ d = 6$

Figure 1: Cumulative Pareto-violation regret on multi-objective Bernoulli bandit instances.



(a) $|\mathcal{A}| = 10,\ d = 4$ (b) $|\mathcal{A}| = 15,\ d = 5$ (c) $|\mathcal{A}| = 20,\ d = 6$

Figure 2: Cumulative gap-based Pareto regret on multi-objective Bernoulli bandit instances.

5 Experiments

We evaluate the proposed algorithms on synthetic multi-objective Bernoulli bandits under both regret criteria. For each action $a \in \mathcal{A}$ and objective $j \in [d]$, we sample $\mu_j(a)$ independently and uniformly from $[0.05, 0.95]$ and generate $Y_{t,j}(a) \sim \text{Bernoulli}(\mu_j(a))$ independently across objectives. Bernoulli rewards provide a canonical bounded-reward model for our analysis; more general bounded rewards can be handled through the Bernoullization argument discussed earlier, which preserves mean reward vectors up to scaling and hence Pareto dominance. We consider $(|\mathcal{A}|, d) \in \{(10, 4), (15, 5), (20, 6)\}$. For Pareto-violation regret, we compare TS-Pareto with explicit exploration (Algorithm 1) and Pareto-IDS (Algorithm 2) against Pareto-UCB1 (Drugan & Nowé, 2013) and the two-phase explore-then-commit MO-MAB algorithm of Davoodi & Maghsudi (2026). The latter first explores all arms and then commits to a reduced cover set, with regret $O(T^{2/3}(|\mathcal{A}| \log T)^{1/3})$ for the exact-cover version and an additional $O(\log |\mathcal{A}|)$ factor for its polynomial-time approximation. We evaluate all four methods using cumulative Pareto-violation regret, i.e., the number of rounds in which a selected action is not Pareto-optimal under the true mean vectors. For each size, we run 200 independent trials and report mean cumulative regret with one-standard-deviation bands. Figure 1 shows that Pareto-IDS achieves the lowest regret across all settings, followed by TS-Pareto, while both baselines incur larger regret. This is consistent with IDS directly balancing violation probability against information gain about the unknown Pareto set, whereas TS-Pareto incurs additional cost from explicit exploration to sample informative actions.

For gap-based Pareto regret, we use the same environments and problem sizes, with $\Delta(a) = \big(\sup_{b \in \mathcal{A}} \min_{j \in [d]} (\mu_j(b) - \mu_j(a))\big)_+$, and compare gap-based TS-Pareto (Algorithm 3) and Pareto-IDS (Algorithm 4) against the same baselines. TS-Pareto samples a posterior parameter and plays uniformly from its Pareto set, whereas Pareto-IDS minimizes the information ratio using posterior expected gap and parameter-level information gain. Figure 2 shows the same ordering: Pareto-IDS achieves the lowest regret, followed by TS-Pareto, while Pareto-UCB1 and the two-phase MO-MAB baseline incur larger regret. This supports the Section 4 principle of balancing expected gap against parameter-level information gain. Continuous-Gaussian experiments in Appendix A.19 show the same ordering, indicating that these empirical conclusions are not specific to Bernoulli rewards.

6 Conclusion

We give an information-theoretic view of Bayesian multi-objective bandits under Pareto-violation and gap-based regret. Avoiding dominated actions mainly requires learning the Pareto set, whereas minimizing Pareto gaps requires finer reward-geometry information. Future work includes extensions to richer structured multi-objective settings with continuous action spaces .

AI USE STATEMENT

Generative AI tools were used solely to assist with editing and polishing the manuscript, including improvements to grammar, sentence structure, clarity, readability, conciseness, and overall presentation. These tools were used only as writing aids and were not relied upon to determine the scientific content of the manuscript. All AI-assisted edits were carefully reviewed by the authors, who take full responsibility for the final wording, presentation, and content of the paper.

ETHICS STATEMENT

This work is primarily theoretical and evaluates the proposed methods using synthetically generated multi-objective bandit environments. The experiments do not involve human participants, personal or sensitive information, proprietary datasets, or data collected from real-world users. Consequently, the work does not raise direct ethical concerns associated with human-subject research, privacy, informed consent, or the collection and use of personal data. The proposed methods are general-purpose sequential decision-making techniques. As with related learning algorithms, however, their use in real-world or high-stakes applications should be accompanied by appropriate domain-specific validation and consideration of issues such as safety, fairness, reliability, and human oversight.

REPRODUCIBILITY STATEMENT

The main paper provides the formal problem formulation, assumptions, regret definitions, proposed algorithms, theoretical results, and experimental setup. The supplementary material contains complete proofs of the main theoretical results together with additional lemmas, derivations, examples, and technical details supporting the analysis. The experimental section specifies the synthetic reward models, problem dimensions, evaluation criteria, and comparison baselines used in the empirical study. Additional implementation details, computational procedures, and information about the computational resources and runtime are provided in the appendix. Together, the main paper and supplementary material are intended to provide sufficient information for reproducing both the theoretical analysis and the reported experimental results.

REFERENCES

Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Pizzato. Multistakeholder recommendation: Survey and research directions. *User Modeling and User-Adapted Interaction*, 30(1):127–158, 2020.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39.1–39.26. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for thompson sampling. *Journal of the ACM (JACM)*, 64(5):1–24, 2017.

Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Artificial intelligence and statistics*, pp. 939–947. PMLR, 2016.

Danton S Char, Nigam H Shah, and David Magnus. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378(11):981–983, 2018.

Mansoor Davoodi and Setareh Maghsudi. Stochastic multi-objective multi-armed bandits: Regret definition and algorithm. *Transactions on Machine Learning Research*, 2026.

Madalina Drugan and Ann Nowé. Designing multi-objective multi-armed bandits algorithms: A study. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2013.

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*, pp. 67–82. PMLR, 2018.

Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. Towards long-term fairness in recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*, pp. 445–453, 2021.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Doruk Öner, Altuğ Karakurt, Atilla Eryılmaz, and Cem Tekin. Combinatorial multi-objective multi-armed bandit problem. *arXiv preprint arXiv:1803.04039*, 2018.

Somangchan Park, Heesang Ann, and Min-hwan Oh. Thompson sampling for multi-objective linear contextual bandit. In *Advances in Neural Information Processing Systems*, 2025.

Mario Rodriguez, Christian Posse, and Ethan Zhang. Multiple objective optimization in recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*, pp. 11–18, 2012.

Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, 27, 2014.

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

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Eralp Turgay, Doruk Oner, and Cem Tekin. Multi-objective contextual bandit problem with similarity information. In *International Conference on Artificial Intelligence and Statistics*, pp. 1673–1681. PMLR, 2018.

Mengfan Xu and Diego Klabjan. Pareto regret analyses in multi-objective multi-armed bandit. In *International Conference on Machine Learning*, pp. 38499–38517. PMLR, 2023.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A.1 OPTIMIZING THE INFORMATION RATIO

At round t, IDS first constructs a sampling distribution $\pi_t^{\mathrm{IDS}} \in \Pi(\mathcal{A})$ that minimizes the information ratio and then draws $A_t \sim \pi_t^{\mathrm{IDS}}$. Specifically, π_t^{IDS} is chosen as a solution to

$$\pi_t^{\mathrm{IDS}} \in \arg\min_{\pi \in \Pi(\mathcal{A})} \frac{\left(\pi^\top \vec{\Delta}_t\right)^2}{\pi^\top \vec{I}_t}. \tag{3}$$

Here, $\vec{\Delta}_t \in \mathbb{R}^{|\mathcal{A}|}$ and $\vec{I}_t \in \mathbb{R}^{|\mathcal{A}|}$ denote the vectors of posterior expected one-step regret and information gain, respectively, with entries $\Delta_t(a)$ and $I_t(a)$ for $a \in \mathcal{A}$. The set

$$\Pi(\mathcal{A}) = \left\{ \pi \in \mathbb{R}_+^{|\mathcal{A}|} : \sum_{a \in \mathcal{A}} \pi(a) = 1 \right\}$$

denotes the probability simplex over the action set. We use the extended-value convention

$$\frac{x^2}{y} = \begin{cases} x^2/y, & y > 0, \\ 0, & x = 0,\ y = 0, \\ +\infty, & x > 0,\ y = 0, \end{cases} \qquad x, y \geq 0.$$

Thus, the information ratio remains well defined even when a candidate sampling distribution yields zero information gain.

Algorithm 5: Computation of the IDS sampling distribution via two-point mixtures

Function `IDSDistribution`$(\mathcal{A}, \vec{\Delta}_t, \vec{I}_t)$**:**
- **if** $|\mathcal{A}| = 1$ **then**
  - **return** the unique distribution in $\Pi(\mathcal{A})$
- **if** $\vec{I}_t = 0$ **then**
  - **if** *there exists* $a \in \mathcal{A}$ *with* $\Delta_t(a) = 0$ **then**
    - **return** the point mass on any such action a
  - **else**
    - **return** any distribution in $\Pi(\mathcal{A})$
- **foreach** $a, a' \in \mathcal{A}$ *with* $a \neq a'$ **do**
  - Compute
    $$q_{a,a'} \in \arg\min_{q \in [0,1]} \frac{\left(q\Delta_t(a) + (1-q)\Delta_t(a')\right)^2}{qI_t(a) + (1-q)I_t(a')},$$
    using the extended-value convention above
- Select
  $$(a^*, a^{**}) \in \arg\min_{a \neq a'} \frac{\left(q_{a,a'}\Delta_t(a) + (1-q_{a,a'})\Delta_t(a')\right)^2}{q_{a,a'}I_t(a) + (1-q_{a,a'})I_t(a')}.$$
- Construct $\pi_t^{\mathrm{IDS}} \in \Pi(\mathcal{A})$ by
  $$\pi_t^{\mathrm{IDS}}(a) = \begin{cases} q_{a^*,a^{**}}, & a = a^*, \\ 1 - q_{a^*,a^{**}}, & a = a^{**}, \\ 0, & \text{otherwise.} \end{cases}$$
- **return** π_t^{IDS}

When $\vec{I}_t \neq 0$, Proposition 1 of Russo & Van Roy (2014) shows that the information-ratio objective is convex on $\{\pi \in \Pi(\mathcal{A}) : \pi^\top \vec{I}_t > 0\}$ and that an optimal solution can be chosen with support on at most two actions. If a zero-regret distribution also has zero information gain, it attains information

ratio 0 under the convention above and is already optimal. Otherwise, since $\vec{I}_t \neq 0$, there exists a feasible distribution with positive information gain and hence finite information ratio, so no zero-information distribution with positive regret can be optimal. Accordingly, in the nontrivial case with $|\mathcal{A}| \geq 2$, Algorithm 5 only needs to consider two-point mixtures. It searches over all pairs of distinct actions and, for each pair, solves the resulting one-dimensional convex optimization problem over the mixing weight $q \in [0, 1]$, using the same extended-value convention. The endpoints $q = 0$ and $q = 1$ include singleton distributions. Among these candidate mixtures, it selects one with the smallest information ratio; the resulting distribution defines π_t^{IDS}, from which A_t is sampled.

The case $|\mathcal{A}| = 1$ is trivial. If $\vec{I}_t = 0$, IDS selects any zero-regret action if one exists; otherwise every distribution has infinite information ratio and the information-ratio criterion is uninformative at that history.

A.2 Proof of Lemma 3.1

Lemma 3.1. ***(Lower bound on information gain via average pairwise distinguishability)*** *Fix a round t, any realizable history $\mathcal{H}_t$, and an action $a \in \mathcal{A}$. Let $S, S' \overset{\text{i.i.d.}}{\sim} p_t$. Then*

$$I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \;\geq\; \frac{1}{2}\,\mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\Big[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\Big]. \tag{1}$$

Proof. Fix a round t, any realizable history $\mathcal{H}_t$, and an action $a \in \mathcal{A}$ throughout the proof. Quantities conditioned on $A_t = a$ are understood under the fixed-action experiment in which action a is selected at round t; whenever $\mathbb{P}(A_t = a \mid \mathcal{H}_t) > 0$, this agrees with ordinary conditioning on the event $A_t = a$.

Let

$$p_t(s) := \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t)$$

denote the posterior distribution of the random Pareto set $\mathcal{S}^*$, and let

$$\mathrm{supp}(p_t) := \{s \in 2^{\mathcal{A}} : p_t(s) > 0\}.$$

Because the policy selects A_t using only $\mathcal{H}_t$ and independent internal randomization, A_t is conditionally independent of $\mathcal{S}^*$ given $\mathcal{H}_t$. Hence selecting action a at round t does not change the posterior distribution of $\mathcal{S}^*$:

$$\mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t, A_t = a) = \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t) = p_t(s).$$

Draw $S, S' \sim p_t$ independently, and let $B \sim \mathrm{Bernoulli}(1/2)$ be independent of (S, S'), all conditional on $\mathcal{H}_t$ and independently of the policy randomization at round t. Define

$$\widehat{S} := \begin{cases} S, & B = 0, \\ S', & B = 1. \end{cases}$$

Conditional on $(\widehat{S}, \mathcal{H}_t, A_t = a)$, generate an auxiliary observation $\widetilde{\mathbf{Y}}$ according to

$$\widetilde{\mathbf{Y}} \mid (\widehat{S} = s, \mathcal{H}_t, A_t = a) \sim P^a_{t,s},$$

conditionally independently of (S, S', B) given $(\widehat{S}, \mathcal{H}_t, A_t = a)$.

Since $\widehat{S} = S$ when $B = 0$ and $\widehat{S} = S'$ when $B = 1$, for every $s \in \mathrm{supp}(p_t)$,

$$\{\widehat{S} = s\} = \{B = 0, S = s\} \cup \{B = 1, S' = s\},$$

where the two events are disjoint. Therefore,

$$\mathbb{P}(\widehat{S} = s \mid \mathcal{H}_t) = \mathbb{P}(B = 0, S = s \mid \mathcal{H}_t) + \mathbb{P}(B = 1, S' = s \mid \mathcal{H}_t).$$

Since B is independent of (S, S') and $S, S' \sim p_t$,

$$\mathbb{P}(B = 0, S = s \mid \mathcal{H}_t) = \mathbb{P}(B = 0)\,\mathbb{P}(S = s \mid \mathcal{H}_t) = \tfrac{1}{2} p_t(s),$$

and similarly,

$$\mathbb{P}(B = 1, S' = s \mid \mathcal{H}_t) = \tfrac{1}{2} p_t(s).$$

Hence,

$$\mathbb{P}(\widehat{S} = s \mid \mathcal{H}_t) = p_t(s).$$

Moreover, because (S, S', B) are constructed independently of the policy randomization conditional on $\mathcal{H}_t$,

$$\mathbb{P}(\widehat{S} = s \mid \mathcal{H}_t, A_t = a) = p_t(s).$$

Next, by construction of $\widetilde{\mathbf{Y}}$, for any measurable set E and every $s \in \mathrm{supp}(p_t)$,

$$\mathbb{P}(\widetilde{\mathbf{Y}} \in E \mid \widehat{S} = s, \mathcal{H}_t, A_t = a) = P^a_{t,s}(E).$$

Therefore,

$$\begin{aligned}&\mathbb{P}(\widehat{S} = s, \widetilde{\mathbf{Y}} \in E \mid \mathcal{H}_t, A_t = a)\\ &\quad = \mathbb{P}(\widehat{S} = s \mid \mathcal{H}_t, A_t = a) P^a_{t,s}(E) = p_t(s) P^a_{t,s}(E).\end{aligned}$$

By the definitions of $p_t(s)$ and $P^a_{t,s}$, together with the conditional independence of A_t and $\mathcal{S}^*$ given $\mathcal{H}_t$, we also have

$$\mathbb{P}(\mathcal{S}^* = s, \mathbf{Y}_{t,a} \in E \mid \mathcal{H}_t, A_t = a) = p_t(s) P^a_{t,s}(E).$$

Thus, the conditional joint distributions coincide:

$$\mathcal{L}(\widehat{S}, \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) = \mathcal{L}(\mathcal{S}^*, \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a).$$

Since conditional mutual information depends only on the corresponding conditional joint distribution,

$$I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) = I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a). \tag{4}$$

Next, by the conditional-independence construction of $\widetilde{\mathbf{Y}}$,

$$\widetilde{\mathbf{Y}} \perp\!\!\!\perp (S, S', B) \mid (\widehat{S}, \mathcal{H}_t, A_t = a),$$

and therefore

$$I(S, S', B; \widetilde{\mathbf{Y}} \mid \widehat{S}, \mathcal{H}_t, A_t = a) = 0. \tag{5}$$

Applying the chain rule gives

$$\begin{aligned}I(\widehat{S}, S, S', B; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) &= I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a)\\ &\quad + I(S, S', B; \widetilde{\mathbf{Y}} \mid \widehat{S}, \mathcal{H}_t, A_t = a)\\ &= I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a),\end{aligned} \tag{6}$$

where the second equality uses Equation (5).

Since $\widehat{S}$ is a deterministic function of (S, S', B), adding $\widehat{S}$ does not introduce any additional information beyond (S, S', B), and hence

$$I(\widehat{S}, S, S', B; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) = I(S, S', B; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a). \tag{7}$$

Combining Equation (6) and Equation (7), we obtain

$$I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) = I(S, S', B; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a). \tag{8}$$

Applying the chain rule once more,

$$I(S, S', B; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) = I(S, S'; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) + I(B; \widetilde{\mathbf{Y}} \mid S, S', \mathcal{H}_t, A_t = a). \tag{9}$$

Thus,

$$I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) = I(S, S'; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) + I(B; \widetilde{\mathbf{Y}} \mid S, S', \mathcal{H}_t, A_t = a). \tag{10}$$

Since mutual information is nonnegative, the first term on the right-hand side of Equation (10) can be dropped, yielding

$$I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) \geq I(B; \widetilde{\mathbf{Y}} \mid S, S', \mathcal{H}_t, A_t = a). \tag{11}$$

By construction, conditional on $(\mathcal{H}_t, A_t = a)$, the auxiliary variables S and S' remain independent with common law p_t. Hence, expanding the conditional mutual information over (S, S') gives

$$\begin{aligned} &I(B; \widetilde{\mathbf{Y}} \mid S, S', \mathcal{H}_t, A_t = a) \\ &\quad = \sum_{s,s' \in \mathrm{supp}(p_t)} p_t(s) p_t(s')\, I(B; \widetilde{\mathbf{Y}} \mid S = s, S' = s', \mathcal{H}_t, A_t = a). \end{aligned} \tag{12}$$

Therefore, from Equation (11),

$$I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) \geq \sum_{s,s' \in \mathrm{supp}(p_t)} p_t(s) p_t(s')\, I(B; \widetilde{\mathbf{Y}} \mid S = s, S' = s', \mathcal{H}_t, A_t = a). \tag{13}$$

Fix $s, s' \in \mathrm{supp}(p_t)$ and define

$$P := P^a_{t,s}, \qquad Q := P^a_{t,s'}, \qquad M := \tfrac{1}{2}(P + Q).$$

Conditional on $(S = s, S' = s', \mathcal{H}_t, A_t = a)$, the auxiliary observation $\widetilde{\mathbf{Y}}$ is drawn from P when $B = 0$ and from Q when $B = 1$. Thus, $\widetilde{\mathbf{Y}}$ has mixture distribution $M = \frac{1}{2}(P + Q)$, where B indicates whether the observation is drawn from P or Q. It follows that

$$I(B; \widetilde{\mathbf{Y}} \mid S = s, S' = s', \mathcal{H}_t, A_t = a) = \tfrac{1}{2}\mathrm{KL}(P \| M) + \tfrac{1}{2}\mathrm{KL}(Q \| M), \tag{14}$$

which is the Jensen–Shannon divergence between P and Q.

Using natural logarithms, Pinsker's inequality gives, for any distributions U, V,

$$\mathrm{KL}(U \| V) \geq 2\,\mathrm{TV}(U, V)^2.$$

Applying this inequality to both terms in Equation (14) yields

$$\tfrac{1}{2}\mathrm{KL}(P \| M) + \tfrac{1}{2}\mathrm{KL}(Q \| M) \geq \mathrm{TV}(P, M)^2 + \mathrm{TV}(Q, M)^2.$$

Since $M = \frac{1}{2}(P + Q)$,

$$\mathrm{TV}(P, M) = \mathrm{TV}(Q, M) = \tfrac{1}{2}\mathrm{TV}(P, Q).$$

Therefore,

$$I(B; \widetilde{\mathbf{Y}} \mid S = s, S' = s', \mathcal{H}_t, A_t = a) \geq \frac{1}{2}\mathrm{TV}\big(P^a_{t,s}, P^a_{t,s'}\big)^2. \tag{15}$$

Substituting Equation (15) into Equation (13), we obtain

$$\begin{aligned} I(\widehat{S}; \widetilde{\mathbf{Y}} \mid \mathcal{H}_t, A_t = a) &\geq \sum_{s,s' \in \mathrm{supp}(p_t)} p_t(s) p_t(s')\, I(B; \widetilde{\mathbf{Y}} \mid S = s, S' = s', \mathcal{H}_t, A_t = a) \\ &\geq \frac{1}{2} \sum_{s,s' \in \mathrm{supp}(p_t)} p_t(s) p_t(s')\, \mathrm{TV}\big(P^a_{t,s}, P^a_{t,s'}\big)^2 \\ &= \frac{1}{2}\mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\Big[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\Big]. \end{aligned} \tag{16}$$

Finally, combining this inequality with Equation (4), we conclude

$$I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \geq \frac{1}{2}\mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\Big[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\Big],$$

which completes the proof. □

A.3 Proof of Theorem 3.1

We restate Theorem 3.1, explicitly specifying the necessary regularity and non-degeneracy conditions, and present a complete formal proof.

Theorem A.1. *Consider a multi-objective bandit model with finite discrete reward alphabet $\mathcal{Y} \subset \mathbb{R}^d$ and finite action set $\mathcal{A}$, leading to Dirichlet posterior updates over the parameter $\boldsymbol{\theta}$. For each $a \in \mathcal{A}$, $\boldsymbol{\theta} \in \Theta$, and $y \in \mathcal{Y}$, write*

$$P^a_{\boldsymbol{\theta}}(y) := \mathbb{P}_{\boldsymbol{\theta}}(\mathbf{Y}_{t,a} = y)$$

for the conditional reward probability under action a. Let $q \in \Pi(\mathcal{A})$ have full support, i.e., $q(a) > 0$ for all $a \in \mathcal{A}$.

Fix a Dirichlet posterior density family $\{p_\alpha : \alpha \in \mathcal{H} \subset \mathbb{R}^m\}$ with $\mathcal{H}$ open, where all densities are taken with respect to a fixed σ-finite reference measure ν on Θ. Under the Dirichlet count-update model, for every round t and every realizable history $\mathcal{H}_t$, the posterior parameter can be written as

$$\alpha_t = \alpha_0 + N_t, \qquad N_t \in \mathbb{N}_0^m,$$

where

$$\mathbb{N}_0 := \{0, 1, 2, \ldots\},$$

and $\alpha_t \in \mathcal{H}$.

Let $\mathcal{S} := 2^{\mathcal{A}}$ denote the collection of all subsets of $\mathcal{A}$, and let $\mathcal{S}^*$ be the random Pareto set taking values in $\mathcal{S}$. Let Θ denote the parameter space of $\boldsymbol{\theta}$, and for each $s \in \mathcal{S}$ define

$$\Theta_s := \{\boldsymbol{\theta} \in \Theta : \mathcal{S}^*(\boldsymbol{\theta}) = s\},$$

i.e., the set of parameters for which the Pareto set equals s. Assume that each Θ_s is measurable. For $\alpha \in \mathcal{H}$, define the posterior probability of Pareto-set candidate s by

$$p_\alpha(s) := p_\alpha(\Theta_s) := \int_{\Theta_s} p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta}).$$

For each pair $s \neq s'$ with $s, s' \in \mathcal{S}$, define the admissible parameter set

$$\mathcal{H}_{s,s'} := \{\alpha \in \mathcal{H} : p_\alpha(s) > 0,\ p_\alpha(s') > 0\}.$$

Assume that for every $s \neq s'$ with $\mathcal{H}_{s,s'} \neq \varnothing$:

1. ***(Analytic parameterization.)*** *For each $(a, y) \in \mathcal{A} \times \mathcal{Y}$ and each $r \in \{s, s'\}$, the maps*

$$\alpha \mapsto \int_{\Theta_r} p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta}), \qquad \alpha \mapsto \int_{\Theta_r} P^a_{\boldsymbol{\theta}}(y) p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})$$

   are real-analytic on $\mathcal{H}$.

2. ***(Witness posterior exists in each component.)*** *For every connected component $\mathcal{C}$ of $\mathcal{H}_{s,s'}$, there exist $\alpha^\star \in \mathcal{C}$ and $a \in \mathcal{A}$ such that*

$$P^a_{p_{\alpha^\star},s} \neq P^a_{p_{\alpha^\star},s'}.$$

   Here, for $\alpha \in \mathcal{H}_{s,s'}$, the predictive reward distribution under action a conditioned on Pareto-set candidate s is

$$P^a_{p_\alpha,s}(y) := \frac{\int_{\Theta_s} P^a_{\boldsymbol{\theta}}(y) p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})}{\int_{\Theta_s} p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})}, \qquad y \in \mathcal{Y}.$$

For a realizable history $\mathcal{H}_t$, let

$$p_t(s) := p_{\alpha_t}(s) = \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t),$$

and let

$$S, S' \overset{\text{i.i.d.}}{\sim} p_t$$

conditional on $\mathcal{H}_t$. Then there exists a Lebesgue measure-zero set $\mathcal{B} \subset \mathcal{H}$ such that, for every $\alpha_0 \in \mathcal{H} \setminus \mathcal{B}$, every round t, and every realizable history $\mathcal{H}_t$,

$$\mathbb{E}\Big[\mathbb{E}_{a\sim q}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right] \,\Big|\, S \neq S', \mathcal{H}_t\Big] > 0 \tag{17}$$

whenever

$$\mathbb{P}(S \neq S' \mid \mathcal{H}_t) > 0,$$

where $P^a_{t,S} := P^a_{p_{\alpha_t},S}$ and TV *denotes total variation distance.*

Proof. Let λ_m denote m-dimensional Lebesgue measure on $\mathbb{R}^m$. Fix $s \neq s'$ with $\mathcal{H}_{s,s'} \neq \varnothing$.

By condition 1, the map

$$\alpha \mapsto p_\alpha(\Theta_r) = \int_{\Theta_r} p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})$$

is real-analytic and hence continuous on $\mathcal{H}$ for $r \in \{s, s'\}$. Therefore, the positivity constraints $p_\alpha(\Theta_s) > 0$ and $p_\alpha(\Theta_{s'}) > 0$ define an open subset of $\mathcal{H}$. Since $\mathcal{H}$ is open in $\mathbb{R}^m$, $\mathcal{H}_{s,s'}$ is open in $\mathbb{R}^m$.

Fix $(a, y) \in \mathcal{A} \times \mathcal{Y}$. For $\alpha \in \mathcal{H}_{s,s'}$, define

$$g_{a,y}(\alpha) := \Big(\int_{\Theta_s} P^a_{\boldsymbol{\theta}}(y) p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})\Big)\Big(\int_{\Theta_{s'}} p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})\Big) - \Big(\int_{\Theta_{s'}} P^a_{\boldsymbol{\theta}}(y) p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})\Big)\Big(\int_{\Theta_s} p_\alpha(\boldsymbol{\theta})\, \nu(d\boldsymbol{\theta})\Big).$$

By condition 1, each integral is real-analytic on $\mathcal{H}$, and hence $g_{a,y}$ is real-analytic on $\mathcal{H}_{s,s'}$.

For $\alpha \in \mathcal{H}_{s,s'}$, we have $p_\alpha(\Theta_s) > 0$ and $p_\alpha(\Theta_{s'}) > 0$, so cross-multiplication is valid. Moreover, since $\mathcal{Y}$ is finite,

$$P^a_{p_\alpha,s} = P^a_{p_\alpha,s'} \iff P^a_{p_\alpha,s}(y) = P^a_{p_\alpha,s'}(y) \quad \forall y \in \mathcal{Y} \iff g_{a,y}(\alpha) = 0 \quad \forall y \in \mathcal{Y}.$$

Thus, the no-separation set can be written as

$$\mathcal{E}_{s,s'} := \Big\{\alpha \in \mathcal{H}_{s,s'} : P^a_{p_\alpha,s} = P^a_{p_\alpha,s'} \ \forall a \in \mathcal{A}\Big\} = \bigcap_{(a,y)\in\mathcal{A}\times\mathcal{Y}} \Big\{\alpha \in \mathcal{H}_{s,s'} : g_{a,y}(\alpha) = 0\Big\}.$$

Let $\{\mathcal{C}_j\}_{j\in J}$ denote the connected components of $\mathcal{H}_{s,s'}$. Since $\mathcal{H}_{s,s'}$ is an open subset of $\mathbb{R}^m$, each connected component $\mathcal{C}_j$ is open, and the collection of connected components is at most countable. Indeed, every nonempty open component contains a point of $\mathbb{Q}^m$. Fix one component $\mathcal{C}_j$.

By condition 2, there exist $\alpha^\star_j \in \mathcal{C}_j$ and $a_j \in \mathcal{A}$ such that

$$P^{a_j}_{p_{\alpha^\star_j},s} \neq P^{a_j}_{p_{\alpha^\star_j},s'}.$$

Since $\mathcal{Y}$ is finite, there exists $y_j \in \mathcal{Y}$ with

$$P^{a_j}_{p_{\alpha^\star_j},s}(y_j) \neq P^{a_j}_{p_{\alpha^\star_j},s'}(y_j),$$

equivalently,

$$g_{a_j,y_j}(\alpha^\star_j) \neq 0.$$

Hence g_{a_j,y_j} is not identically zero on the connected open set $\mathcal{C}_j$.

Let

$$Z_j := \{\alpha \in \mathcal{C}_j : g_{a_j,y_j}(\alpha) = 0\}.$$

Because $\mathcal{C}_j$ is open and connected and g_{a_j,y_j} is real-analytic and not identically zero on $\mathcal{C}_j$, the standard analytic zero-set property implies

$$\lambda_m(Z_j) = 0.$$

Since

$$\mathcal{E}_{s,s'} \cap \mathcal{C}_j \subseteq Z_j,$$

we conclude

$$\lambda_m(\mathcal{E}_{s,s'} \cap \mathcal{C}_j) = 0.$$

Since $\mathcal{H}_{s,s'}$ has at most countably many connected components,

$$\mathcal{E}_{s,s'} = \bigcup_{j\in J} (\mathcal{E}_{s,s'} \cap \mathcal{C}_j)$$

is a countable union of measure-zero sets. Therefore,

$$\lambda_m(\mathcal{E}_{s,s'}) = 0.$$

For each $n \in \mathbb{N}_0^m$, define

$$\mathcal{B}_n^{(s,s')} := \{\alpha_0 \in \mathcal{H} : \alpha_0 + n \in \mathcal{E}_{s,s'}\}.$$

Equivalently,

$$\mathcal{B}_n^{(s,s')} = (\mathcal{E}_{s,s'} - n) \cap \mathcal{H}.$$

Translation by n preserves Lebesgue-null sets, and taking an intersection with $\mathcal{H}$ preserves nullity. Hence

$$\lambda_m(\mathcal{B}_n^{(s,s')}) = 0.$$

Now define

$$\mathcal{B}^{(s,s')} := \bigcup_{n \in \mathbb{N}_0^m} \mathcal{B}_n^{(s,s')}.$$

Since $\mathbb{N}_0^m$ is countable and each $\mathcal{B}_n^{(s,s')}$ has measure zero,

$$\lambda_m(\mathcal{B}^{(s,s')}) = 0.$$

Finally, define

$$\mathcal{B} := \bigcup_{\substack{s \neq s' \\ \mathcal{H}_{s,s'} \neq \varnothing}} \mathcal{B}^{(s,s')}.$$

Because $\mathcal{S} = 2^{\mathcal{A}}$ is finite, this is a finite union of measure-zero sets, hence

$$\lambda_m(\mathcal{B}) = 0.$$

Fix $\alpha_0 \in \mathcal{H} \setminus \mathcal{B}$ and consider any round t and any realizable history $\mathcal{H}_t$, with posterior parameter

$$\alpha_t = \alpha_0 + N_t.$$

Let $s \neq s'$ be any pair with

$$p_t(s)p_t(s') > 0.$$

Then necessarily

$$\alpha_t \in \mathcal{H}_{s,s'},$$

since

$$p_{\alpha_t}(\Theta_s) = p_t(s) > 0, \qquad p_{\alpha_t}(\Theta_{s'}) = p_t(s') > 0.$$

In particular, $\mathcal{H}_{s,s'} \neq \varnothing$, so $\mathcal{B}^{(s,s')} \subseteq \mathcal{B}$. Hence $\alpha_0 \notin \mathcal{B}^{(s,s')}$. If $\alpha_t \in \mathcal{E}_{s,s'}$, then $\alpha_t = \alpha_0 + N_t$ with $N_t \in \mathbb{N}_0^m$ would imply

$$\alpha_0 \in \mathcal{B}_{N_t}^{(s,s')} \subseteq \mathcal{B}^{(s,s')},$$

a contradiction. Therefore,

$$\alpha_t \notin \mathcal{E}_{s,s'}.$$

By the definition of $\mathcal{E}_{s,s'}$, it is therefore not the case that

$$P^a_{p_{\alpha_t},s} = P^a_{p_{\alpha_t},s'} \qquad \forall a \in \mathcal{A}.$$

Hence, there exists some $a^\star \in \mathcal{A}$ such that

$$P^{a^\star}_{p_{\alpha_t},s} \neq P^{a^\star}_{p_{\alpha_t},s'}.$$

Since total variation distance is zero if and only if the two probability distributions are equal,

$$\mathrm{TV}\Big(P^{a^\star}_{p_{\alpha_t},s}, P^{a^\star}_{p_{\alpha_t},s'}\Big) > 0.$$

Since q has full support,

$$q(a^\star) > 0.$$

Therefore,

$$\mathbb{E}_{a\sim q}\left[\mathrm{TV}\Big(P^a_{p_{\alpha_t},s}, P^a_{p_{\alpha_t},s'}\Big)^2\right] \geq q(a^\star)\mathrm{TV}\Big(P^{a^\star}_{p_{\alpha_t},s}, P^{a^\star}_{p_{\alpha_t},s'}\Big)^2 > 0.$$

Thus, for every pair $s \neq s'$ with $p_t(s)p_t(s') > 0$, the average squared predictive separation is strictly positive.

For $s, s' \in \mathrm{supp}(p_t)$, define

$$D_t(s, s') := \mathbb{E}_{a\sim q}\left[\mathrm{TV}\Big(P^a_{p_{\alpha_t},s}, P^a_{p_{\alpha_t},s'}\Big)^2\right],$$

where

$$\mathrm{supp}(p_t) := \{s \in \mathcal{S} : p_t(s) > 0\}.$$

This quantity is well defined because $p_t(s) > 0$ and $p_t(s') > 0$. We have shown that

$$D_t(s, s') > 0$$

whenever

$$s, s' \in \mathrm{supp}(p_t), \qquad s \neq s'.$$

Let

$$S, S' \overset{\text{i.i.d.}}{\sim} p_t$$

conditional on $\mathcal{H}_t$, and assume

$$G_t := \mathbb{P}(S \neq S' \mid \mathcal{H}_t) > 0.$$

Because S and S' are conditionally independent draws from p_t,

$$G_t = \sum_{\substack{s,s'\in\mathrm{supp}(p_t)\\ s\neq s'}} p_t(s)p_t(s').$$

Furthermore, for $s, s' \in \mathrm{supp}(p_t)$ with $s \neq s'$,

$$\begin{aligned}\mathbb{P}(S = s, S' = s' \mid S \neq S', \mathcal{H}_t) &= \frac{\mathbb{P}(S = s, S' = s', S \neq S' \mid \mathcal{H}_t)}{\mathbb{P}(S \neq S' \mid \mathcal{H}_t)} \\ &= \frac{\mathbb{P}(S = s, S' = s' \mid \mathcal{H}_t)}{G_t} = \frac{p_t(s)p_t(s')}{G_t},\end{aligned}$$

where the second equality uses $\{S = s, S' = s'\} \subseteq \{S \neq S'\}$ whenever $s \neq s'$.

Therefore,

$$\begin{aligned}\mathbb{E}[D_t(S, S') \mid S \neq S', \mathcal{H}_t] &= \sum_{\substack{s,s'\in\mathrm{supp}(p_t)\\ s\neq s'}} D_t(s, s')\,\mathbb{P}(S = s, S' = s' \mid S \neq S', \mathcal{H}_t) \\ &= \frac{1}{G_t}\sum_{\substack{s,s'\in\mathrm{supp}(p_t)\\ s\neq s'}} p_t(s)p_t(s')D_t(s, s').\end{aligned}$$

Every term in the final sum is strictly positive. Moreover, $G_t > 0$ implies that at least one distinct pair in $\mathrm{supp}(p_t)$ exists. Therefore,

$$\mathbb{E}[D_t(S, S') \mid S \neq S', \mathcal{H}_t] > 0.$$

Substituting the definition of $D_t(S, S')$, we obtain

$$\mathbb{E}\left[\mathbb{E}_{a\sim q}\left[\mathrm{TV}\Big(P^a_{p_{\alpha_t},S}, P^a_{p_{\alpha_t},S'}\Big)^2\right] \,\middle|\, S \neq S', \mathcal{H}_t\right] > 0,$$

which is exactly Equation (17).

Since $\lambda_m(\mathcal{B}) = 0$, the above argument holds for every Dirichlet prior parameter $\alpha_0 \in \mathcal{H} \setminus \mathcal{B}$. By construction, $\mathcal{B}$ excludes every initial parameter α_0 for which some count update $\alpha_0 + n$, with $n \in \mathbb{N}_0^m$, lies in a no-separation set for any admissible pair (s, s'). Since every realizable posterior parameter has the form

$$\alpha_t = \alpha_0 + N_t, \qquad N_t \in \mathbb{N}_0^m,$$

the preceding argument applies simultaneously to every round t and every realizable history $\mathcal{H}_t$. This completes the proof. □

A.4 A CONCRETE BETA–BERNOULLI MODEL SATISFYING THEOREM 3.1

We verify the regularity and witness conditions of Theorem A.1 in a standard two-action, two-objective Bayesian bandit. Let

$$\mathcal{A} = \{a_1, a_2\}, \qquad \mathbf{Y}_{t,a_i} = (Y_{t,a_i,1}, Y_{t,a_i,2}) \in \{0,1\}^2.$$

Conditional on

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}),$$

assume the reward coordinates are independent and

$$Y_{t,a_i,j} \mid \boldsymbol{\theta} \sim \text{Bernoulli}(\theta_{ij}), \qquad i,j \in \{1,2\}.$$

Place independent Beta priors

$$\theta_{ij} \sim \text{Beta}(\alpha_{ij}, \beta_{ij}), \qquad \alpha_{ij}, \beta_{ij} > 0.$$

Recall that the Beta distribution is the two-category special case of the Dirichlet distribution. Thus, each Bernoulli parameter has a Dirichlet conjugate factor. Collect the hyperparameters into

$$\alpha = (\alpha_{11}, \beta_{11}, \alpha_{12}, \beta_{12}, \alpha_{21}, \beta_{21}, \alpha_{22}, \beta_{22}) \in \mathcal{H}, \qquad \mathcal{H} = (0,\infty)^8.$$

The domain $\mathcal{H}$ is open and connected. After any realizable history, the corresponding success and failure counts can be collected into $N_t \in \mathbb{N}_0^8$, and conjugacy gives the posterior update

$$\alpha_t = \alpha_0 + N_t,$$

which is exactly the count-update parameterization required by Theorem A.1. Here, the reference measure ν in Theorem A.1 is ordinary Lebesgue measure on $(0,1)^4$.

The three possible Pareto sets are

$$s_1 = \{a_1\}, \qquad s_2 = \{a_2\}, \qquad s_{12} = \{a_1, a_2\}.$$

The interiors of their corresponding parameter regions are, respectively,

$$\{\theta_{11} > \theta_{21},\ \theta_{12} > \theta_{22}\},$$

$$\{\theta_{11} < \theta_{21},\ \theta_{12} < \theta_{22}\},$$

and

$$\{\theta_{11} > \theta_{21},\ \theta_{12} < \theta_{22}\} \cup \{\theta_{11} < \theta_{21},\ \theta_{12} > \theta_{22}\}.$$

The omitted tie boundaries have Lebesgue measure zero, and each displayed region has positive Lebesgue measure. Since every Beta density with positive parameters is strictly positive on $(0,1)$, it follows that

$$p_\alpha(s_1) > 0, \qquad p_\alpha(s_2) > 0, \qquad p_\alpha(s_{12}) > 0$$

for every $\alpha \in \mathcal{H}$. Hence, for every distinct $s, s' \in \{s_1, s_2, s_{12}\}$,

$$\mathcal{H}_{s,s'} = \mathcal{H},$$

so every admissible set $\mathcal{H}_{s,s'}$ has a single connected component.

The joint prior density has the form

$$p_\alpha(\boldsymbol{\theta}) = \prod_{i,j\in\{1,2\}} \frac{\theta_{ij}^{\alpha_{ij}-1}(1-\theta_{ij})^{\beta_{ij}-1}}{B(\alpha_{ij},\beta_{ij})}.$$

We next verify condition 1. Fix any $\alpha^0 \in \mathcal{H}$. Because all coordinates of α^0 are strictly positive, there is a sufficiently small complex polydisc $\mathcal{D}$ around α^0 on which every hyperparameter has strictly positive real part. The Beta density factors and their reciprocal normalizing constants admit holomorphic extensions in the hyperparameters on $\mathcal{D}$. Moreover, after shrinking $\mathcal{D}$ if necessary, there exist constants $C < \infty$ and $\varepsilon > 0$ such that, uniformly over the hyperparameters in $\mathcal{D}$,

$$|p_\alpha(\boldsymbol{\theta})| \le C \prod_{i,j\in\{1,2\}} \theta_{ij}^{\varepsilon-1}(1-\theta_{ij})^{\varepsilon-1},$$

and the right-hand side is integrable on $(0,1)^4$. Since $0 \le P^a_{\boldsymbol{\theta}}(y) \le 1$ for $y \in \{0,1\}^2$, the same envelope applies to the predictive numerators. The dominated holomorphic integration theorem therefore implies that, for every Pareto region Θ_s, action a, and reward $y \in \{0,1\}^2$, the maps

$$\alpha \mapsto \int_{\Theta_s} p_\alpha(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad \alpha \mapsto \int_{\Theta_s} P^a_{\boldsymbol{\theta}}(y) p_\alpha(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

are real-analytic on $\mathcal{H}$. Thus, condition 1 holds.

It remains to verify the witness condition. Consider the symmetric prior

$$\alpha_{ij} = \beta_{ij} = 1 \qquad \text{for all } i, j,$$

under which the four parameters are independent $\mathrm{Unif}(0,1)$ random variables. If U, V are independent $\mathrm{Unif}(0,1)$ variables, then $\mathbb{P}(U > V) = 1/2$ and

$$\mathbb{E}[U \mid U > V] = \frac{\int_0^1 \int_0^u u\, dv\, du}{1/2} = \frac{1/3}{1/2} = \frac{2}{3}.$$

By symmetry,

$$\mathbb{E}[U \mid U < V] = \frac{1}{3}.$$

Therefore,

$$\mathbb{E}[\theta_{11} \mid \mathcal{S}^* = s_1] = \frac{2}{3}, \qquad \mathbb{E}[\theta_{11} \mid \mathcal{S}^* = s_2] = \frac{1}{3}.$$

Under s_{12}, the two trade-off regions have equal probability by symmetry. In one region $\theta_{11} > \theta_{21}$ and in the other $\theta_{11} < \theta_{21}$, so

$$\mathbb{E}[\theta_{11} \mid \mathcal{S}^* = s_{12}] = \frac{1}{2}\left(\frac{2}{3} + \frac{1}{3}\right) = \frac{1}{2}.$$

For action a_1, the first reward coordinate is Bernoulli conditional on θ_{11}. After conditioning only on a Pareto-set candidate s, its posterior-predictive marginal is therefore Bernoulli with success probability $\mathbb{E}[\theta_{11} \mid \mathcal{S}^* = s]$. Hence, the first-coordinate posterior-predictive marginals have distinct success probabilities $2/3$, $1/3$, and $1/2$. Therefore, the corresponding joint posterior-predictive reward laws are pairwise distinct:

$$P^{a_1}_{p_{\alpha^\star}, s_1} \neq P^{a_1}_{p_{\alpha^\star}, s_2}, \qquad P^{a_1}_{p_{\alpha^\star}, s_1} \neq P^{a_1}_{p_{\alpha^\star}, s_{12}}, \qquad P^{a_1}_{p_{\alpha^\star}, s_2} \neq P^{a_1}_{p_{\alpha^\star}, s_{12}},$$

where $\alpha^\star$ denotes the symmetric hyperparameter vector. Thus, action a_1 is a witness for every pair of competing Pareto-set candidate at $\alpha^\star$. Because each $\mathcal{H}_{s,s'} = \mathcal{H}$ is connected, condition 2 holds for every pair.

Consequently, this standard Beta–Bernoulli multi-objective bandit satisfies the regularity and non-degeneracy conditions used in Theorem A.1.

A.5 A PRACTICAL SUFFICIENT CONDITION FOR QUANTITATIVE SEPARATION

For bounded rewards, Assumptions 3.1–3.2 can be verified through posterior-predictive mean differences, which are often easier to interpret than total variation distances. All conditional expectations involving $A_t = a$ use the fixed-action interpretation from Section 3.

Proposition A.1. *Suppose $R_{\max} > 0$ and $\mathbf{Y}_{t,a} \in [0, R_{\max}]^d$ almost surely for every $a \in \mathcal{A}$. Let $(\delta_t)_{t \ge 1}$ be a deterministic positive sequence. Assume that, at every realizable history $\mathcal{H}_t$ with $G_t > 0$, for every pair $s \neq s'$ with $p_t(s)p_t(s') > 0$, there exist an action $a \in \mathcal{A}$ and coordinate $j \in [d]$ such that*

$$|\mathbb{E}[Y_{t,a,j} \mid \mathcal{H}_t, \mathcal{S}^* = s, A_t = a] - \mathbb{E}[Y_{t,a,j} \mid \mathcal{H}_t, \mathcal{S}^* = s', A_t = a]| \ge \delta_t.$$

Then

$$\kappa_t(\mathcal{H}_t) \ge \frac{\delta_t^2}{R_{\max}^2 |\mathcal{A}|}.$$

Consequently, if $\delta_t \ge \delta > 0$, then Assumption 3.1 holds with

$$\kappa = \frac{\delta^2}{R_{\max}^2 |\mathcal{A}|}.$$

If instead $\delta_t \ge \delta t^{-\gamma}$ for some $\delta > 0$ and $\gamma \ge 0$, then Assumption 3.2 holds with

$$c = \frac{\delta^2}{R_{\max}^2 |\mathcal{A}|}, \qquad \beta = 2\gamma.$$

Proof. Fix a realizable history $\mathcal{H}_t$ with $G_t > 0$ and a pair $s \neq s'$ with $p_t(s)p_t(s') > 0$. By assumption, there exist an action a and coordinate j for which the corresponding posterior-predictive means differ by at least δ_t. Let

$$P = P^a_{t,s}, \qquad Q = P^a_{t,s'}.$$

Since $0 \leq Y_{t,a,j} \leq R_{\max}$, the function $f(\mathbf{y}) = y_j / R_{\max}$ takes values in $[0, 1]$. By the variational characterization of total variation,

$$\mathrm{TV}(P, Q) \geq |\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \geq \frac{\delta_t}{R_{\max}}.$$

Therefore, for this pair,

$$\mathbb{E}_{a' \sim \mathrm{Unif}(\mathcal{A})} \left[\mathrm{TV}\Big(P^{a'}_{t,s}, P^{a'}_{t,s'} \Big)^2 \right] \geq \frac{\delta_t^2}{R_{\max}^2 |\mathcal{A}|}.$$

The same lower bound holds for every posterior-supported pair $s \neq s'$. Averaging conditional on $S \neq S'$ gives

$$\kappa_t(\mathcal{H}_t) \geq \frac{\delta_t^2}{R_{\max}^2 |\mathcal{A}|}.$$

The two stated consequences follow immediately. □

A.6 Proof of Theorem 3.2

Theorem 3.2. *The Bayesian cumulative Pareto-violation regret of Algorithm 1 satisfies:*

- *Under Assumption 3.1 (Uniform separation)* $\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \leq \left(2 + \frac{2H(\mathcal{S}^*)}{\kappa}\right)\sqrt{T}$.
- *Under Assumption 3.2 (Decaying separation)* $\mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \leq 2\sqrt{T} + \frac{2H(\mathcal{S}^*)}{c}\, T^{\beta+\frac{1}{2}}$. *In particular, under Assumption 3.2, the regret is sublinear whenever* $\beta < \frac{1}{2}$.

Proof. Recall that the instantaneous Pareto-violation loss is

$$\ell_t := \mathbb{1}\{A_t \notin \mathcal{S}^*\},$$

and that the Bayesian cumulative Pareto-violation regret is

$$\mathbb{E}\left[\sum_{t=1}^T \ell_t\right] = \mathbb{E}\left[\sum_{t=1}^T \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right].$$

Also recall the disagreement probability

$$G_t := \mathbb{P}(S \neq S' \mid \mathcal{H}_t),$$

where $S, S' \overset{\text{i.i.d.}}{\sim} p_t$, and $p_t(s) = \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t)$ denotes the posterior distribution over candidate Pareto sets, with $\mathcal{S}^* := \mathcal{S}^*(\boldsymbol{\theta})$ representing the Pareto set induced by the random parameter $\boldsymbol{\theta}$.

Fix versions of the relevant regular conditional distributions as probability kernels in the history. For every realizable history h_t, define the history-dependent one-step information gain

$$I_t(h_t) := I(\mathcal{S}^*; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t = h_t, A_t),$$

equivalently,

$$I_t(h_t) = \sum_{a \in \mathcal{A}} \mathbb{P}(A_t = a \mid \mathcal{H}_t = h_t)\, I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t = h_t, A_t = a).$$

We write $I_t := I_t(\mathcal{H}_t)$ for the corresponding $\mathcal{H}_t$-measurable random quantity.

Under Algorithm 1, we first draw

$$B_t \sim \mathrm{Bernoulli}(1 - \varepsilon_t).$$

At $t = 1$, we have $\varepsilon_1 = 1$, so $B_1 = 0$ almost surely and

$$\mathbb{E}[\ell_1 \mid \mathcal{H}_1] \le 1 = \varepsilon_1.$$

Thus, the loss bound established below also holds at $t = 1$. For the branch-conditioning argument, we may therefore assume $t \ge 2$, for which $1 - \varepsilon_t > 0$.

If $B_t = 0$, then $S_t = \mathcal{A}$ and $A_t \sim \mathrm{Unif}(\mathcal{A})$, and hence trivially

$$\ell_t \le 1.$$

If $B_t = 1$, then $\boldsymbol{\theta}_t \sim p_t(\boldsymbol{\theta})$, $S_t = \mathcal{S}^*(\boldsymbol{\theta}_t)$, and $A_t \sim \mathrm{Unif}(S_t)$. In this case, if $S_t = \mathcal{S}^*$, then $A_t \in \mathcal{S}^*$ almost surely, and thus $\ell_t = 0$. Therefore,

$$\ell_t \le \mathbb{1}\{B_t = 0\} + \mathbb{1}\{B_t = 1,\ S_t \ne \mathcal{S}^*\}.$$

Taking conditional expectation given $\mathcal{H}_t$, for $t \ge 2$ we obtain

$$\mathbb{E}[\ell_t \mid \mathcal{H}_t] \le \varepsilon_t + (1 - \varepsilon_t)\mathbb{P}(S_t \ne \mathcal{S}^* \mid \mathcal{H}_t, B_t = 1).$$

Conditional on $\mathcal{H}_t$, the branch variable B_t is generated using fresh algorithmic randomness independent of the true parameter $\boldsymbol{\theta}$. Hence, conditioning additionally on $B_t = 1$ does not change the posterior distribution of $\boldsymbol{\theta}$, so

$$\boldsymbol{\theta} \mid (\mathcal{H}_t, B_t = 1) \sim p_t(\boldsymbol{\theta}).$$

On this branch, the Thompson sample $\boldsymbol{\theta}_t \sim p_t(\boldsymbol{\theta})$ is generated using fresh randomness independent of $\boldsymbol{\theta}$ conditional on $(\mathcal{H}_t, B_t = 1)$. Therefore, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_t$ are conditionally independent and identically distributed according to $p_t(\boldsymbol{\theta})$. Consequently,

$$\mathcal{S}^* = \mathcal{S}^*(\boldsymbol{\theta}) \quad \text{and} \quad S_t = \mathcal{S}^*(\boldsymbol{\theta}_t)$$

are conditionally independent and identically distributed according to the induced posterior

$$p_t(s) = \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t).$$

Thus,

$$\mathbb{P}(S_t \ne \mathcal{S}^* \mid \mathcal{H}_t, B_t = 1) = \mathbb{P}(S \ne S' \mid \mathcal{H}_t) = G_t,$$

where $S, S' \overset{\text{i.i.d.}}{\sim} p_t$. Together with the already verified $t = 1$ case, this gives, for every t,

$$\mathbb{E}[\ell_t \mid \mathcal{H}_t] \le \varepsilon_t + (1 - \varepsilon_t)G_t. \tag{18}$$

Define $r_t := (1 - \varepsilon_t)G_t$. Then

$$\mathbb{E}[\ell_t \mid \mathcal{H}_t] \le \varepsilon_t + r_t, \qquad \text{with} \qquad r_t \le G_t. \tag{19}$$

Fix t, and let

$$\pi_t(a) := \mathbb{P}(A_t = a \mid \mathcal{H}_t), \qquad a \in \mathcal{A},$$

denote the conditional action distribution induced by Algorithm 1. Since with probability ε_t the algorithm explores uniformly over $\mathcal{A}$, we have

$$\pi_t(a) \ge \frac{\varepsilon_t}{|\mathcal{A}|} \qquad \text{for all } a \in \mathcal{A}. \tag{20}$$

By the definition of I_t,

$$I_t = \sum_{a \in \mathcal{A}} \pi_t(a)\, I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a).$$

Applying Lemma 3.1 to each action a, we obtain

$$I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \ge \frac{1}{2}\mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right].$$

Therefore,

$$I_t \ge \frac{1}{2}\sum_{a \in \mathcal{A}} \pi_t(a)\, \mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right].$$

Using Equation (20) and exchanging the finite sum and expectation, we obtain

$$I_t \geq \frac{\varepsilon_t}{2|\mathcal{A}|} \sum_{a \in \mathcal{A}} \mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t} \left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \right] = \frac{\varepsilon_t}{2} \mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t} \left[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})} \mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \right].$$

If $G_t = 0$, then $r_t = 0$, so all subsequent upper bounds on r_t hold trivially. We may therefore suppose $G_t > 0$. Since the total variation term vanishes whenever $S = S'$, conditioning on the event $\{S \neq S'\}$ gives

$$I_t \geq \frac{\varepsilon_t}{2} \mathbb{P}(S \neq S' \mid \mathcal{H}_t) \mathbb{E}\Big[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})} \mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t \Big].$$

Recalling that $\mathbb{P}(S \neq S' \mid \mathcal{H}_t) = G_t$, this yields

$$I_t \geq \frac{\varepsilon_t}{2} G_t \mathbb{E}\Big[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})} \mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t \Big].$$

Case 1: By Assumption 3.1 (Uniform separation),

$$\mathbb{E}\Big[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})} \mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t \Big] \geq \kappa$$

whenever $G_t > 0$. Therefore,

$$I_t \geq \frac{\varepsilon_t \kappa}{2} G_t. \tag{21}$$

Since $r_t \leq G_t$ by Equation (19),

$$r_t \leq \frac{2}{\varepsilon_t \kappa} I_t.$$

Together with the trivial $G_t = 0$ case, this inequality holds at every realizable history. With $\varepsilon_t = t^{-1/2}$,

$$r_t \leq \frac{2\sqrt{t}}{\kappa} I_t \leq \frac{2\sqrt{T}}{\kappa} I_t. \tag{22}$$

Case 2: By Assumption 3.2 (Decaying separation),

$$\mathbb{E}\Big[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})} \mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t \Big] \geq c\, t^{-\beta}$$

whenever $G_t > 0$. Therefore,

$$I_t \geq \frac{\varepsilon_t c}{2} t^{-\beta} G_t. \tag{23}$$

Again using $r_t \leq G_t$,

$$r_t \leq \frac{2t^{\beta}}{\varepsilon_t c} I_t.$$

Together with the trivial $G_t = 0$ case, this inequality holds at every realizable history. With $\varepsilon_t = t^{-1/2}$,

$$r_t \leq \frac{2}{c} t^{\beta + \frac{1}{2}} I_t \leq \frac{2}{c} T^{\beta + \frac{1}{2}} I_t. \tag{24}$$

We now control the expected cumulative information gain about the unknown Pareto set $\mathcal{S}^*$. Since $\mathcal{H}_1$ is empty and

$$\mathcal{H}_{t+1} = (\mathcal{H}_t, A_t, \mathbf{Y}_{t,A_t}),$$

the chain rule for mutual information gives

$$I(\mathcal{S}^*; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t).$$

The action A_t is selected based on $\mathcal{H}_t$ using only independent internal randomization of the algorithm. Hence

$$I(\mathcal{S}^*; A_t \mid \mathcal{H}_t) = 0.$$

Applying the chain rule again,

$$I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) = I(\mathcal{S}^*; A_t \mid \mathcal{H}_t) + I(\mathcal{S}^*; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t) = \mathbb{E}[I_t],$$

where the final equality follows by averaging the history-dependent quantity $I_t = I_t(\mathcal{H}_t)$ over $\mathcal{H}_t$. Consequently,

$$I(\mathcal{S}^*; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} \mathbb{E}[I_t] = \mathbb{E}\left[\sum_{t=1}^{T} I_t\right].$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} I_t\right] = I(\mathcal{S}^*; \mathcal{H}_{T+1}) \le H(\mathcal{S}^*), \tag{25}$$

where the inequality follows because $\mathcal{S}^*$ is discrete and mutual information is upper bounded by entropy.

We now proceed as follows:

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\ell_t \mid \mathcal{H}_t]\right] \le \mathbb{E}\left[\sum_{t=1}^{T} (\varepsilon_t + r_t)\right] = \sum_{t=1}^{T} \varepsilon_t + \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]. \tag{26}$$

Here, the first equality follows from the tower property of conditional expectation, and the inequality uses Equation (19).

Since $\varepsilon_t = t^{-1/2}$,

$$\sum_{t=1}^{T} \varepsilon_t = \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \le 2\sqrt{T}. \tag{27}$$

Under Assumption 3.1. Using Equation (22),

$$r_t \le \frac{2\sqrt{T}}{\kappa} I_t.$$

Summing over $t = 1, \ldots, T$ and taking expectations gives

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \le \frac{2\sqrt{T}}{\kappa} \mathbb{E}\left[\sum_{t=1}^{T} I_t\right].$$

Applying Equation (25),

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \le \frac{2H(\mathcal{S}^*)}{\kappa} \sqrt{T}.$$

Combining this with Equation (26) and Equation (27), we conclude

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le \left(2 + \frac{2H(\mathcal{S}^*)}{\kappa}\right) \sqrt{T}.$$

Under Assumption 3.2. Using Equation (24),

$$r_t \le \frac{2}{c} t^{\beta+\frac{1}{2}} I_t \le \frac{2}{c} T^{\beta+\frac{1}{2}} I_t,$$

where the second inequality uses $t \le T$ and $\beta \ge 0$. Summing over t and taking expectations gives

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \le \frac{2T^{\beta+\frac{1}{2}}}{c} \mathbb{E}\left[\sum_{t=1}^{T} I_t\right].$$

Applying Equation (25),

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \le \frac{2H(\mathcal{S}^*)}{c} T^{\beta+\frac{1}{2}}.$$

Combining this with Equation (26) and Equation (27) yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \leq 2\sqrt{T} + \frac{2H(\mathcal{S}^*)}{c} T^{\beta+\frac{1}{2}}.$$

Finally, if $\beta < \frac{1}{2}$, then $\beta + \frac{1}{2} < 1$, so both terms on the right-hand side are $o(T)$, implying sublinear regret.

This completes the proof. □

A.7 Proof of Proposition 3.1

Proposition 3.1. *For Algorithm 1, the expected cumulative Pareto-set mismatch, defined as* $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\right]$, *satisfies* $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\right] \leq \left(2 + \frac{2H(\mathcal{S}^*)}{\kappa}\right)\sqrt{T}$ *under Assumption 3.1, and* $\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\right] \leq 2\sqrt{T} + \frac{2H(\mathcal{S}^*)}{c} T^{\beta+\frac{1}{2}}$ *under Assumption 3.2.*

Proof. Recall that under Algorithm 1, the sampled set S_t is defined by

$$S_t = \begin{cases} \mathcal{S}^*(\boldsymbol{\theta}_t), & \text{if } B_t = 1, \\ \mathcal{A}, & \text{if } B_t = 0, \end{cases}$$

where $B_t \sim \text{Bernoulli}(1 - \varepsilon_t)$ and, on the branch $B_t = 1$, $\boldsymbol{\theta}_t \sim p_t(\boldsymbol{\theta})$.

Define the instantaneous Pareto-set mismatch loss

$$m_t := \mathbb{1}\{S_t \neq \mathcal{S}^*\}.$$

We first bound its conditional expectation. At $t = 1$, $\varepsilon_1 = 1$, so $B_1 = 0$ almost surely and therefore

$$\mathbb{E}[m_1 \mid \mathcal{H}_1] \leq 1 = \varepsilon_1.$$

Thus the bound derived below already holds at $t = 1$. For the branch-conditioning argument, we may therefore assume $t \geq 2$, for which $1 - \varepsilon_t > 0$.

If $B_t = 0$, then $S_t = \mathcal{A}$, and hence trivially

$$m_t \leq 1.$$

If $B_t = 1$, then

$$S_t = \mathcal{S}^*(\boldsymbol{\theta}_t).$$

Conditional on $\mathcal{H}_t$, the branch variable B_t is generated using fresh algorithmic randomness independent of the true parameter $\boldsymbol{\theta}$. Hence conditioning additionally on $B_t = 1$ does not change the posterior distribution of $\boldsymbol{\theta}$, so

$$\boldsymbol{\theta} \mid (\mathcal{H}_t, B_t = 1) \sim p_t(\boldsymbol{\theta}).$$

On this branch, the Thompson sample $\boldsymbol{\theta}_t \sim p_t(\boldsymbol{\theta})$ is generated using fresh randomness independent of $\boldsymbol{\theta}$ conditional on $(\mathcal{H}_t, B_t = 1)$. Therefore, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_t$ are conditionally independent and identically distributed according to $p_t(\boldsymbol{\theta})$. Consequently,

$$\mathcal{S}^* = \mathcal{S}^*(\boldsymbol{\theta}) \quad \text{and} \quad S_t = \mathcal{S}^*(\boldsymbol{\theta}_t)$$

are conditionally independent and identically distributed according to the induced posterior

$$p_t(s) = \mathbb{P}(\mathcal{S}^* = s \mid \mathcal{H}_t).$$

Hence, for $t \geq 2$,

$$\mathbb{P}(S_t \neq \mathcal{S}^* \mid \mathcal{H}_t, B_t = 1) = \mathbb{P}(S \neq S' \mid \mathcal{H}_t) = G_t,$$

where $S, S' \overset{\text{i.i.d.}}{\sim} p_t$.

Moreover,

$$m_t = \mathbb{1}\{S_t \neq \mathcal{S}^*\} \leq \mathbb{1}\{B_t = 0\} + \mathbb{1}\{B_t = 1,\ S_t \neq \mathcal{S}^*\}.$$

Taking conditional expectation given $\mathcal{H}_t$, for $t \geq 2$, gives

$$\mathbb{E}[m_t \mid \mathcal{H}_t] \leq \varepsilon_t + (1-\varepsilon_t)\mathbb{P}(S_t \neq \mathcal{S}^* \mid \mathcal{H}_t, B_t = 1).$$

Substituting the preceding identity, and combining with the already verified $t = 1$ case, yields for every t,

$$\mathbb{E}[m_t \mid \mathcal{H}_t] \leq \varepsilon_t + (1-\varepsilon_t)G_t.$$

Define

$$r_t := (1-\varepsilon_t)G_t.$$

Then

$$\mathbb{E}[m_t \mid \mathcal{H}_t] \leq \varepsilon_t + r_t, \qquad \text{with} \qquad r_t \leq G_t. \tag{28}$$

We next relate r_t to information gain. Fix versions of the relevant regular conditional distributions as probability kernels in the history. For every realizable history h_t, define

$$I_t(h_t) := I(\mathcal{S}^*; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t = h_t, A_t),$$

and write $I_t := I_t(\mathcal{H}_t)$. Let

$$\pi_t(a) := \mathbb{P}(A_t = a \mid \mathcal{H}_t), \qquad a \in \mathcal{A}.$$

Since Algorithm 1 explores uniformly over $\mathcal{A}$ with probability ε_t,

$$\pi_t(a) \geq \frac{\varepsilon_t}{|\mathcal{A}|} \qquad \text{for all } a \in \mathcal{A}.$$

By the definition of I_t,

$$I_t = \sum_{a \in \mathcal{A}} \pi_t(a)\, I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a).$$

Applying Lemma 3.1 to each action a,

$$I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \geq \frac{1}{2}\mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right].$$

Therefore,

$$I_t \geq \frac{1}{2}\sum_{a \in \mathcal{A}} \pi_t(a)\, \mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right].$$

Using the lower bound on $\pi_t(a)$,

$$I_t \geq \frac{\varepsilon_t}{2|\mathcal{A}|}\sum_{a \in \mathcal{A}} \mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right],$$

and hence

$$I_t \geq \frac{\varepsilon_t}{2}\mathbb{E}_{S,S' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\right].$$

If $G_t = 0$, then $r_t = 0$, so all subsequent upper bounds on r_t hold trivially. We may therefore assume $G_t > 0$. Since the total variation term vanishes whenever $S = S'$, conditioning on $\{S \neq S'\}$ gives

$$I_t \geq \frac{\varepsilon_t}{2}G_t\mathbb{E}\Big[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t\Big].$$

Under Assumption 3.1,

$$\mathbb{E}\Big[\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t\Big] \geq \kappa$$

whenever $G_t > 0$. Consequently,

$$I_t \geq \frac{\varepsilon_t \kappa}{2}G_t.$$

Using $r_t \leq G_t$ and $\varepsilon_t = t^{-1/2}$, we obtain

$$r_t \leq \frac{2}{\varepsilon_t \kappa}I_t = \frac{2\sqrt{t}}{\kappa}I_t \leq \frac{2\sqrt{T}}{\kappa}I_t. \tag{29}$$

Together with the trivial $G_t = 0$ case, this inequality holds at every realizable history.

Similarly, under Assumption 3.2,

$$\mathbb{E}\Big[\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t\Big] \geq c\, t^{-\beta}$$

whenever $G_t > 0$. Thus

$$I_t \geq \frac{\varepsilon_t c}{2} t^{-\beta} G_t.$$

Using again $r_t \leq G_t$ and $\varepsilon_t = t^{-1/2}$, we obtain

$$r_t \leq \frac{2t^\beta}{\varepsilon_t c} I_t = \frac{2}{c} t^{\beta+\frac{1}{2}} I_t \leq \frac{2}{c} T^{\beta+\frac{1}{2}} I_t. \tag{30}$$

Again, together with the trivial $G_t = 0$ case, this holds at every realizable history.

We now control the expected cumulative information gain. Since $\mathcal{H}_1$ is empty and

$$\mathcal{H}_{t+1} = (\mathcal{H}_t, A_t, \mathbf{Y}_{t,A_t}),$$

the chain rule for mutual information gives

$$I(\mathcal{S}^*; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t).$$

Since A_t is selected based on $\mathcal{H}_t$ using independent internal randomization,

$$I(\mathcal{S}^*; A_t \mid \mathcal{H}_t) = 0.$$

Therefore,

$$I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) = I(\mathcal{S}^*; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t) = \mathbb{E}[I_t],$$

where the final equality follows by averaging the history-dependent quantity $I_t = I_t(\mathcal{H}_t)$ over $\mathcal{H}_t$. Consequently,

$$I(\mathcal{S}^*; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} \mathbb{E}[I_t] = \mathbb{E}\left[\sum_{t=1}^{T} I_t\right].$$

Since $\mathcal{S}^*$ is finite-valued,

$$\mathbb{E}\left[\sum_{t=1}^{T} I_t\right] = I(\mathcal{S}^*; \mathcal{H}_{T+1}) \leq H(\mathcal{S}^*). \tag{31}$$

Summing Equation (28) over $t = 1, \ldots, T$ and taking expectations gives

$$\mathbb{E}\left[\sum_{t=1}^{T} m_t\right] \leq \sum_{t=1}^{T} \varepsilon_t + \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]. \tag{32}$$

Since $\varepsilon_t = t^{-1/2}$,

$$\sum_{t=1}^{T} \varepsilon_t = \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}. \tag{33}$$

Under Assumption 3.1, using Equation (29) and Equation (31),

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \leq \frac{2\sqrt{T}}{\kappa} \mathbb{E}\left[\sum_{t=1}^{T} I_t\right] \leq \frac{2H(\mathcal{S}^*)}{\kappa} \sqrt{T}.$$

Combining this with Equation (32) and Equation (33), we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\right] \leq \left(2 + \frac{2H(\mathcal{S}^*)}{\kappa}\right) \sqrt{T}.$$

Under Assumption 3.2, using Equation (30) and Equation (31),

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] \le \frac{2T^{\beta+\frac{1}{2}}}{c}\mathbb{E}\left[\sum_{t=1}^{T} I_t\right] \le \frac{2H(\mathcal{S}^*)}{c}T^{\beta+\frac{1}{2}}.$$

Combining this with Equation (32) and Equation (33), we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{S_t \neq \mathcal{S}^*\}\right] \le 2\sqrt{T} + \frac{2H(\mathcal{S}^*)}{c}T^{\beta+\frac{1}{2}}.$$

Finally, if $\beta < \frac{1}{2}$, then $\beta + \frac{1}{2} < 1$, so both terms on the right-hand side are $o(T)$, implying sublinear cumulative Pareto-set mismatch.

This completes the proof. □

A.8 Proof of Theorem 3.3

Theorem 3.3. *Assume that, for some $\Psi > 0$, $\psi_t\big(\pi_t^{\mathrm{IDS}}\big) \le \Psi$ for all realizable $\mathcal{H}_t$ and $\forall t \in [T]$. Then, for Algorithm 2, the Bayesian cumulative Pareto-violation regret satisfies $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\Big] = \mathbb{E}\Big[\sum_{t=1}^{T} \Delta_t\big(\pi_t^{\mathrm{IDS}}\big)\Big] \le \sqrt{\Psi\, H(\mathcal{S}^*)\, T}$. Furthermore, under Assumption 3.1, $\psi_t\big(\pi_t^{\mathrm{IDS}}\big) \le \frac{8}{\kappa}$ for all $t \in [T]$ and $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\Big] \le \sqrt{\frac{8}{\kappa}H(\mathcal{S}^*)\, T}$. Under Assumption 3.2, $\psi_t\big(\pi_t^{\mathrm{IDS}}\big) \le \frac{8}{c}t^{\beta}$ for all $t \in [T]$, and $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\Big] \le \sqrt{\frac{8}{c}H(\mathcal{S}^*)\, T^{\beta+1}}$, which is sublinear whenever $\beta < 1$.*

Proof. Recall that the instantaneous Pareto-violation loss is

$$\ell_t := \mathbb{1}\{A_t \notin \mathcal{S}^*\}.$$

For Algorithm 2, the Bayesian cumulative Pareto-violation regret is

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right].$$

Fix versions of the relevant regular conditional distributions as probability kernels in the history. For every realizable history h_t and action $a \in \mathcal{A}$, define

$$\Delta_t(a; h_t) := \mathbb{P}(a \notin \mathcal{S}^* \mid \mathcal{H}_t = h_t)$$

and

$$I_t(a; h_t) := I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t = h_t, A_t = a),$$

where the latter uses the fixed-action interpretation from Section 3. Since $\mathcal{S}^* \in 2^{\mathcal{A}}$ and $\mathcal{A}$ is finite, $I_t(a; h_t)$ is finite for every a and realizable h_t. For $\pi \in \Pi(\mathcal{A})$, define

$$\Delta_t(\pi; h_t) := \sum_{a\in\mathcal{A}} \pi(a)\Delta_t(a; h_t), \qquad I_t(\pi; h_t) := \sum_{a\in\mathcal{A}} \pi(a) I_t(a; h_t).$$

For $x, y \ge 0$, define the extended quadratic-over-linear function

$$\varphi(x, y) := \begin{cases} x^2/y, & y > 0, \\ 0, & x = 0,\ y = 0, \\ +\infty, & x > 0,\ y = 0. \end{cases}$$

Thus the history-wise IDS objective can be written as

$$\psi_t(\pi; h_t) := \varphi(\Delta_t(\pi; h_t), I_t(\pi; h_t)).$$

For fixed h_t, both $\Delta_t(\pi; h_t)$ and $I_t(\pi; h_t)$ are affine, hence continuous, functions of π. Moreover, φ is lower semicontinuous on $\mathbb{R}^2_+$, so $\pi \mapsto \psi_t(\pi; h_t)$ is lower semicontinuous. Since $\mathcal{A}$ is finite, $\Pi(\mathcal{A})$

is a compact metric simplex. Therefore, for every realizable history h_t, the minimum is attained and the set

$$\Gamma_t(h_t) := \arg\min_{\pi \in \Pi(\mathcal{A})} \psi_t(\pi; h_t)$$

is nonempty and compact.

The history space is standard Borel, and, under the fixed regular conditional probability kernels above, the maps $h_t \mapsto \Delta_t(a; h_t)$ and $h_t \mapsto I_t(a; h_t)$ are measurable for every $a \in \mathcal{A}$. Since $\mathcal{A}$ is finite, it follows that $(h_t, \pi) \mapsto \psi_t(\pi; h_t)$ is jointly measurable and is lower semicontinuous in π. Hence the measurable minimum/selection theorem for a measurable lower-semicontinuous objective over a compact metric action space yields a measurable selector

$$h_t \longmapsto \pi_t^{\mathrm{IDS}}(h_t) \in \Gamma_t(h_t).$$

We fix one such selector. Consequently, $\pi_t^{\mathrm{IDS}} = \pi_t^{\mathrm{IDS}}(\mathcal{H}_t)$ is $\mathcal{H}_t$-measurable.

We suppress the explicit dependence on h_t and write $\Delta_t(a)$, $I_t(a)$, $\Delta_t(\pi)$, and $I_t(\pi)$ for the corresponding $\mathcal{H}_t$-measurable random quantities evaluated at $h_t = \mathcal{H}_t$.

Conditional on $\mathcal{H}_t$, the distribution π_t^{IDS} is fixed, and A_t is generated from it using fresh algorithmic randomness independent of the true parameter and hence of $\mathcal{S}^*$. Therefore,

$$A_t \perp\!\!\!\perp \mathcal{S}^* \mid \mathcal{H}_t.$$

Consequently,

$$\mathbb{E}[\ell_t \mid \mathcal{H}_t] = \sum_{a \in \mathcal{A}} \pi_t^{\mathrm{IDS}}(a)\, \mathbb{P}(a \notin \mathcal{S}^* \mid \mathcal{H}_t) = \Delta_t(\pi_t^{\mathrm{IDS}}).$$

Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_t(\pi_t^{\mathrm{IDS}})\right]. \tag{34}$$

Suppose first that

$$\psi_t(\pi_t^{\mathrm{IDS}}) \le \Psi$$

at every realizable history and every $t \in [T]$, for some finite $\Psi > 0$. Then, at every such history,

$$\Delta_t(\pi_t^{\mathrm{IDS}})^2 \le \Psi\, I_t(\pi_t^{\mathrm{IDS}}). \tag{35}$$

Indeed, when $I_t(\pi_t^{\mathrm{IDS}}) > 0$, this follows directly from the definition of the information ratio. When $I_t(\pi_t^{\mathrm{IDS}}) = 0$, finiteness of the ratio implies $\Delta_t(\pi_t^{\mathrm{IDS}}) = 0$, so Equation (35) also holds.

Taking square roots,

$$\Delta_t(\pi_t^{\mathrm{IDS}}) \le \sqrt{\Psi}\, \sqrt{I_t(\pi_t^{\mathrm{IDS}})}.$$

Substituting this into Equation (34) gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right] \le \sqrt{\Psi} \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{I_t(\pi_t^{\mathrm{IDS}})}\right].$$

Since $x \mapsto \sqrt{x}$ is concave, Jensen's inequality gives

$$\mathbb{E}\left[\sqrt{I_t(\pi_t^{\mathrm{IDS}})}\right] \le \sqrt{\mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]}.$$

Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right] \le \sqrt{\Psi} \sum_{t=1}^{T} \sqrt{\mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]}.$$

Applying Cauchy–Schwarz,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right] \le \sqrt{\Psi T \sum_{t=1}^{T} \mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]}. \tag{36}$$

We now control the cumulative expected information gain about $\mathcal{S}^*$. Since $\mathcal{H}_1$ is empty and

$$\mathcal{H}_{t+1} = (\mathcal{H}_t, A_t, \mathbf{Y}_{t,A_t}),$$

the chain rule for mutual information gives

$$I(\mathcal{S}^*; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t)\,.$$

Because π_t^{IDS} is $\mathcal{H}_t$-measurable and A_t is sampled from it using fresh algorithmic randomness independent of $\mathcal{S}^*$ conditional on $\mathcal{H}_t$,

$$I(\mathcal{S}^*; A_t \mid \mathcal{H}_t) = 0.$$

Therefore,

$$I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) = I(\mathcal{S}^*; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t)\,.$$

By the definition of the history-dependent action information quantities and the law of total expectation,

$$\begin{aligned} I(\mathcal{S}^*; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t) &= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \mathbb{P}(A_t = a \mid \mathcal{H}_t)\, I_t(a; \mathcal{H}_t)\right] \\ &= \mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]\,. \end{aligned}$$

Consequently,

$$\sum_{t=1}^{T} \mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big] = I(\mathcal{S}^*; \mathcal{H}_{T+1}) \le H(\mathcal{S}^*), \tag{37}$$

where the inequality follows because $\mathcal{S}^*$ is finite-valued and mutual information is upper bounded by entropy.

Substituting Equation (37) into Equation (36), we obtain the generic bound

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le \sqrt{\Psi\, H(\mathcal{S}^*)\, T}. \tag{38}$$

We now specialize this bound under Assumptions 3.1 and 3.2.

Uniform separation. Fix t and a realizable history $\mathcal{H}_t$, and write

$$G_t := \mathbb{P}(S \neq S' \mid \mathcal{H}_t), \qquad S, S' \overset{\text{i.i.d.}}{\sim} p_t.$$

If $G_t = 0$, then

$$G_t = 1 - \sum_s p_t(s)^2 = 0,$$

so the posterior over $\mathcal{S}^*$ is concentrated on a single Pareto set, say s_0, with $p_t(s_0) = 1$. Since s_0 is a realizable Pareto set, it is nonempty. Choose any $a_0 \in s_0$. Then

$$\Delta_t(a_0) = \mathbb{P}(a_0 \notin \mathcal{S}^* \mid \mathcal{H}_t) = 0.$$

Moreover, $\mathcal{S}^* = s_0$ almost surely conditional on $\mathcal{H}_t$, and hence

$$I_t(a_0) = I(\mathcal{S}^*; \mathbf{Y}_{t,a_0} \mid \mathcal{H}_t, A_t = a_0) = 0.$$

Thus, under the extended-value convention,

$$\psi_t(\delta_{a_0}) = 0.$$

Since information ratios are nonnegative and π_t^{IDS} minimizes the information ratio,

$$0 \le \psi_t(\pi_t^{\mathrm{IDS}}) \le \psi_t(\delta_{a_0}) = 0.$$

Therefore,

$$\psi_t(\pi_t^{\mathrm{IDS}}) = 0 \le \frac{8}{\kappa}.$$

It remains only to consider the case $G_t > 0$.

For each $a \in \mathcal{A}$, define

$$g_t(a) := \mathbb{E}\Big[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t\Big].$$

By Assumption 3.1,

$$\mathbb{E}_{a\sim\mathrm{Unif}(\mathcal{A})}[g_t(a)] \geq \kappa.$$

Choose, using a fixed deterministic tie-breaking rule,

$$a_t^{\inf} \in \arg\max_{a\in\mathcal{A}} g_t(a).$$

Then

$$g_t(a_t^{\inf}) \geq \kappa.$$

By Lemma 3.1,

$$I_t(a) \geq \frac{1}{2}\mathbb{E}_{S,S'\overset{\text{i.i.d.}}{\sim}p_t}\Big[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2\Big].$$

Since the total variation term vanishes when $S = S'$, and $G_t > 0$,

$$I_t(a) \geq \frac{1}{2}G_t\mathbb{E}\Big[\mathrm{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t\Big] = \frac{1}{2}G_t g_t(a).$$

Hence

$$I_t(a_t^{\inf}) \geq \frac{\kappa}{2}G_t. \tag{39}$$

Let

$$s_{\max} \in \arg\max_s p_t(s), \qquad p_{\max} := p_t(s_{\max}),$$

using a fixed deterministic tie-breaking rule if the maximizer is not unique. Since $p_{\max} > 0$, $s_{\max}$ is a posterior-supported Pareto set and hence is nonempty. Choose $a_t^{\exp} \in s_{\max}$ using a fixed deterministic rule. Since $a_t^{\exp}$ is Pareto-optimal whenever $\mathcal{S}^* = s_{\max}$,

$$\mathbb{P}(a_t^{\exp} \in \mathcal{S}^* \mid \mathcal{H}_t) \geq p_{\max}.$$

Therefore,

$$\Delta_t(a_t^{\exp}) \leq 1 - p_{\max}.$$

Moreover, since $p_t(s) \leq p_{\max}$ for every s,

$$\sum_s p_t(s)^2 \leq p_{\max}\sum_s p_t(s) = p_{\max}.$$

Thus

$$G_t = 1 - \sum_s p_t(s)^2 \geq 1 - p_{\max},$$

and hence

$$\Delta_t(a_t^{\exp}) \leq G_t. \tag{40}$$

Since G_t is a probability and we are in the case $G_t > 0$,

$$0 < G_t \leq 1.$$

Therefore the comparison distribution

$$\pi_t^{\mathrm{mix}} := (1 - G_t)\delta_{a_t^{\exp}} + G_t\delta_{a_t^{\inf}}$$

is a valid probability distribution on $\mathcal{A}$, where δ_a denotes the point mass on action a. Since every Pareto-violation loss is at most one,

$$\Delta_t(a_t^{\inf}) \leq 1.$$

Using Equation (40),

$$\begin{aligned}\Delta_t(\pi_t^{\mathrm{mix}}) &= (1 - G_t)\Delta_t(a_t^{\exp}) + G_t\Delta_t(a_t^{\inf}) \\ &\leq (1 - G_t)G_t + G_t \\ &= 2G_t - G_t^2 \\ &\leq 2G_t.\end{aligned}$$

Its information gain satisfies, using Equation (39),

$$\begin{aligned} I_t(\pi_t^{\mathrm{mix}}) &= (1-G_t)I_t(a_t^{\mathrm{exp}}) + G_t I_t(a_t^{\mathrm{inf}}) \\ &\geq G_t I_t(a_t^{\mathrm{inf}}) \\ &\geq \frac{\kappa}{2} G_t^2. \end{aligned}$$

Since $G_t > 0$ and $\kappa > 0$, the denominator above is strictly positive. Therefore,

$$\psi_t(\pi_t^{\mathrm{mix}}) \leq \frac{(2G_t)^2}{(\kappa/2)G_t^2} = \frac{8}{\kappa}.$$

Because π_t^{IDS} minimizes the information ratio,

$$\psi_t(\pi_t^{\mathrm{IDS}}) \leq \psi_t(\pi_t^{\mathrm{mix}}) \leq \frac{8}{\kappa}.$$

Together with the already treated $G_t = 0$ case, this holds at every realizable history. Taking

$$\Psi = \frac{8}{\kappa}$$

in Equation (38) yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \leq \sqrt{\frac{8}{\kappa} H(\mathcal{S}^*) T}.$$

Decaying separation. Fix t and a realizable history $\mathcal{H}_t$. If $G_t = 0$, the same argument as above yields

$$\psi_t(\pi_t^{\mathrm{IDS}}) = 0 \leq \frac{8}{c} t^{\beta}.$$

Suppose therefore that $G_t > 0$, and define

$$g_t(a) := \mathbb{E}\Big[\mathrm{TV}\big(P_{t,S}^a, P_{t,S'}^a\big)^2 \;\Big|\; S \neq S', \mathcal{H}_t\Big].$$

By Assumption 3.2,

$$\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}[g_t(a)] \geq c\, t^{-\beta}.$$

Choose, using a fixed deterministic tie-breaking rule,

$$a_t^{\mathrm{inf}} \in \arg\max_{a \in \mathcal{A}} g_t(a).$$

Then

$$g_t(a_t^{\mathrm{inf}}) \geq c\, t^{-\beta}.$$

By Lemma 3.1 and the same conditioning argument,

$$I_t(a_t^{\mathrm{inf}}) \geq \frac{c}{2} t^{-\beta} G_t.$$

Let

$$s_{\max} \in \arg\max_{s} p_t(s),$$

using the same fixed deterministic tie-breaking rule. Since $p_t(s_{\max}) > 0$, $s_{\max}$ is a posterior-supported Pareto set and hence is nonempty. Choose $a_t^{\mathrm{exp}} \in s_{\max}$ using the same fixed deterministic rule. As proved above,

$$\Delta_t(a_t^{\mathrm{exp}}) \leq G_t.$$

Since $0 < G_t \leq 1$, define the valid comparison distribution

$$\pi_t^{\mathrm{mix}} := (1-G_t)\delta_{a_t^{\mathrm{exp}}} + G_t \delta_{a_t^{\mathrm{inf}}}.$$

Then

$$\Delta_t(\pi_t^{\mathrm{mix}}) \leq (1-G_t)G_t + G_t \leq 2G_t,$$

while

$$I_t(\pi_t^{\mathrm{mix}}) \geq G_t I_t(a_t^{\mathrm{inf}}) \geq \frac{c}{2} t^{-\beta} G_t^2.$$

Since $G_t > 0$ and $c > 0$, the denominator is strictly positive. Therefore,

$$\psi_t(\pi_t^{\mathrm{mix}}) \leq \frac{(2G_t)^2}{(c/2)t^{-\beta}G_t^2} = \frac{8}{c} t^{\beta}.$$

Since IDS minimizes the information ratio,

$$\psi_t(\pi_t^{\mathrm{IDS}}) \leq \psi_t(\pi_t^{\mathrm{mix}}) \leq \frac{8}{c} t^{\beta} \qquad \text{for every } t \in [T].$$

Consequently, since $\beta \geq 0$ and $t \leq T$,

$$\psi_t(\pi_t^{\mathrm{IDS}}) \leq \frac{8}{c} T^{\beta} \qquad \text{for every } t \in [T].$$

Taking

$$\Psi = \frac{8}{c} T^{\beta}$$

in Equation (38) gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \leq \sqrt{\frac{8}{c} H(\mathcal{S}^*) T^{\beta+1}}.$$

Finally, if $\beta < 1$, then

$$\frac{\beta+1}{2} < 1,$$

so the bound is $o(T)$, implying sublinear Pareto-violation regret.

This completes the proof. □

A.9 Proof of Proposition 3.2

Proposition 3.2. *For every $\kappa \in (0, 1/8]$, there exists a two-armed, two-objective Bayesian bandit instance for which, at every finite realizable history $\mathcal{H}_t$ with $G_t > 0$, the average squared-TV separation satisfies $\kappa_t(\mathcal{H}_t) = \kappa$. Furthermore, for every horizon $T \geq 1$, every possibly randomized, adaptive algorithm satisfies $\mathbb{E}\Big[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*(\boldsymbol{\theta})\}\Big] \geq \frac{1}{32} \min\{T, \frac{1}{\kappa}\}$.*

Proof. We construct a two-armed, two-objective Bayesian bandit for which the average squared-TV separation is exactly κ at every finite realizable history, and then prove that every possibly randomized and adaptive algorithm incurs Bayesian cumulative Pareto-violation regret at least $\frac{1}{32} \min\{T, 1/\kappa\}$.

Fix any $\kappa \in (0, 1/8]$ and set

$$\delta := \sqrt{\frac{\kappa}{2}}.$$

Then $\delta \in (0, 1/4]$ and $\kappa = 2\delta^2$. Let

$$\mathcal{A} = \{a_0, a_1\}, \qquad \Theta = \{\theta_0, \theta_1\},$$

with the uniform prior

$$\mathbb{P}(\boldsymbol{\theta} = \theta_0) = \mathbb{P}(\boldsymbol{\theta} = \theta_1) = \frac{1}{2}.$$

Step 1: Construction and Pareto sets. Under both environments, action a_0 yields the deterministic reward

$$\mathbf{Y}_{t,a_0} \equiv (1/2, 1/2).$$

Under θ_0, action a_1 yields

$$\mathbf{Y}_{t,a_1} = (Z_t, Z_t), \qquad Z_t \sim \mathrm{Bernoulli}(1/2 - \delta),$$

whereas under θ_1 it yields

$$\mathbf{Y}_{t,a_1} = (Z_t, Z_t), \qquad Z_t \sim \text{Bernoulli}(1/2 + \delta).$$

Observations are independent across rounds conditional on the environment.

Under θ_0,

$$\boldsymbol{\mu}_{\theta_0}(a_0) = (1/2, 1/2), \qquad \boldsymbol{\mu}_{\theta_0}(a_1) = (1/2 - \delta, 1/2 - \delta),$$

so a_0 strictly dominates a_1 and

$$\mathcal{S}^*(\theta_0) = \{a_0\}.$$

Similarly, under θ_1,

$$\boldsymbol{\mu}_{\theta_1}(a_0) = (1/2, 1/2), \qquad \boldsymbol{\mu}_{\theta_1}(a_1) = (1/2 + \delta, 1/2 + \delta),$$

so a_1 strictly dominates a_0 and

$$\mathcal{S}^*(\theta_1) = \{a_1\}.$$

Write

$$s_0 := \{a_0\}, \qquad s_1 := \{a_1\}.$$

Thus $\mathcal{S}^* = s_i$ if and only if $\boldsymbol{\theta} = \theta_i$.

Because

$$0 < 1/2 - \delta < 1/2 + \delta < 1,$$

each outcome in $\{(0,0),(1,1)\}$ from action a_1 has strictly positive probability under both environments, while a_0 has the same deterministic law under both. Since the algorithm uses the same policy kernel at any common history, it follows inductively that every finite history realizable under one environment is also realizable under the other. Hence both environments retain strictly positive posterior probability after every finite realizable history. In particular,

$$G_t = \mathbb{P}(S \neq S' \mid \mathcal{H}_t) > 0$$

at every finite realizable history.

Step 2: The separation is exactly κ. Since each Pareto-set candidate uniquely identifies the environment and future rewards are conditionally independent of the past given the environment, for every finite realizable history $\mathcal{H}_t$,

$$P^{a_0}_{t,s_0} = P^{a_0}_{t,s_1},$$

whereas

$$P^{a_1}_{t,s_0} = \mathcal{L}((Z,Z)), \quad Z \sim \text{Bernoulli}(1/2 - \delta),$$

and

$$P^{a_1}_{t,s_1} = \mathcal{L}((Z,Z)), \quad Z \sim \text{Bernoulli}(1/2 + \delta).$$

Fix such a history and let $S, S' \overset{\text{i.i.d.}}{\sim} p_t$. Conditional on $\{S \neq S'\}$, the ordered pair (S, S') is either (s_0, s_1) or (s_1, s_0). Since total variation is symmetric,

$$\begin{aligned}\kappa_t(\mathcal{H}_t) &= \mathbb{E}\Big[\mathbb{E}_{a\sim\text{Unif}(\mathcal{A})}\,\text{TV}\big(P^a_{t,S}, P^a_{t,S'}\big)^2 \,\Big|\, S \neq S', \mathcal{H}_t\Big] \\ &= \frac{1}{2}\,\text{TV}\big(P^{a_0}_{t,s_0}, P^{a_0}_{t,s_1}\big)^2 + \frac{1}{2}\,\text{TV}\big(P^{a_1}_{t,s_0}, P^{a_1}_{t,s_1}\big)^2.\end{aligned}$$

The first term is zero. For a_1, the reward takes values only in $\{(0,0),(1,1)\}$, and therefore

$$\begin{aligned}\text{TV}\big(P^{a_1}_{t,s_0}, P^{a_1}_{t,s_1}\big) &= \frac{1}{2}\Big(|(1/2 - \delta) - (1/2 + \delta)| + |(1/2 + \delta) - (1/2 - \delta)|\Big) \\ &= 2\delta.\end{aligned}$$

Consequently,

$$\kappa_t(\mathcal{H}_t) = \frac{1}{2}(2\delta)^2 = 2\delta^2 = \kappa.$$

Thus the first claim of the proposition holds at every finite realizable history.

Step 3: Express the Bayesian regret through the number of pulls of a_1. Fix an arbitrary horizon $T \geq 1$ and an arbitrary possibly randomized and adaptive algorithm. For $i \in \{0, 1\}$, let $\mathbb{P}_i$ and $\mathbb{E}_i$ denote probability and expectation under environment θ_i, and define

$$N_1(T) := \sum_{t=1}^{T} \mathbb{1}\{A_t = a_1\}.$$

Under θ_0, only a_0 is Pareto optimal, so the cumulative violation loss is $N_1(T)$. Under θ_1, only a_1 is Pareto optimal, so the cumulative violation loss is $T - N_1(T)$. Hence, under the uniform prior,

$$\begin{aligned} R_T &:= \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*(\boldsymbol{\theta})\}\right] \\ &= \frac{1}{2}\mathbb{E}_0[N_1(T)] + \frac{1}{2}\big(T - \mathbb{E}_1[N_1(T)]\big) \\ &= \frac{1}{2}\Big(T - \big(\mathbb{E}_1[N_1(T)] - \mathbb{E}_0[N_1(T)]\big)\Big). \end{aligned} \tag{41}$$

Step 4: Relate the difference in pull counts to trajectory TV. Let $\mathbb{P}_i^{\mathrm{traj}}$ be the law under θ_i of

$$\mathcal{T}_{T+1} := (A_1, \mathbf{Y}_{1,A_1}, \ldots, A_T, \mathbf{Y}_{T,A_T}).$$

Since

$$f(\mathcal{T}_{T+1}) := \frac{N_1(T)}{T}$$

takes values in $[0, 1]$, the variational characterization of total variation gives

$$|\mathbb{E}_1[f] - \mathbb{E}_0[f]| \leq \mathrm{TV}\Big(\mathbb{P}_0^{\mathrm{traj}}, \mathbb{P}_1^{\mathrm{traj}}\Big).$$

Therefore,

$$|\mathbb{E}_1[N_1(T)] - \mathbb{E}_0[N_1(T)]| \leq T\ \mathrm{TV}\Big(\mathbb{P}_0^{\mathrm{traj}}, \mathbb{P}_1^{\mathrm{traj}}\Big). \tag{42}$$

Step 5: Upper bound the trajectory KL divergence. By the chain rule for KL divergence,

$$\begin{aligned} &\mathrm{KL}\Big(\mathbb{P}_0^{\mathrm{traj}} \,\|\, \mathbb{P}_1^{\mathrm{traj}}\Big) \\ &\quad = \sum_{t=1}^{T} \mathbb{E}_0\big[\mathrm{KL}(\mathbb{P}_0(A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) \,\|\, \mathbb{P}_1(A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t))\big]. \end{aligned}$$

At every history in the common support of the two environments, the algorithm uses the same conditional action kernel; write it as $\pi_t(\cdot \mid \mathcal{H}_t)$. Thus

$$\mathbb{P}_0(A_t = a \mid \mathcal{H}_t) = \mathbb{P}_1(A_t = a \mid \mathcal{H}_t) = \pi_t(a \mid \mathcal{H}_t).$$

Consequently, the conditional chain rule for KL gives

$$\begin{aligned} &\mathrm{KL}(\mathbb{P}_0(A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) \,\|\, \mathbb{P}_1(A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t)) \\ &\quad = \sum_{a \in \mathcal{A}} \pi_t(a \mid \mathcal{H}_t)\, \mathrm{KL}(\mathbb{P}_0(\mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \,\|\, \mathbb{P}_1(\mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)). \end{aligned}$$

If $a = a_0$, the observation laws are identical, so the corresponding KL term is zero. If $a = a_1$, conditional independence across rounds implies that the KL term is

$$\begin{aligned} D_\delta &:= \left(\frac{1}{2} - \delta\right) \log \frac{1/2 - \delta}{1/2 + \delta} + \left(\frac{1}{2} + \delta\right) \log \frac{1/2 + \delta}{1/2 - \delta} \\ &= 2\delta \log\left(\frac{1 + 2\delta}{1 - 2\delta}\right). \end{aligned}$$

Therefore,

$$\mathrm{KL}\Big(\mathbb{P}_0^{\mathrm{traj}} \,\|\, \mathbb{P}_1^{\mathrm{traj}}\Big) = D_\delta\, \mathbb{E}_0[N_1(T)].$$

For $x \in [0, 1)$,

$$\log\left(\frac{1+x}{1-x}\right) = \log(1+x) - \log(1-x) \le x + \frac{x}{1-x} \le \frac{2x}{1-x}.$$

Since $2\delta \le 1/2$, taking $x = 2\delta$ gives

$$D_\delta \le \frac{8\delta^2}{1-2\delta} \le 16\delta^2.$$

Therefore,

$$\mathrm{KL}\Big(\mathbb{P}_0^{\mathrm{traj}} \,\|\, \mathbb{P}_1^{\mathrm{traj}}\Big) \le 16\delta^2\, \mathbb{E}_0[N_1(T)]. \tag{43}$$

From Equation (41),

$$R_T = \frac{1}{2}\mathbb{E}_0[N_1(T)] + \frac{1}{2}\big(T - \mathbb{E}_1[N_1(T)]\big) \ge \frac{1}{2}\mathbb{E}_0[N_1(T)],$$

because $N_1(T) \le T$ almost surely. Hence

$$\mathbb{E}_0[N_1(T)] \le 2R_T,$$

and therefore

$$\mathrm{KL}\Big(\mathbb{P}_0^{\mathrm{traj}} \,\|\, \mathbb{P}_1^{\mathrm{traj}}\Big) \le 32\delta^2 R_T. \tag{44}$$

Step 6: Conclude by a two-case argument. If $R_T \ge T/4$, then

$$R_T \ge \frac{T}{4} \ge \frac{1}{32}\min\left\{T, \frac{1}{\kappa}\right\},$$

so the desired bound holds.

Now suppose $R_T < T/4$. By Equation (41),

$$\mathbb{E}_1[N_1(T)] - \mathbb{E}_0[N_1(T)] = T - 2R_T > \frac{T}{2}.$$

Combining this with Equation (42) gives

$$\mathrm{TV}\Big(\mathbb{P}_0^{\mathrm{traj}}, \mathbb{P}_1^{\mathrm{traj}}\Big) > \frac{1}{2}.$$

Pinsker's inequality, with natural logarithms, implies

$$\mathrm{KL}\Big(\mathbb{P}_0^{\mathrm{traj}} \,\|\, \mathbb{P}_1^{\mathrm{traj}}\Big) \ge 2\ \mathrm{TV}\Big(\mathbb{P}_0^{\mathrm{traj}}, \mathbb{P}_1^{\mathrm{traj}}\Big)^2 > \frac{1}{2}.$$

Together with Equation (44), this yields

$$32\delta^2 R_T > \frac{1}{2},$$

and hence

$$R_T > \frac{1}{64\delta^2} = \frac{1}{32\kappa}.$$

Since

$$\min\left\{T, \frac{1}{\kappa}\right\} \le \frac{1}{\kappa},$$

we obtain

$$R_T > \frac{1}{32\kappa} \ge \frac{1}{32}\min\left\{T, \frac{1}{\kappa}\right\}.$$

Thus the desired lower bound also holds in the second case.

Combining the two cases, for every $T \ge 1$ and every possibly randomized and adaptive algorithm,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*(\boldsymbol{\theta})\}\right] \ge \frac{1}{32}\min\left\{T, \frac{1}{\kappa}\right\}.$$

Together with $\kappa_t(\mathcal{H}_t) = \kappa$ at every finite realizable history, this proves Proposition 3.2. □

A.10 Refined Thompson Sampling Guarantees for Pareto-Violation Regret with Separation-Dependent Exploration Schedules

The regret bound in Theorem 3.2 uses the generic, separation-agnostic schedule $\varepsilon_t = t^{-1/2}$. We now show that, when the separation profile is known, the same sampling rule can be equipped with a separation-dependent exploration schedule to obtain a sharper guarantee. Under polynomially decaying separation $\kappa_t(\mathcal{H}_t) \gtrsim t^{-\beta}$, a schedule matched to β yields sublinear regret for every $\beta < 1$. Uniform separation is recovered as the special case $\beta = 0$. Thus, the $\beta < 1/2$ condition in Theorem 3.2 is a consequence of the generic exploration schedule rather than an intrinsic limitation of Thompson Sampling.

Proposition A.2. *Suppose Assumption 3.2 holds with constants $c > 0$ and $\beta \in [0, 1)$. Consider the sampling rule of Algorithm 1, but replace its default exploration schedule by*

$$\varepsilon_t := \min\Big\{\alpha\, t^{-\frac{1-\beta}{2}},\, 1\Big\}, \qquad \alpha := \sqrt{\frac{H(\mathcal{S}^*)}{c}}.$$

Then

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le 4\sqrt{\frac{H(\mathcal{S}^*)}{c}}\, T^{\frac{1+\beta}{2}}.$$

Hence the regret is sublinear for every $\beta < 1$.

In particular, under Assumption 3.1, taking $c = \kappa$ and $\beta = 0$ gives

$$\varepsilon_t = \min\left\{\frac{\alpha}{\sqrt{t}},\, 1\right\}, \qquad \alpha = \sqrt{\frac{H(\mathcal{S}^*)}{\kappa}},$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le 4\sqrt{\frac{T\, H(\mathcal{S}^*)}{\kappa}}.$$

Therefore, for κ in the lower-bound regime of Proposition 3.2, when $T \asymp 1/\kappa$, the upper bound has the same $1/\kappa$ dependence as the lower bound, up to universal constants and the $\sqrt{H(\mathcal{S}^*)}$ factor.

Proof. Let

$$\ell_t := \mathbb{1}\{A_t \notin \mathcal{S}^*\}, \qquad G_t := \mathbb{P}(S \neq S' \mid \mathcal{H}_t), \qquad r_t := (1 - \varepsilon_t)G_t,$$

where $S, S' \overset{\text{i.i.d.}}{\sim} p_t$ are independent posterior draws of $\mathcal{S}^*$.

We first treat the degenerate case $H(\mathcal{S}^*) = 0$. Then $\mathcal{S}^*$ is almost surely constant under the prior and consequently under every posterior arising from a realizable history. Since $\alpha = 0$, we have $\varepsilon_t = 0$ for every t. The Thompson Sampling branch therefore samples a parameter whose Pareto set equals the true deterministic Pareto set almost surely and selects an action from that set. Hence $\ell_t = 0$ almost surely for every t, and the claimed bound is immediate. Henceforth assume $H(\mathcal{S}^*) > 0$, so $\alpha > 0$.

Conditional on $\mathcal{H}_t$, the uniform-exploration branch incurs expected loss at most one. On the Thompson Sampling branch, the sampled Pareto set and the true Pareto set are independent draws from the same posterior distribution. If they agree, every action selected from the sampled set is Pareto optimal under the true environment; if they disagree, the violation loss is at most one. Hence the TS-branch violation probability is at most G_t, and

$$\mathbb{E}[\ell_t \mid \mathcal{H}_t] \le \varepsilon_t + (1 - \varepsilon_t)G_t = \varepsilon_t + r_t. \tag{45}$$

For each $a \in \mathcal{A}$, define

$$I_t(a) := I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a),$$

using the fixed-action interpretation, and let

$$I_t := \sum_{a \in \mathcal{A}} \mathbb{P}(A_t = a \mid \mathcal{H}_t)\, I_t(a)$$

denote the history-wise one-step information gain under the actual sampling distribution. Because the exploration branch chooses uniformly from $\mathcal{A}$,

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t) \geq \frac{\varepsilon_t}{|\mathcal{A}|} \qquad \text{for every } a \in \mathcal{A}.$$

Therefore,

$$I_t \geq \varepsilon_t \, \mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}[I_t(a)].$$

By Lemma 3.1 and the definition of $\kappa_t(\mathcal{H}_t)$,

$$\mathbb{E}_{a \sim \mathrm{Unif}(\mathcal{A})}[I_t(a)] \geq \frac{1}{2} G_t \kappa_t(\mathcal{H}_t).$$

Hence, under Assumption 3.2, whenever $G_t > 0$,

$$I_t \geq \frac{\varepsilon_t c}{2} t^{-\beta} G_t. \tag{46}$$

If $G_t = 0$, then $r_t = 0$. If $G_t > 0$ and $\varepsilon_t < 1$, then $r_t = (1 - \varepsilon_t) G_t \leq G_t$, and Equation (46) gives

$$r_t \leq \frac{2t^\beta}{c \, \varepsilon_t} I_t. \tag{47}$$

When $\varepsilon_t = 1$, we instead have $r_t = 0$ directly.

We will also use the total information budget

$$\mathbb{E}\left[\sum_{t=1}^{T} I_t\right] = I(\mathcal{S}^*; \mathcal{H}_{T+1}) \leq H(\mathcal{S}^*). \tag{48}$$

Indeed, the mutual-information chain rule gives

$$I(\mathcal{S}^*; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t).$$

Conditional on $\mathcal{H}_t$, all posterior samples and branch/action randomization used by the algorithm are generated using fresh internal randomness independent of the true parameter and therefore of $\mathcal{S}^*$. Hence

$$I(\mathcal{S}^*; A_t \mid \mathcal{H}_t) = 0.$$

The conditional chain rule and the definition of I_t then imply

$$I(\mathcal{S}^*; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) = \mathbb{E}[I_t].$$

Summing over t yields Equation (48).

Taking expectations in Equation (45) and summing gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_t\right] \leq \sum_{t=1}^{T} \varepsilon_t + \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]. \tag{49}$$

Set

$$\gamma := \frac{1 - \beta}{2}.$$

Since $\beta \in [0, 1)$, we have $\gamma \in (0, 1/2]$ and

$$1 - \gamma = \beta + \gamma = \frac{1 + \beta}{2}. \tag{50}$$

The schedule is

$$\varepsilon_t = \min\{\alpha t^{-\gamma}, 1\}.$$

We first bound the exploration term. Since $\varepsilon_t \leq \alpha t^{-\gamma}$,

$$\sum_{t=1}^{T} \varepsilon_t \leq \alpha \sum_{t=1}^{T} t^{-\gamma}.$$

For $\gamma \in (0,1)$,

$$\sum_{t=1}^{T} t^{-\gamma} \le 1 + \int_1^T x^{-\gamma}\, dx = 1 + \frac{T^{1-\gamma}-1}{1-\gamma} \le \frac{T^{1-\gamma}}{1-\gamma}.$$

Since

$$\frac{1}{1-\gamma} = \frac{2}{1+\beta} \le 2,$$

we obtain

$$\sum_{t=1}^{T} \varepsilon_t \le 2\alpha T^{\frac{1+\beta}{2}} = 2\sqrt{\frac{H(\mathcal{S}^*)}{c}}\, T^{\frac{1+\beta}{2}}. \tag{51}$$

For the second term in Equation (49), define

$$\mathcal{U} := \{t \in [T] : \varepsilon_t < 1\}.$$

If $t \notin \mathcal{U}$, then $\varepsilon_t = 1$ and hence $r_t = 0$. Therefore,

$$\sum_{t=1}^{T} r_t = \sum_{t \in \mathcal{U}} r_t.$$

For every $t \in \mathcal{U}$, $\varepsilon_t = \alpha t^{-\gamma}$, so Equation (47) gives

$$r_t \le \frac{2}{\alpha c} t^{\beta+\gamma} I_t.$$

Using Equation (50) and $t \le T$,

$$r_t \le \frac{2}{\alpha c} T^{\frac{1+\beta}{2}} I_t, \qquad t \in \mathcal{U}.$$

Therefore, by Equation (48),

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} r_t\right] &= \mathbb{E}\left[\sum_{t \in \mathcal{U}} r_t\right] \\
&\le \frac{2}{\alpha c} T^{\frac{1+\beta}{2}} \mathbb{E}\left[\sum_{t \in \mathcal{U}} I_t\right] \\
&\le \frac{2H(\mathcal{S}^*)}{\alpha c} T^{\frac{1+\beta}{2}} \\
&= 2\sqrt{\frac{H(\mathcal{S}^*)}{c}}\, T^{\frac{1+\beta}{2}}.
\end{aligned} \tag{52}$$

Combining Equation (49), Equation (51), and Equation (52) yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le 4\sqrt{\frac{H(\mathcal{S}^*)}{c}}\, T^{\frac{1+\beta}{2}}.$$

Since $\beta < 1$, the exponent $(1+\beta)/2$ is strictly smaller than one, so the regret is sublinear.

Finally, under Assumption 3.1, Assumption 3.2 holds with $c = \kappa$ and $\beta = 0$. Thus

$$\varepsilon_t = \min\left\{\frac{\alpha}{\sqrt{t}}, 1\right\}, \qquad \alpha = \sqrt{\frac{H(\mathcal{S}^*)}{\kappa}},$$

and

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t \notin \mathcal{S}^*\}\right] \le 4\sqrt{\frac{T\,H(\mathcal{S}^*)}{\kappa}}.$$

If $T \asymp 1/\kappa$, then

$$\sqrt{\frac{T\,H(\mathcal{S}^*)}{\kappa}} \asymp \frac{\sqrt{H(\mathcal{S}^*)}}{\kappa}.$$

Hence the tuned Thompson Sampling upper bound has the same $1/\kappa$ dependence as the lower bound in Proposition 3.2, up to universal constants and the $\sqrt{H(\mathcal{S}^*)}$ factor. □

A.11 Equivalent Form of the Pareto Gap for Gap-Based Regret

Proposition A.3 (Equivalent form of the Pareto gap)**.** *Let $\mathcal{A}$ be finite. For any action $a \in \mathcal{A}$, define*

$$\Delta_{\boldsymbol{\theta}}(a) := \sup\left\{\varepsilon \geq 0 : \exists b \in \mathcal{A} \text{ such that } \boldsymbol{\mu}_{\boldsymbol{\theta}}(b) \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \varepsilon \mathbf{1}\right\}.$$

Then

$$\Delta_{\boldsymbol{\theta}}(a) = \left(\sup_{b \in \mathcal{A}} \min_{j \in [d]} \big(\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a)\big)\right)_+.$$

Proof. Fix $a \in \mathcal{A}$. For each $b \in \mathcal{A}$, define

$$m_b := \min_{j \in [d]} \big(\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a)\big).$$

For any $b \in \mathcal{A}$ and $\varepsilon \geq 0$,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(b) \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \varepsilon \mathbf{1}$$

if and only if

$$\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a) \geq \varepsilon, \qquad \forall j \in [d],$$

which is equivalent to

$$\varepsilon \leq \min_{j \in [d]} \big(\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a)\big) = m_b.$$

Let

$$S_a := \left\{\varepsilon \geq 0 : \exists b \in \mathcal{A} \text{ such that } \boldsymbol{\mu}_{\boldsymbol{\theta}}(b) \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \varepsilon \mathbf{1}\right\}.$$

The preceding equivalence yields

$$S_a = \left\{\varepsilon \geq 0 : \exists b \in \mathcal{A} \text{ such that } \varepsilon \leq m_b\right\}.$$

Since $\mathcal{A}$ is finite,

$$M := \max_{b \in \mathcal{A}} m_b$$

is attained by some $b^* \in \mathcal{A}$. Moreover, taking $b = a$ gives

$$m_a = \min_{j \in [d]} \big(\mu_{\boldsymbol{\theta},j}(a) - \mu_{\boldsymbol{\theta},j}(a)\big) = 0,$$

and therefore $M \geq 0$.

For every $\varepsilon \in [0, M]$, choosing $b = b^*$ gives

$$\varepsilon \leq m_{b^*} = M,$$

so $\varepsilon \in S_a$. Conversely, if $\varepsilon \in S_a$, then there exists $b \in \mathcal{A}$ such that

$$\varepsilon \leq m_b \leq M.$$

Hence

$$S_a = [0, M].$$

It follows that

$$\begin{aligned}\Delta_{\boldsymbol{\theta}}(a) &= \sup S_a \\ &= M \\ &= \max_{b \in \mathcal{A}} \min_{j \in [d]} \big(\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a)\big).\end{aligned}$$

Because $\mathcal{A}$ is finite, the maximum equals the corresponding supremum. Moreover, this quantity is nonnegative because the choice $b = a$ gives value zero. It therefore equals its positive part, and hence

$$\Delta_{\boldsymbol{\theta}}(a) = \left(\sup_{b \in \mathcal{A}} \min_{j \in [d]} \big(\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a)\big)\right)_+.$$

□

A.12 Proof of Lemma 4.1

Lemma 4.1. *Fix a round t, a realizable history $\mathcal{H}_t$, and an action $a \in \mathcal{A}$. Let p_t denote the posterior distribution of $\boldsymbol{\theta}$ given $\mathcal{H}_t$, and let $\boldsymbol{\theta}, \boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t$. Under the fixed-action interpretation,*

$$I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \,|\, \mathcal{H}_t, A_t = a) \geq \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t}\left[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2\right]. \tag{2}$$

Proof. Fix a round t, a realizable history $\mathcal{H}_t$, and an action $a \in \mathcal{A}$. Fix versions of the relevant regular conditional distributions throughout. Quantities conditioned on $A_t = a$ are understood under the fixed-action experiment in which action a is selected at round t. Whenever $\mathbb{P}(A_t = a \mid \mathcal{H}_t) > 0$, this coincides with ordinary conditioning on the event $A_t = a$.

Let p_t denote the posterior law of $\boldsymbol{\theta}$ given $\mathcal{H}_t$. Since A_t is selected using only $\mathcal{H}_t$ and fresh algorithmic randomness independent of the true parameter $\boldsymbol{\theta}$,

$$A_t \perp\!\!\!\perp \boldsymbol{\theta} \mid \mathcal{H}_t.$$

Hence fixing $A_t = a$ does not alter the posterior law of $\boldsymbol{\theta}$, and under the fixed-action experiment,

$$\boldsymbol{\theta} \sim p_t.$$

For each $\boldsymbol{\vartheta} \in \Theta$, define the action-specific reward law

$$P^a_{\boldsymbol{\vartheta}} := \mathcal{L}(\mathbf{Y}_{t,a} \mid \boldsymbol{\theta} = \boldsymbol{\vartheta}).$$

Because the reward arrays are i.i.d. across rounds conditional on $\boldsymbol{\theta}$, the current reward $\mathbf{Y}_{t,a}$ is conditionally independent of the past given $\boldsymbol{\theta}$. Moreover, the fresh action-selection randomization is independent of the current reward conditional on $(\mathcal{H}_t, \boldsymbol{\theta})$. Therefore,

$$\mathcal{L}(\mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a, \boldsymbol{\theta} = \boldsymbol{\vartheta}) = P^a_{\boldsymbol{\vartheta}}.$$

Define the posterior-predictive law for action a by

$$P^a_t := \int_\Theta P^a_{\boldsymbol{\vartheta}}\, p_t(d\boldsymbol{\vartheta}).$$

Since $\boldsymbol{\theta} \sim p_t$ under the fixed-action law, marginalizing over $\boldsymbol{\theta}$ gives, for every measurable set E,

$$\mathbb{P}(\mathbf{Y}_{t,a} \in E \mid \mathcal{H}_t, A_t = a) = \int_\Theta P^a_{\boldsymbol{\vartheta}}(E)\, p_t(d\boldsymbol{\vartheta}) = P^a_t(E).$$

Thus P^a_t is precisely the conditional marginal law of $\mathbf{Y}_{t,a}$ given $\mathcal{H}_t$ under the fixed action a.

By the standard representation of mutual information as the expected KL divergence between the conditional observation law and its marginal law,

$$\begin{aligned} I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) &= \int_\Theta \mathrm{KL}(P^a_{\boldsymbol{\vartheta}} \,\|\, P^a_t)\, p_t(d\boldsymbol{\vartheta}) \\ &= \mathbb{E}_{\boldsymbol{\theta}\sim p_t}[\mathrm{KL}(P^a_{\boldsymbol{\theta}} \,\|\, P^a_t)]. \end{aligned} \tag{53}$$

This identity is understood in the extended sense if the mutual information is infinite.

By Pinsker's inequality, using natural logarithms,

$$\mathrm{KL}(P^a_{\boldsymbol{\theta}} \,\|\, P^a_t) \geq 2\ \mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_t)^2.$$

Taking expectation with respect to $\boldsymbol{\theta} \sim p_t$ and using Equation (53) yields

$$I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \geq 2\,\mathbb{E}_{\boldsymbol{\theta}\sim p_t}\left[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_t)^2\right]. \tag{54}$$

Now let $\boldsymbol{\theta}, \boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t$. For every realization of $(\boldsymbol{\theta}, \boldsymbol{\theta}')$, the triangle inequality for total variation gives

$$\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'}) \leq \mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_t) + \mathrm{TV}(P^a_{\boldsymbol{\theta}'}, P^a_t).$$

Using $(x+y)^2 \le 2x^2 + 2y^2$, we obtain

$$\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2 \le 2\,\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_t)^2 + 2\,\mathrm{TV}(P^a_{\boldsymbol{\theta}'}, P^a_t)^2 .$$

Taking expectation over the independent posterior draws and using that both have marginal law p_t gives

$$\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t}\Big[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2\Big] \le 4\,\mathbb{E}_{\boldsymbol{\theta}\sim p_t}\Big[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_t)^2\Big] .$$

Equivalently,

$$\mathbb{E}_{\boldsymbol{\theta}\sim p_t}\Big[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_t)^2\Big] \ge \frac{1}{4}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t}\Big[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2\Big] .$$

Combining this inequality with Equation (54) yields

$$I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a) \ge \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t}\Big[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2\Big] .$$

This proves the claim. □

A.13 Proof of Theorem 4.1

Theorem 4.1. *Under bounded rewards, for Algorithm 3, the Bayesian cumulative gap-based Pareto regret satisfies* $\mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\Big] = O\Big(\sqrt{|\mathcal{A}|\, M\, \mathrm{IB}_T\, T}\Big)$*, where* $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$*,* $\mathcal{H}_{T+1} := (A_1, \mathbf{Y}_{1,A_1}, \ldots, A_T, \mathbf{Y}_{T,A_T})$ *denotes the observed interaction history up to horizon* T*, and* $M := \sup_{\boldsymbol{\theta}\in\mathrm{supp}(p_1)} |\mathcal{S}^*(\boldsymbol{\theta})| \le |\mathcal{A}|$*. Since* $M \le |\mathcal{A}|$*, the worst-case bound is* $O\big(|\mathcal{A}|\sqrt{\mathrm{IB}_T\, T}\big)$*.*

Lemma A.1. *Fix a round* t *and condition on* $\mathcal{H}_t$*. Let* $\boldsymbol{\theta}, \boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t$*, where* $p_t(\boldsymbol{\theta}) = \mathbb{P}(\boldsymbol{\theta} \mid \mathcal{H}_t)$*. Define* $S := \mathcal{S}^*(\boldsymbol{\theta})$ *and* $S' := \mathcal{S}^*(\boldsymbol{\theta}')$*. Conditional on* $(\boldsymbol{\theta}, \boldsymbol{\theta}')$*, let* $A \sim \mathrm{Unif}(S')$*. Suppose that every reward coordinate lies in* $[0,1]$*. Then*

$$\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A) \mid \mathcal{H}_t] \le 4 \sum_{a\in\mathcal{A}} \mathbb{E}[\mathbb{1}\{a \in S'\}\, \mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'}) | \mathcal{H}_t] . \tag{55}$$

Proof. Fix $\mathcal{H}_t$ throughout and suppress this conditioning whenever no ambiguity arises. Fix realizations of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, and let $a \in S' = \mathcal{S}^*(\boldsymbol{\theta}')$.

We first construct a Pareto-optimal witness under $\boldsymbol{\theta}$. Since $\mathcal{A}$ is finite, the Pareto gap admits the finite maximum representation

$$\Delta_{\boldsymbol{\theta}}(a) = \max_{u\in\mathcal{A}} \left(\min_{j\in[d]} [\mu_{\boldsymbol{\theta},j}(u) - \mu_{\boldsymbol{\theta},j}(a)] \right)_+ .$$

If $\Delta_{\boldsymbol{\theta}}(a) > 0$, choose, using a fixed deterministic tie-breaking rule, an action u attaining this maximum. Then

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(u) \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \Delta_{\boldsymbol{\theta}}(a)\mathbf{1}.$$

If $\Delta_{\boldsymbol{\theta}}(a) = 0$, take $u = a$, for which the same inequality holds. Thus, in all cases,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(u) \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \Delta_{\boldsymbol{\theta}}(a)\mathbf{1}. \tag{56}$$

Among the actions that weakly dominate u under $\boldsymbol{\theta}$, choose, again using a fixed deterministic tie-breaking rule, a $\succeq$-maximal action $b(a)$. Such an action exists because $\mathcal{A}$ is finite. If some action strictly Pareto dominated $b(a)$, then, since $b(a) \succeq u$, that action would also weakly dominate u, contradicting the maximality of $b(a)$. Hence

$$b(a) \in S = \mathcal{S}^*(\boldsymbol{\theta}).$$

Moreover,

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(b(a)) \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(u),$$

and therefore Equation (56) implies

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(b(a)) \succeq \boldsymbol{\mu}_{\boldsymbol{\theta}}(a) + \Delta_{\boldsymbol{\theta}}(a)\mathbf{1}. \tag{57}$$

Since $a \in S' = \mathcal{S}^*(\boldsymbol{\theta}')$, the action $b(a)$ cannot strictly Pareto dominate a under $\boldsymbol{\theta}'$. Consequently, there exists at least one coordinate $i = i(a) \in [d]$ for which

$$\mu_{\boldsymbol{\theta}',i}(b(a)) \le \mu_{\boldsymbol{\theta}',i}(a). \tag{58}$$

At the same coordinate, Equation (57) gives

$$\mu_{\boldsymbol{\theta},i}(b(a)) \ge \mu_{\boldsymbol{\theta},i}(a) + \Delta_{\boldsymbol{\theta}}(a). \tag{59}$$

Combining Equation (58) and Equation (59),

$$\begin{aligned}\Delta_{\boldsymbol{\theta}}(a) &\le \mu_{\boldsymbol{\theta},i}(b(a)) - \mu_{\boldsymbol{\theta},i}(a) \\ &= (\mu_{\boldsymbol{\theta},i}(b(a)) - \mu_{\boldsymbol{\theta}',i}(b(a))) + (\mu_{\boldsymbol{\theta}',i}(b(a)) - \mu_{\boldsymbol{\theta}',i}(a)) \\ &\quad + (\mu_{\boldsymbol{\theta}',i}(a) - \mu_{\boldsymbol{\theta},i}(a)) \\ &\le |\mu_{\boldsymbol{\theta},i}(b(a)) - \mu_{\boldsymbol{\theta}',i}(b(a))| + |\mu_{\boldsymbol{\theta}',i}(a) - \mu_{\boldsymbol{\theta},i}(a)|\,. \end{aligned} \tag{60}$$

We next relate these mean differences to total variation. For any two probability laws P, Q and any measurable function f satisfying $\|f\|_\infty \le 1$, the dual characterization of total variation implies

$$|\mathbb{E}_P[f] - \mathbb{E}_Q[f]| \le 2\,\mathrm{TV}(P, Q).$$

Since every reward coordinate lies in $[0, 1]$, this applies to each coordinate projection. Hence

$$|\mu_{\boldsymbol{\theta},i}(b(a)) - \mu_{\boldsymbol{\theta}',i}(b(a))| \le 2\,\mathrm{TV}\Big(P_{\boldsymbol{\theta}}^{b(a)}, P_{\boldsymbol{\theta}'}^{b(a)}\Big)\,,$$

and

$$|\mu_{\boldsymbol{\theta}',i}(a) - \mu_{\boldsymbol{\theta},i}(a)| \le 2\,\mathrm{TV}(P_{\boldsymbol{\theta}}^{a}, P_{\boldsymbol{\theta}'}^{a})\,.$$

Substituting these bounds into Equation (60) gives

$$\Delta_{\boldsymbol{\theta}}(a) \le 2\,\mathrm{TV}\Big(P_{\boldsymbol{\theta}}^{b(a)}, P_{\boldsymbol{\theta}'}^{b(a)}\Big) + 2\,\mathrm{TV}(P_{\boldsymbol{\theta}}^{a}, P_{\boldsymbol{\theta}'}^{a})\,. \tag{61}$$

Conditional on $(\boldsymbol{\theta}, \boldsymbol{\theta}', \mathcal{H}_t)$, the action A is uniform on S'. Averaging Equation (61) therefore gives

$$\begin{aligned}&\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A) \mid \boldsymbol{\theta}, \boldsymbol{\theta}', \mathcal{H}_t] \\ &\le \frac{2}{|S'|}\sum_{a\in S'} \mathrm{TV}\Big(P_{\boldsymbol{\theta}}^{b(a)}, P_{\boldsymbol{\theta}'}^{b(a)}\Big) + \frac{2}{|S'|}\sum_{a\in S'} \mathrm{TV}(P_{\boldsymbol{\theta}}^{a}, P_{\boldsymbol{\theta}'}^{a})\,. \end{aligned} \tag{62}$$

For every $a \in S'$, the witness $b(a)$ belongs to S. Consequently,

$$\begin{aligned}\frac{1}{|S'|}\sum_{a\in S'} \mathrm{TV}\Big(P_{\boldsymbol{\theta}}^{b(a)}, P_{\boldsymbol{\theta}'}^{b(a)}\Big) &\le \max_{u\in S} \mathrm{TV}(P_{\boldsymbol{\theta}}^{u}, P_{\boldsymbol{\theta}'}^{u}) \\ &\le \sum_{u\in S} \mathrm{TV}(P_{\boldsymbol{\theta}}^{u}, P_{\boldsymbol{\theta}'}^{u})\,. \end{aligned} \tag{63}$$

Substituting Equation (63) into Equation (62),

$$\begin{aligned}&\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A) \mid \boldsymbol{\theta}, \boldsymbol{\theta}', \mathcal{H}_t] \\ &\le 2\sum_{u\in S} \mathrm{TV}(P_{\boldsymbol{\theta}}^{u}, P_{\boldsymbol{\theta}'}^{u}) + \frac{2}{|S'|}\sum_{a\in S'} \mathrm{TV}(P_{\boldsymbol{\theta}}^{a}, P_{\boldsymbol{\theta}'}^{a})\,. \end{aligned} \tag{64}$$

We now average over $\boldsymbol{\theta}, \boldsymbol{\theta}' \overset{\text{i.i.d.}}{\sim} p_t$. Since the two posterior draws are exchangeable and total variation is symmetric,

$$\begin{aligned}&\mathbb{E}\left[\sum_{u\in S} \mathrm{TV}(P_{\boldsymbol{\theta}}^{u}, P_{\boldsymbol{\theta}'}^{u}) \,\middle|\, \mathcal{H}_t\right] \\ &= \mathbb{E}\left[\sum_{a\in S'} \mathrm{TV}(P_{\boldsymbol{\theta}}^{a}, P_{\boldsymbol{\theta}'}^{a}) \,\middle|\, \mathcal{H}_t\right] \\ &= \sum_{a\in\mathcal{A}} \mathbb{E}[\mathbb{1}\{a \in S'\}\,\mathrm{TV}(P_{\boldsymbol{\theta}}^{a}, P_{\boldsymbol{\theta}'}^{a}) | \mathcal{H}_t]\,. \end{aligned} \tag{65}$$

For the second term in Equation (64), the Pareto set S' is nonempty because $\mathcal{A}$ is finite and nonempty. Therefore $|S'| \geq 1$, and

$$\frac{\mathbb{1}\{a \in S'\}}{|S'|} \leq \mathbb{1}\{a \in S'\}.$$

Hence

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{|S'|}\sum_{a\in S'} \mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})\middle|\mathcal{H}_t\right] \\
&= \sum_{a\in\mathcal{A}} \mathbb{E}\left[\frac{\mathbb{1}\{a \in S'\}}{|S'|}\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})\middle|\mathcal{H}_t\right] \\
&\leq \sum_{a\in\mathcal{A}} \mathbb{E}[\mathbb{1}\{a \in S'\}\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})|\mathcal{H}_t]\,. \qquad (66)
\end{aligned}$$

Taking expectation in Equation (64) and applying Equation (65) and Equation (66) yields

$$\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A) \mid \mathcal{H}_t] \leq 4\sum_{a\in\mathcal{A}} \mathbb{E}[\mathbb{1}\{a \in S'\}\,\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})|\mathcal{H}_t]\,.$$

This proves the lemma A.1. □

Lemma A.2. *Fix a round t and condition on $\mathcal{H}_t$. Suppose every reward coordinate lies in $[0, R_{\max}]$ almost surely for some finite $R_{\max}$. Let $\boldsymbol{\theta} \sim p_t$ denote the true posterior parameter and let $\boldsymbol{\theta}_t \sim p_t$ be an independent Thompson sample. Define*

$$S_t := \mathcal{S}^*(\boldsymbol{\theta}_t), \qquad A_t \sim \mathrm{Unif}(S_t),$$

and

$$\pi_t(a) := \mathbb{P}(A_t = a \mid \mathcal{H}_t).$$

For each realized history, define the fixed-action one-step information gain

$$I_t(a) := I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)\,,$$

where this quantity is understood pointwise in the realized history $\mathcal{H}_t$. Define

$$J_t := \sum_{a\in\mathcal{A}} \pi_t(a) I_t(a).$$

Then

$$(\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t])^2 \leq 32 R^2_{\max}|\mathcal{A}|M J_t. \qquad (67)$$

Proof. If $R_{\max} = 0$, then every reward coordinate is identically zero and every Pareto gap is zero, so the result is immediate. Assume $R_{\max} > 0$.

Apply Lemma A.1 to the normalized rewards

$$\overline{\mathbf{Y}}_{t,a} := \frac{\mathbf{Y}_{t,a}}{R_{\max}}.$$

Every coordinate of $\overline{\mathbf{Y}}_{t,a}$ lies in $[0, 1]$. Positive rescaling preserves all Pareto-dominance relations, so

$$\overline{\mathcal{S}}^*(\boldsymbol{\theta}) = \mathcal{S}^*(\boldsymbol{\theta})$$

for every parameter value, and the normalized Pareto gap satisfies

$$\overline{\Delta}_{\boldsymbol{\theta}}(a) = \frac{\Delta_{\boldsymbol{\theta}}(a)}{R_{\max}}.$$

Moreover, $y \mapsto y/R_{\max}$ is a measurable bijection with measurable inverse, so total variation is invariant under this common transformation:

$$\mathrm{TV}\left(\overline{P}^a_{\boldsymbol{\theta}}, \overline{P}^a_{\boldsymbol{\theta}_t}\right) = \mathrm{TV}\left(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\right).$$

Conditional on $\mathcal{H}_t$, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_t$ are independent draws from p_t, and conditional on $(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$, $A_t \sim \mathrm{Unif}(S_t)$. Therefore Lemma A.1 applies with

$$(\boldsymbol{\theta}, \boldsymbol{\theta}') = (\boldsymbol{\theta}, \boldsymbol{\theta}_t)$$

and $A = A_t$. It gives

$$\frac{1}{R_{\max}} \mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t] \le 4 \sum_{a \in \mathcal{A}} \mathbb{E}\big[\mathbb{1}\{a \in S_t\} \, \mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big) \big| \mathcal{H}_t\big] .$$

Thus

$$\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t] \le 4R_{\max} \sum_{a \in \mathcal{A}} \mathbb{E}\big[\mathbb{1}\{a \in S_t\} \, \mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big) \big| \mathcal{H}_t\big] . \tag{68}$$

For each $a \in \mathcal{A}$, define

$$q_t(a) := \mathbb{P}(a \in S_t \mid \mathcal{H}_t) = \mathbb{E}[\mathbb{1}\{a \in S_t\} \mid \mathcal{H}_t] .$$

Since A_t is drawn uniformly from S_t,

$$\begin{aligned} \pi_t(a) &= \mathbb{P}(A_t = a \mid \mathcal{H}_t) \\ &= \mathbb{E}[\mathbb{P}(A_t = a \mid \boldsymbol{\theta}_t, \mathcal{H}_t) | \mathcal{H}_t] \\ &= \mathbb{E}\left[\frac{\mathbb{1}\{a \in S_t\}}{|S_t|} \middle| \mathcal{H}_t \right] . \end{aligned} \tag{69}$$

Recall

$$M := \sup_{\vartheta \in \mathrm{supp}(p_1)} |\mathcal{S}^*(\vartheta)|.$$

Since

$$\mathbb{P}(\boldsymbol{\theta} \in \mathrm{supp}(p_1)) = 1,$$

we also have

$$\mathbb{P}(\boldsymbol{\theta} \in \mathrm{supp}(p_1) \mid \mathcal{H}_t) = 1 \qquad \text{almost surely.}$$

Since $\boldsymbol{\theta}_t$ is sampled from this posterior conditional on $\mathcal{H}_t$, it follows that

$$\boldsymbol{\theta}_t \in \mathrm{supp}(p_1) \qquad \text{almost surely.}$$

Hence

$$|S_t| = |\mathcal{S}^*(\boldsymbol{\theta}_t)| \le M \qquad \text{almost surely.}$$

Consequently,

$$\frac{\mathbb{1}\{a \in S_t\}}{|S_t|} \ge \frac{\mathbb{1}\{a \in S_t\}}{M}.$$

Taking conditional expectation and using Equation (69),

$$\pi_t(a) \ge \frac{q_t(a)}{M}. \tag{70}$$

Hence, whenever $q_t(a) > 0$, $\pi_t(a) > 0$ and

$$\frac{q_t(a)}{\pi_t(a)} \le M. \tag{71}$$

If $q_t(a) = 0$, then $\mathbb{1}\{a \in S_t\} = 0$ almost surely conditional on $\mathcal{H}_t$, so its corresponding term below vanishes.

For a fixed action a, conditional Cauchy–Schwarz gives

$$\begin{aligned} &\mathbb{E}\big[\mathbb{1}\{a \in S_t\} \, \mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big) \big| \mathcal{H}_t\big] \\ &\quad \le \sqrt{q_t(a)} \sqrt{\mathbb{E}\Big[\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)^2 \Big| \mathcal{H}_t\Big]}. \end{aligned} \tag{72}$$

For $q_t(a) > 0$, insert $\pi_t(a)$:

$$\sqrt{q_t(a)}\sqrt{\mathbb{E}\Big[\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)^2\Big|\mathcal{H}_t\Big]} = \sqrt{\frac{q_t(a)}{\pi_t(a)}}\sqrt{\pi_t(a)\mathbb{E}\Big[\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)^2\Big|\mathcal{H}_t\Big]}. \tag{73}$$

Summing over actions and applying Cauchy–Schwarz gives

$$\left(\sum_{a\in\mathcal{A}}\mathbb{E}\big[\mathbb{1}\{a\in S_t\}\,\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)\big|\mathcal{H}_t\big]\right)^2 \leq \left(\sum_{\substack{a\in\mathcal{A}\\ q_t(a)>0}}\frac{q_t(a)}{\pi_t(a)}\right)\left(\sum_{a\in\mathcal{A}}\pi_t(a)\mathbb{E}\Big[\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)^2\Big|\mathcal{H}_t\Big]\right). \tag{74}$$

By Equation (71),

$$\sum_{\substack{a\in\mathcal{A}\\ q_t(a)>0}}\frac{q_t(a)}{\pi_t(a)} \leq |\mathcal{A}|M. \tag{75}$$

For the second factor, conditional on $\mathcal{H}_t$, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_t$ are independent draws from p_t. Therefore, for every fixed a,

$$\mathbb{E}\Big[\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)^2\Big|\mathcal{H}_t\Big] = \mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'\overset{\text{i.i.d.}}{\sim}p_t}\Big[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2\Big]. \tag{76}$$

Lemma 4.1 gives

$$I_t(a) \geq \frac{1}{2}\mathbb{E}_{\boldsymbol{\theta},\boldsymbol{\theta}'\overset{\text{i.i.d.}}{\sim}p_t}\Big[\mathrm{TV}(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}'})^2\Big].$$

Combining this with Equation (76),

$$\mathbb{E}\Big[\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)^2\Big|\mathcal{H}_t\Big] \leq 2I_t(a).$$

Multiplying by $\pi_t(a)$ and summing gives

$$\sum_{a\in\mathcal{A}}\pi_t(a)\mathbb{E}\Big[\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)^2\Big|\mathcal{H}_t\Big] \leq 2J_t. \tag{77}$$

Combining Equation (74), Equation (75), and Equation (77),

$$\left(\sum_{a\in\mathcal{A}}\mathbb{E}\big[\mathbb{1}\{a\in S_t\}\,\mathrm{TV}\big(P^a_{\boldsymbol{\theta}}, P^a_{\boldsymbol{\theta}_t}\big)\big|\mathcal{H}_t\big]\right)^2 \leq 2|\mathcal{A}|MJ_t. \tag{78}$$

Squaring Equation (68) and applying Equation (78),

$$\left(\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t)\mid\mathcal{H}_t]\right)^2 \leq 32R^2_{\max}|\mathcal{A}|MJ_t.$$

This proves the lemma A.2. □

Proof of Theorem 4.1. Without loss of generality, suppose that the rewards are uniformly bounded and let $R_{\max} < \infty$ be such that every reward coordinate lies almost surely in $[0, R_{\max}]$. Indeed, a fixed coordinate-wise shift is a bijective transformation of the observations and leaves all reward differences, Pareto-dominance relations, Pareto gaps, and the mutual information with the observed history unchanged. The hidden constant in the bound below may depend on this fixed reward bound.

We prove the nontrivial case

$$\mathrm{IB}_T = I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) < \infty.$$

If $\mathrm{IB}_T = \infty$, the explicit upper bound below is understood in the extended sense and is vacuous.

For each round t, define

$$r_t := \mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t]\,.$$

By Lemma A.2,

$$r_t^2 \le 32 R_{\max}^2 |\mathcal{A}| M J_t.$$

Since $r_t \ge 0$,

$$r_t \le 4\sqrt{2}\, R_{\max} \sqrt{|\mathcal{A}| M J_t}. \tag{79}$$

By linearity of expectation and the tower property,

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] &= \sum_{t=1}^{T} \mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t)] \\ &= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t]] \\ &= \sum_{t=1}^{T} \mathbb{E}[r_t]. \end{aligned} \tag{80}$$

Applying Equation (79),

$$\mathbb{E}[r_t] \le 4\sqrt{2}\, R_{\max} \sqrt{|\mathcal{A}| M}\, \mathbb{E}\left[\sqrt{J_t}\right].$$

Hence

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] \le 4\sqrt{2}\, R_{\max} \sqrt{|\mathcal{A}| M} \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{J_t}\right]. \tag{81}$$

Since $J_t \ge 0$ and $x \mapsto \sqrt{x}$ is concave,

$$\mathbb{E}\left[\sqrt{J_t}\right] \le \sqrt{\mathbb{E}[J_t]}.$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] \le 4\sqrt{2}\, R_{\max} \sqrt{|\mathcal{A}| M} \sum_{t=1}^{T} \sqrt{\mathbb{E}[J_t]}.$$

Applying Cauchy–Schwarz across time,

$$\sum_{t=1}^{T} \sqrt{\mathbb{E}[J_t]} \le \sqrt{T \sum_{t=1}^{T} \mathbb{E}[J_t]}.$$

Thus

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] \le 4\sqrt{2}\, R_{\max} \sqrt{|\mathcal{A}| M\, T \sum_{t=1}^{T} \mathbb{E}[J_t]}. \tag{82}$$

It remains to identify the cumulative one-step information gain. Conditional on $\mathcal{H}_t$, the true parameter has law p_t, and Algorithm 3 independently draws

$$\boldsymbol{\theta}_t \sim p_t.$$

Hence

$$\boldsymbol{\theta} \perp\!\!\!\perp \boldsymbol{\theta}_t \mid \mathcal{H}_t.$$

The algorithm then forms

$$S_t = \mathcal{S}^*(\boldsymbol{\theta}_t)$$

and samples

$$A_t \sim \mathrm{Unif}(S_t)$$

using fresh internal randomization independent of the true parameter. Therefore

$$\boldsymbol{\theta} \perp\!\!\!\perp A_t \mid \mathcal{H}_t,$$

and consequently

$$I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) = 0. \tag{83}$$

For each realized history,

$$J_t = \sum_{a \in \mathcal{A}} \pi_t(a) I_t(a).$$

For every action with $\pi_t(a) > 0$, the fixed-action interpretation of $I_t(a)$ agrees with ordinary conditioning on $A_t = a$; actions with $\pi_t(a) = 0$ contribute zero to the sum. Hence the standard decomposition of conditional mutual information over the discrete action A_t gives

$$\begin{aligned} \mathbb{E}[J_t] &= \mathbb{E}\left[\sum_{a \in \mathcal{A}} \pi_t(a) I_t(a)\right] \\ &= I(\boldsymbol{\theta}; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t)\,. \end{aligned} \tag{84}$$

Applying the mutual-information chain rule,

$$\begin{aligned} I(\boldsymbol{\theta}; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) &= I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) + I(\boldsymbol{\theta}; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t) \\ &= \mathbb{E}[J_t], \end{aligned} \tag{85}$$

where the final equality follows from Equation (83) and Equation (84).

Since

$$\mathcal{H}_{t+1} = (\mathcal{H}_t, A_t, \mathbf{Y}_{t,A_t})\,,$$

and $\mathcal{H}_1$ is empty, applying the mutual-information chain rule sequentially over $t = 1, \ldots, T$ gives

$$\begin{aligned} I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) &= \sum_{t=1}^{T} I(\boldsymbol{\theta}; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) \\ &= \sum_{t=1}^{T} \mathbb{E}[J_t]. \end{aligned} \tag{86}$$

By definition,

$$\mathrm{IB}_T = I(\boldsymbol{\theta}; \mathcal{H}_{T+1})\,,$$

so

$$\sum_{t=1}^{T} \mathbb{E}[J_t] = \mathrm{IB}_T.$$

Substituting into Equation (82),

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] \le 4\sqrt{2}\, R_{\max} \sqrt{|\mathcal{A}| M\, \mathrm{IB}_T\, T}.$$

Since $R_{\max}$ is a fixed finite reward bound, its multiplicative contribution is absorbed into the hidden constant. Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] = O\Big(\sqrt{|\mathcal{A}|\, M\, \mathrm{IB}_T\, T}\Big)\,.$$

Finally,

$$M = \sup_{\boldsymbol{\theta} \in \mathrm{supp}(p_1)} |\mathcal{S}^*(\boldsymbol{\theta})| \le |\mathcal{A}|.$$

Hence

$$\sqrt{|\mathcal{A}| M} \le |\mathcal{A}|,$$

and consequently

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] = O\Big(|\mathcal{A}| \sqrt{\mathrm{IB}_T\, T}\Big)\,.$$

This completes the proof. □

A.14 Proof of Corollary 4.1

Corollary 4.1. *Under bounded rewards, for Algorithm 3, let $S_t := \mathcal{S}^*(\boldsymbol{\theta}_t)$ denote the sampled Pareto set at round t. Then $\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{a\in S_t}\Delta_{\boldsymbol{\theta}}(a)\Big] = O\Big(\sqrt{|\mathcal{A}|\,M^3\,\mathrm{IB}_T\,T}\Big)$, where $\mathrm{IB}_T := I(\boldsymbol{\theta};\mathcal{H}_{T+1})$. In particular, since $M \le |\mathcal{A}|$, the worst-case bound is $O\big(|\mathcal{A}|^2\sqrt{\mathrm{IB}_T T}\big)$.*

Proof. Fix a round t. Under Algorithm 3, after drawing $\boldsymbol{\theta}_t$ from the posterior, the algorithm forms $S_t = \mathcal{S}^*(\boldsymbol{\theta}_t)$ and samples A_t uniformly from S_t. Hence, conditional on $\mathcal{H}_t, \boldsymbol{\theta}_t, \boldsymbol{\theta}$,

$$\mathbb{P}(A_t = a \mid \mathcal{H}_t, \boldsymbol{\theta}_t, \boldsymbol{\theta}) = \frac{1}{|S_t|} \qquad \text{for every } a \in S_t.$$

Therefore,

$$\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t, \boldsymbol{\theta}_t, \boldsymbol{\theta}] = \frac{1}{|S_t|}\sum_{a\in S_t}\Delta_{\boldsymbol{\theta}}(a),$$

and thus

$$\sum_{a\in S_t}\Delta_{\boldsymbol{\theta}}(a) = |S_t|\,\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t, \boldsymbol{\theta}_t, \boldsymbol{\theta}]\,.$$

Recall that

$$M := \sup_{\boldsymbol{\vartheta}\in\mathrm{supp}(p_1)} |\mathcal{S}^*(\boldsymbol{\vartheta})|\,.$$

Since

$$\mathbb{P}(\boldsymbol{\theta} \in \mathrm{supp}(p_1)) = 1,$$

we also have

$$\mathbb{P}(\boldsymbol{\theta} \in \mathrm{supp}(p_1) \mid \mathcal{H}_t) = 1 \qquad \text{almost surely.}$$

Because $\boldsymbol{\theta}_t$ is sampled from this posterior conditional on $\mathcal{H}_t$, it follows that

$$\boldsymbol{\theta}_t \in \mathrm{supp}(p_1) \qquad \text{almost surely,}$$

and hence

$$|S_t| = |\mathcal{S}^*(\boldsymbol{\theta}_t)| \le M \qquad \text{almost surely.}$$

Since $\Delta_{\boldsymbol{\theta}}(A_t) \ge 0$,

$$\sum_{a\in S_t}\Delta_{\boldsymbol{\theta}}(a) \le M\,\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t, \boldsymbol{\theta}_t, \boldsymbol{\theta}]\,.$$

Summing over $t = 1, \dots, T$, taking expectations, and applying the tower property term by term gives

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in S_t}\Delta_{\boldsymbol{\theta}}(a)\right] &\le M\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t, \boldsymbol{\theta}_t, \boldsymbol{\theta}]] \\
&= M\sum_{t=1}^{T}\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t)] \\
&= M\,\mathbb{E}\left[\sum_{t=1}^{T}\Delta_{\boldsymbol{\theta}}(A_t)\right].
\end{aligned}$$

By Theorem 4.1,

$$\mathbb{E}\left[\sum_{t=1}^{T}\Delta_{\boldsymbol{\theta}}(A_t)\right] = O\Big(\sqrt{|\mathcal{A}|\,M\,\mathrm{IB}_T\,T}\Big)\,.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\in S_t}\Delta_{\boldsymbol{\theta}}(a)\right] &= O\Big(M\sqrt{|\mathcal{A}|\,M\,\mathrm{IB}_T\,T}\Big) \\
&= O\Big(\sqrt{|\mathcal{A}|\,M^3\,\mathrm{IB}_T\,T}\Big)\,.
\end{aligned}$$

Finally, since $M \le |\mathcal{A}|$,

$$\sqrt{|\mathcal{A}|\, M^3\, \mathrm{IB}_T\, T} \le \sqrt{|\mathcal{A}|^4\, \mathrm{IB}_T\, T} = |\mathcal{A}|^2 \sqrt{\mathrm{IB}_T\, T}.$$

Hence, in the worst case,

$$\mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in S_t} \Delta_{\boldsymbol{\theta}}(a)\right] = O\Big(|\mathcal{A}|^2 \sqrt{\mathrm{IB}_T\, T}\Big).$$

This proves the corollary. □

A.15 Proof of Theorem 4.2

Theorem 4.2. *Assume that rewards are bounded and that, for some $\Psi > 0$, $\psi_t(\pi_t^{\mathrm{IDS}}) \le \Psi$ for every realizable $\mathcal{H}_t$, $\forall t \in [T]$. Then Algorithm 4 satisfies $\mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\Big] \le \sqrt{\Psi\, \mathrm{IB}_T\, T}$, where $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$. Moreover, there exists a constant $C > 0$, depending only on the fixed reward bound, such that $\psi_t(\pi_t^{\mathrm{IDS}}) \le C|\mathcal{A}|M$ for all $t \in [T]$. Consequently, $\mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\Big] = O\Big(\sqrt{|\mathcal{A}|\, M\, \mathrm{IB}_T\, T}\Big)$. Since $M \le |\mathcal{A}|$, the worst-case bound is $O\big(|\mathcal{A}|\sqrt{\mathrm{IB}_T T}\big)$.*

Proof. Fix a round t and condition on the realized history $\mathcal{H}_t$. Recall that

$$\Delta_t(a) := \mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t]$$

and

$$I_t(a) := I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)\,,$$

where $I_t(a)$ is understood pointwise for the realized history $\mathcal{H}_t$ under the fixed-action interpretation. For any $\pi \in \Pi(\mathcal{A})$, define

$$\Delta_t(\pi) := \sum_{a \in \mathcal{A}} \pi(a)\Delta_t(a), \qquad I_t(\pi) := \sum_{a \in \mathcal{A}} \pi(a) I_t(a),$$

and

$$\psi_t(\pi) := \frac{\Delta_t(\pi)^2}{I_t(\pi)}.$$

We use the standard extended-value convention that $\psi_t(\pi) = 0$ if $\Delta_t(\pi) = I_t(\pi) = 0$, and $\psi_t(\pi) = +\infty$ if $I_t(\pi) = 0$ while $\Delta_t(\pi) > 0$.

We first establish the generic information-ratio bound. Fix an $\mathcal{H}_t$-measurable minimizer π_t^{IDS}, using a fixed deterministic tie-breaking rule when the minimizer is not unique. Conditional on $\mathcal{H}_t$, Algorithm 4 samples $A_t \sim \pi_t^{\mathrm{IDS}}$ using fresh internal randomness. Hence

$$\boldsymbol{\theta} \perp\!\!\!\perp A_t \mid \mathcal{H}_t.$$

Therefore,

$$\begin{aligned}
\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t] &= \sum_{a \in \mathcal{A}} \mathbb{P}(A_t = a \mid \mathcal{H}_t)\mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t, A_t = a] \\
&= \sum_{a \in \mathcal{A}} \pi_t^{\mathrm{IDS}}(a)\mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t] \\
&= \sum_{a \in \mathcal{A}} \pi_t^{\mathrm{IDS}}(a)\Delta_t(a) \\
&= \Delta_t(\pi_t^{\mathrm{IDS}}).
\end{aligned} \tag{87}$$

Under the assumed bound

$$\psi_t(\pi_t^{\mathrm{IDS}}) \le \Psi < \infty,$$

if $I_t(\pi_t^{\mathrm{IDS}}) = 0$, then necessarily

$$\Delta_t(\pi_t^{\mathrm{IDS}}) = 0,$$

since otherwise the extended-value convention would give $\psi_t(\pi_t^{\mathrm{IDS}}) = +\infty$. If $I_t(\pi_t^{\mathrm{IDS}}) > 0$, then

$$\Delta_t(\pi_t^{\mathrm{IDS}})^2 = \psi_t(\pi_t^{\mathrm{IDS}}) I_t(\pi_t^{\mathrm{IDS}}) \le \Psi I_t(\pi_t^{\mathrm{IDS}}).$$

Hence, in either case,

$$\Delta_t(\pi_t^{\mathrm{IDS}}) \le \sqrt{\Psi\, I_t(\pi_t^{\mathrm{IDS}})}.$$

Using Equation (87), linearity of expectation, and the tower property,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] &= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[\Delta_{\boldsymbol{\theta}}(A_t) \mid \mathcal{H}_t]] \\
&= \sum_{t=1}^{T} \mathbb{E}\big[\Delta_t(\pi_t^{\mathrm{IDS}})\big] \\
&\le \sqrt{\Psi} \sum_{t=1}^{T} \mathbb{E}\left[\sqrt{I_t(\pi_t^{\mathrm{IDS}})}\right]. \qquad (88)
\end{aligned}$$

Since $x \mapsto \sqrt{x}$ is concave, Jensen's inequality gives

$$\mathbb{E}\left[\sqrt{I_t(\pi_t^{\mathrm{IDS}})}\right] \le \sqrt{\mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]}.$$

Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] \le \sqrt{\Psi} \sum_{t=1}^{T} \sqrt{\mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]}.$$

Applying Cauchy–Schwarz across time,

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] \le \sqrt{\Psi T \sum_{t=1}^{T} \mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]}. \qquad (89)$$

We now identify the information budget. Since A_t is generated from $\mathcal{H}_t$ and fresh internal randomness,

$$\boldsymbol{\theta} \perp\!\!\!\perp A_t \mid \mathcal{H}_t,$$

and therefore

$$I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) = 0. \qquad (90)$$

For every action with $\pi_t^{\mathrm{IDS}}(a) > 0$, the fixed-action interpretation of $I_t(a)$ agrees with ordinary conditioning on $A_t = a$; actions with zero selection probability contribute zero. Therefore, averaging over the selected action and then over the random history gives

$$\mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big] = I(\boldsymbol{\theta}; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t)\,. \qquad (91)$$

Thus, by the conditional mutual-information chain rule and Equation (90),

$$\begin{aligned}
I(\boldsymbol{\theta}; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) &= I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) + I(\boldsymbol{\theta}; \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t, A_t) \\
&= \mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big]\,. \qquad (92)
\end{aligned}$$

Since

$$\mathcal{H}_{T+1} = (A_1, \mathbf{Y}_{1,A_1}, \ldots, A_T, \mathbf{Y}_{T,A_T})\,,$$

the chain rule for mutual information yields

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\big[I_t(\pi_t^{\mathrm{IDS}})\big] &= \sum_{t=1}^{T} I(\boldsymbol{\theta}; A_t, \mathbf{Y}_{t,A_t} \mid \mathcal{H}_t) \\
&= I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) \\
&= \mathrm{IB}_T. \qquad (93)
\end{aligned}$$

Combining Equation (89) and Equation (93) gives

$$\mathbb{E}\left[\sum_{t=1}^{T}\Delta_{\boldsymbol{\theta}}(A_t)\right] \le \sqrt{\Psi\,\mathrm{IB}_T\,T}.$$

This proves the first claim.

We next establish the uniform information-ratio bound. Fix a round t and a realized history $\mathcal{H}_t$. Let π_t^{TS} denote the action distribution obtained by applying one step of Algorithm 3 to the current posterior p_t: independently draw $\boldsymbol{\theta}_t \sim p_t$, form

$$S_t = \mathcal{S}^*(\boldsymbol{\theta}_t),$$

and sample

$$A_t^{\mathrm{TS}} \sim \mathrm{Unif}(S_t).$$

Thus,

$$\pi_t^{\mathrm{TS}}(a) = \mathbb{P}\big(A_t^{\mathrm{TS}} = a \mid \mathcal{H}_t\big).$$

Lemma A.2 is a pointwise statement conditional on the current realized history and posterior. Therefore, it applies to this Thompson-sampling comparator even though the history $\mathcal{H}_t$ was generated by Algorithm 4. Moreover, the quantities $I_t(a)$ are fixed-action information gains determined by the current posterior and observation kernels, so the same $I_t(a)$ are used for both the IDS policy and this comparator. Let the fixed bound on the reward coordinates be $[0, R_{\max}]$. Then Lemma A.2 gives

$$\big(\mathbb{E}\big[\Delta_{\boldsymbol{\theta}}(A_t^{\mathrm{TS}}) \mid \mathcal{H}_t\big]\big)^2 \le 32R_{\max}^2|\mathcal{A}|M\sum_{a\in\mathcal{A}}\pi_t^{\mathrm{TS}}(a)I_t(a). \tag{94}$$

By definition,

$$\begin{aligned}\Delta_t(\pi_t^{\mathrm{TS}}) &= \sum_{a\in\mathcal{A}}\pi_t^{\mathrm{TS}}(a)\Delta_t(a)\\ &= \sum_{a\in\mathcal{A}}\pi_t^{\mathrm{TS}}(a)\mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t]\\ &= \mathbb{E}\big[\Delta_{\boldsymbol{\theta}}(A_t^{\mathrm{TS}}) \mid \mathcal{H}_t\big].\end{aligned} \tag{95}$$

Also,

$$I_t(\pi_t^{\mathrm{TS}}) = \sum_{a\in\mathcal{A}}\pi_t^{\mathrm{TS}}(a)I_t(a).$$

Therefore, Equation (94) becomes

$$\Delta_t(\pi_t^{\mathrm{TS}})^2 \le 32R_{\max}^2|\mathcal{A}|MI_t(\pi_t^{\mathrm{TS}}). \tag{96}$$

If $I_t(\pi_t^{\mathrm{TS}}) > 0$, dividing by $I_t(\pi_t^{\mathrm{TS}})$ gives

$$\psi_t(\pi_t^{\mathrm{TS}}) \le 32R_{\max}^2|\mathcal{A}|M.$$

If $I_t(\pi_t^{\mathrm{TS}}) = 0$, then Equation (96) implies

$$\Delta_t(\pi_t^{\mathrm{TS}}) = 0,$$

and hence, by the extended-value convention,

$$\psi_t(\pi_t^{\mathrm{TS}}) = 0.$$

Thus, in all cases,

$$\psi_t(\pi_t^{\mathrm{TS}}) \le 32R_{\max}^2|\mathcal{A}|M. \tag{97}$$

Since π_t^{IDS} minimizes $\psi_t(\pi)$ over all $\pi \in \Pi(\mathcal{A})$, and $\pi_t^{\mathrm{TS}} \in \Pi(\mathcal{A})$ is feasible,

$$\psi_t(\pi_t^{\mathrm{IDS}}) \le \psi_t(\pi_t^{\mathrm{TS}}).$$

Combining this with Equation (97),

$$\psi_t(\pi_t^{\mathrm{IDS}}) \le 32R_{\max}^2|\mathcal{A}|M.$$

Therefore, there exists a constant

$$C = 32R_{\max}^2,$$

depending only on the fixed reward bound, such that

$$\psi_t(\pi_t^{\mathrm{IDS}}) \le C|\mathcal{A}|M$$

for every $t \in [T]$. This proves the second claim.

Substituting $\Psi = C|\mathcal{A}|M$ into the generic information-ratio bound gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] \le \sqrt{C|\mathcal{A}|M\,\mathrm{IB}_T\,T} = O\Big(\sqrt{|\mathcal{A}|M\,\mathrm{IB}_T\,T}\Big),$$

where the hidden constant depends only on the fixed reward bound.

Finally, since $M \le |\mathcal{A}|$,

$$\sqrt{|\mathcal{A}|M} \le |\mathcal{A}|,$$

and therefore

$$\mathbb{E}\left[\sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t)\right] = O\Big(|\mathcal{A}|\sqrt{\mathrm{IB}_T\,T}\Big)$$

in the worst case. This completes the proof. □

A.16 Proof of Proposition 4.1

Proposition 4.1. *For finite-action Bayesian models, the following bounds hold for* $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$. *For finite parameter space* Θ*:* $\mathrm{IB}_T \le H(\boldsymbol{\theta}) \le \log|\Theta|$. *In the multi-objective categorical reward model with* d *objectives and* K *categories per objective, suppose that* $\boldsymbol{\theta}$ *consists of the categorical probability vectors and that each action–objective pair has an independent Dirichlet prior with parameters* $(\alpha_{a,j,1}, \dots, \alpha_{a,j,K})$ *satisfying* $\alpha_{a,j,k} \ge 1$ *for all* $a \in \mathcal{A}$, $j \in [d]$, *and* $k \in [K]$. *Then* $\mathrm{IB}_T = O(dK|\mathcal{A}|\log T)$. *In the finite-action Gaussian reward model, suppose that* $\boldsymbol{\theta} = (\{\mu_a\}_{a\in\mathcal{A}}, \Sigma)$, $\mathbf{Y}_{t,A_t} \mid (A_t = a, \boldsymbol{\theta}) \sim \mathcal{N}(\mu_a, \Sigma)$, Σ *has an inverse-Wishart prior, and* $\{\mu_a\}_{a\in\mathcal{A}}$ *have independent Gaussian priors conditional on* Σ. *Then* $\mathrm{IB}_T = O(d|\mathcal{A}|\log T + d^2 \log T)$.

Proof. The three bounds follow from Lemmas A.3, A.4, and A.5, respectively. □

Lemma A.3 (Information budget for finite parameter spaces)**.** *Suppose that the unknown model parameter* $\boldsymbol{\theta}$ *takes values in a finite parameter space* Θ. *Then, for every* $T \ge 1$,

$$\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) \le H(\boldsymbol{\theta}) \le \log|\Theta|.$$

Proof. The result follows from the entropy bound on mutual information. Since $\boldsymbol{\theta}$ is finite-valued,

$$I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) = H(\boldsymbol{\theta}) - H(\boldsymbol{\theta} \mid \mathcal{H}_{T+1}) \le H(\boldsymbol{\theta}),$$

because $H(\boldsymbol{\theta} \mid \mathcal{H}_{T+1}) \ge 0$. Also, since $\boldsymbol{\theta}$ is supported on Θ, we have

$$H(\boldsymbol{\theta}) \le \log|\Theta|.$$

Therefore,

$$\mathrm{IB}_T = I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) \le H(\boldsymbol{\theta}) \le \log|\Theta|.$$

□

Lemma A.4 (Logarithmic information growth for the multi-objective categorical reward model)**.** *Fix a finite action set* $\mathcal{A}$, *an objective dimension* $d \in \mathbb{N}$, *and a category size* $K \in \mathbb{N}$. *For each action* $a \in \mathcal{A}$ *and coordinate* $j \in [d] := \{1, \dots, d\}$, *let*

$$p_{a,j} = (p_{a,j,1}, \dots, p_{a,j,K}) \in \Delta^{K-1}, \qquad \Delta^{K-1} := \Big\{x \in [0,1]^K : \sum_{k=1}^{K} x_k = 1\Big\}.$$

Let $\boldsymbol{\theta} := (p_{a,j})_{a\in\mathcal{A},\, j\in[d]}$. *At round* t, *an action* $A_t \in \mathcal{A}$ *is selected, and the observed outcome is* $Y_t = (Y_{t,1}, \dots, Y_{t,d}) \in [K]^d$. *Assume that conditional on* $(\boldsymbol{\theta}, A_t = a, \mathcal{H}_t)$, *the coordinates are*

independent and satisfy $Y_{t,j} \mid (\boldsymbol{\theta}, A_t = a, \mathcal{H}_t) \sim \mathrm{Cat}(p_{a,j}), \forall j \in [d]$. *Assume independent Dirichlet priors across* (a, j)*:*

$$p_{a,j} \sim \mathrm{Dir}(\alpha_{a,j}), \qquad \alpha_{a,j} = (\alpha_{a,j,1}, \ldots, \alpha_{a,j,K}) \in (0, \infty)^K,$$

and $\alpha_{a,j,0} := \sum_{k=1}^{K} \alpha_{a,j,k}$. *Also assume that* $\alpha_{a,j,k} \geq 1$ *for all* $a \in \mathcal{A},\ j \in [d],\ k \in [K]$. *Let* $\mathcal{H}_t := (A_1, Y_1, \ldots, A_{t-1}, Y_{t-1})$ *denote the history before round* t*, and define the information budget* $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$. *Assume that the action-selection rule uses only* $\mathcal{H}_t$ *and fresh randomness independent of* $\boldsymbol{\theta}$. *Then*

$$\mathrm{IB}_T \leq d(K+1)|\mathcal{A}|(1 + \log T) = O(dK|\mathcal{A}| \log T).$$

Proof. We proceed in four steps.

Step 1: Posterior update.

For each $a \in \mathcal{A}$, $j \in [d]$, and $k \in [K]$, define

$$N_{a,j,k}(t) := \sum_{s=1}^{t} \mathbb{1}\{A_s = a,\ Y_{s,j} = k\}, \qquad N_a(t) := \sum_{s=1}^{t} \mathbb{1}\{A_s = a\}.$$

Fix a time $t \geq 1$. We derive the posterior of $p_{a,j}$ given the history $\mathcal{H}_t$. Because the prior factorizes independently across pairs (a, j) and the observation model factorizes across coordinates conditional on $(\boldsymbol{\theta}, A_s, \mathcal{H}_s)$, it is enough to consider a fixed pair (a, j). Moreover, since the policy selects A_s using only $\mathcal{H}_s$ and fresh randomness independent of $\boldsymbol{\theta}$, the policy factors do not contribute any $\boldsymbol{\theta}$-dependent terms to the posterior kernel.

Condition on the realized history $\mathcal{H}_t$. Among the first $t-1$ rounds, the observations relevant to $p_{a,j}$ are exactly those rounds $s \leq t-1$ for which $A_s = a$, in which case the coordinate $Y_{s,j}$ is drawn from $\mathrm{Cat}(p_{a,j})$. Hence the likelihood contribution of these observations is

$$L_{a,j}(p_{a,j}; \mathcal{H}_t) = \prod_{k=1}^{K} p_{a,j,k}^{N_{a,j,k}(t-1)}.$$

The prior density of $p_{a,j}$ is Dirichlet:

$$p_{a,j} \sim \mathrm{Dir}(\alpha_{a,j}),$$

so, up to normalization,

$$\pi_{a,j}(p_{a,j}) \propto \prod_{k=1}^{K} p_{a,j,k}^{\alpha_{a,j,k}-1}.$$

Multiplying prior and likelihood, the posterior density is proportional to

$$\pi_{a,j}(p_{a,j} \mid \mathcal{H}_t) \propto \prod_{k=1}^{K} p_{a,j,k}^{\alpha_{a,j,k}+N_{a,j,k}(t-1)-1}.$$

This is exactly the kernel of a Dirichlet distribution with parameter vector

$$\beta_{a,j}(t) = \big(\beta_{a,j,1}(t), \ldots, \beta_{a,j,K}(t)\big), \qquad \beta_{a,j,k}(t) := \alpha_{a,j,k} + N_{a,j,k}(t-1).$$

Therefore,

$$p_{a,j} \mid \mathcal{H}_t \sim \mathrm{Dir}(\beta_{a,j}(t)).$$

Its total concentration is

$$\beta_{a,j,0}(t) = \sum_{k=1}^{K} \beta_{a,j,k}(t) = \alpha_{a,j,0} + \sum_{k=1}^{K} N_{a,j,k}(t-1).$$

Since whenever action a is played, exactly one category is observed in coordinate j,

$$\sum_{k=1}^{K} N_{a,j,k}(t-1) = N_a(t-1).$$

Hence,

$$\beta_{a,j,0}(t) = \alpha_{a,j,0} + N_a(t-1).$$

Step 2: Information decomposition.

By definition,

$$\mathrm{IB}_T = I(\boldsymbol{\theta}; \mathcal{H}_{T+1}), \qquad \mathcal{H}_{T+1} = (A_1, Y_1, \ldots, A_T, Y_T).$$

Applying the chain rule for mutual information,

$$I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} I(\boldsymbol{\theta}; (A_t, Y_t) \mid \mathcal{H}_t).$$

Applying the conditional chain rule again,

$$I(\boldsymbol{\theta}; (A_t, Y_t) \mid \mathcal{H}_t) = I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) + I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t).$$

By assumption, A_t is generated using only $\mathcal{H}_t$ and fresh randomness independent of $\boldsymbol{\theta}$. Hence,

$$I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) = 0.$$

Therefore,

$$\mathrm{IB}_T = \sum_{t=1}^{T} I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t).$$

Step 3: Reduction to the played action and one-step mutual-information bound.

Fix t. By the chain rule for mutual information,

$$I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t) = \sum_{j=1}^{d} I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t, A_t, Y_{t,1:j-1}).$$

For each j,

$$I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t, A_t, Y_{t,1:j-1}) \le I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t, A_t).$$

Indeed,

$$\begin{aligned} &I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t, A_t, Y_{t,1:j-1}) \\ &= H(Y_{t,j} \mid \mathcal{H}_t, A_t, Y_{t,1:j-1}) - H(Y_{t,j} \mid \boldsymbol{\theta}, \mathcal{H}_t, A_t, Y_{t,1:j-1}). \end{aligned}$$

The first entropy is at most

$$H(Y_{t,j} \mid \mathcal{H}_t, A_t),$$

since conditioning reduces entropy. The second entropy equals

$$H(Y_{t,j} \mid \boldsymbol{\theta}, \mathcal{H}_t, A_t),$$

since the coordinates are conditionally independent given $(\boldsymbol{\theta}, A_t, \mathcal{H}_t)$. Hence,

$$I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t) \le \sum_{j=1}^{d} I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t, A_t).$$

Now fix $j \in [d]$ and condition on $\{A_t = a\}$. Because $Y_{t,j}$ depends only on the block $p_{a,j}$ and is conditionally independent of all other components of $\boldsymbol{\theta}$ given $(p_{a,j}, \mathcal{H}_t, A_t = a)$, we have

$$I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t, A_t = a) = I(p_{a,j}; Y_{t,j} \mid \mathcal{H}_t, A_t = a).$$

Moreover, since A_t is conditionally independent of $\boldsymbol{\theta}$ given $\mathcal{H}_t$, conditioning additionally on $A_t = a$ does not alter the posterior. Thus,

$$p_{a,j} \mid \mathcal{H}_t, A_t = a \sim \mathrm{Dir}(\beta_{a,j}(t)).$$

We now derive the one-step Dirichlet–categorical information bound. Let

$$p \sim \mathrm{Dir}(\beta),$$

where

$$\beta = (\beta_1, \dots, \beta_K), \qquad \beta_k \geq 1, \qquad \beta_0 := \sum_{k=1}^{K} \beta_k.$$

Given p, let

$$Y \mid p \sim \mathrm{Cat}(p).$$

The posterior predictive distribution of Y is categorical with probabilities

$$\bar{p}_k := \mathbb{E}[p_k] = \frac{\beta_k}{\beta_0}, \qquad k \in [K].$$

Therefore,

$$I(p; Y) = \mathbb{E}_p \left[\mathrm{KL} \left(\mathrm{Cat}(p) \,\|\, \mathrm{Cat}(\bar{p}) \right) \right] = \mathbb{E}_p \left[\sum_{k=1}^{K} p_k \log \frac{p_k}{\bar{p}_k} \right].$$

Expanding the expectation gives

$$I(p; Y) = \sum_{k=1}^{K} \mathbb{E}[p_k \log p_k] - \sum_{k=1}^{K} \mathbb{E}[p_k] \log \bar{p}_k.$$

For a Dirichlet random vector,

$$\mathbb{E}[p_k] = \frac{\beta_k}{\beta_0}$$

and

$$\mathbb{E}[p_k \log p_k] = \frac{\beta_k}{\beta_0} \left(\psi(\beta_k + 1) - \psi(\beta_0 + 1) \right),$$

where ψ is the digamma function. Hence,

$$I(p; Y) = \sum_{k=1}^{K} \frac{\beta_k}{\beta_0} \left(\psi(\beta_k + 1) - \psi(\beta_0 + 1) - \log \frac{\beta_k}{\beta_0} \right).$$

We now upper bound the term inside the parentheses. Using the standard digamma bounds

$$\psi(x) \leq \log x \qquad \text{and} \qquad \psi(x) \geq \log x - \frac{1}{x}, \qquad x \geq 1,$$

we get

$$\psi(\beta_k + 1) - \psi(\beta_0 + 1) \leq \log(\beta_k + 1) - \log(\beta_0 + 1) + \frac{1}{\beta_0 + 1}.$$

Therefore,

$$\begin{aligned} &\psi(\beta_k + 1) - \psi(\beta_0 + 1) - \log \frac{\beta_k}{\beta_0} \\ &\leq \log \frac{\beta_k + 1}{\beta_k} - \log \frac{\beta_0 + 1}{\beta_0} + \frac{1}{\beta_0 + 1} \\ &= \log \left(1 + \frac{1}{\beta_k} \right) - \log \left(1 + \frac{1}{\beta_0} \right) + \frac{1}{\beta_0 + 1}. \end{aligned}$$

Since

$$- \log \left(1 + \frac{1}{\beta_0} \right) \leq 0,$$

dropping this negative term gives

$$\psi(\beta_k + 1) - \psi(\beta_0 + 1) - \log \frac{\beta_k}{\beta_0} \leq \log \left(1 + \frac{1}{\beta_k} \right) + \frac{1}{\beta_0 + 1}.$$

Using

$$\log(1 + u) \leq u, \qquad u \geq 0,$$

we obtain

$$\psi(\beta_k + 1) - \psi(\beta_0 + 1) - \log \frac{\beta_k}{\beta_0} \leq \frac{1}{\beta_k} + \frac{1}{\beta_0 + 1}.$$

Substituting this into the exact expression for $I(p; Y)$, we get

$$I(p; Y) \leq \sum_{k=1}^{K} \frac{\beta_k}{\beta_0} \left(\frac{1}{\beta_k} + \frac{1}{\beta_0 + 1} \right).$$

Now,

$$\sum_{k=1}^{K} \frac{\beta_k}{\beta_0} \frac{1}{\beta_k} = \frac{K}{\beta_0},$$

and

$$\sum_{k=1}^{K} \frac{\beta_k}{\beta_0} \frac{1}{\beta_0 + 1} = \frac{1}{\beta_0 + 1}.$$

Therefore,

$$I(p; Y) \leq \frac{K}{\beta_0} + \frac{1}{\beta_0 + 1} \leq \frac{K + 1}{\beta_0}.$$

Thus, if

$$p \sim \mathrm{Dir}(\beta), \qquad \beta_k \geq 1, \qquad Y \mid p \sim \mathrm{Cat}(p),$$

then

$$I(p; Y) \leq \frac{K + 1}{\beta_0}.$$

Now fix a realized history $\mathcal{H}_t = h_t$. Applying the one-step bound conditionally on $(\mathcal{H}_t = h_t, A_t = a)$, we take

$$p = p_{a,j}, \qquad \beta = \beta_{a,j}(t).$$

Since

$$\beta_{a,j,k}(t) = \alpha_{a,j,k} + N_{a,j,k}(t - 1) \geq 1,$$

the assumptions of the one-step bound hold. Moreover,

$$\beta_{a,j,0}(t) = \alpha_{a,j,0} + N_a(t - 1).$$

Hence,

$$I(p_{a,j}; Y_{t,j} \mid \mathcal{H}_t = h_t, A_t = a) \leq \frac{K + 1}{\alpha_{a,j,0} + N_a(t - 1)}.$$

Since

$$\alpha_{a,j,0} = \sum_{k=1}^{K} \alpha_{a,j,k} \geq K \geq 1,$$

we obtain

$$I(p_{a,j}; Y_{t,j} \mid \mathcal{H}_t = h_t, A_t = a) \leq \frac{K + 1}{1 + N_a(t - 1)}.$$

We now remove the conditioning on the realized action. Averaging over the conditional distribution of A_t given this history,

$$\begin{aligned} &I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t = h_t, A_t) \\ &= \sum_{a \in \mathcal{A}} \Pr(A_t = a \mid \mathcal{H}_t = h_t) I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t = h_t, A_t = a). \end{aligned}$$

Therefore,

$$I(\boldsymbol{\theta}; Y_{t,j} \mid \mathcal{H}_t = h_t, A_t) \leq (K + 1) \sum_{a \in \mathcal{A}} \Pr(A_t = a \mid \mathcal{H}_t = h_t) \frac{1}{1 + N_a(t - 1)}.$$

Using the earlier coordinate-wise decomposition,

$$I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t = h_t, A_t) \le d(K+1) \sum_{a \in \mathcal{A}} \Pr(A_t = a \mid \mathcal{H}_t = h_t) \frac{1}{1 + N_a(t-1)}.$$

Finally, averaging over $\mathcal{H}_t$ gives

$$I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t) \le d(K+1) \sum_{a \in \mathcal{A}} \mathbb{E}\left[\Pr(A_t = a \mid \mathcal{H}_t) \frac{1}{1 + N_a(t-1)}\right].$$

Step 4: Summation over time via harmonic counts.

Combining the previous steps gives

$$\mathrm{IB}_T \le d(K+1) \sum_{t=1}^{T} \sum_{a \in \mathcal{A}} \mathbb{E}\left[\Pr(A_t = a \mid \mathcal{H}_t) \frac{1}{1 + N_a(t-1)}\right].$$

For fixed $a \in \mathcal{A}$ and t, define

$$f_t(\mathcal{H}_t) := \frac{1}{1 + N_a(t-1)}.$$

This quantity is $\mathcal{H}_t$-measurable because $N_a(t-1)$ depends only on the actions taken before round t, all of which are contained in $\mathcal{H}_t$. Hence, by the tower property,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\{A_t = a\} f_t(\mathcal{H}_t)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{A_t = a\} f_t(\mathcal{H}_t) \mid \mathcal{H}_t\right]\right] \\
&= \mathbb{E}\left[f_t(\mathcal{H}_t) \mathbb{E}\left[\mathbb{1}\{A_t = a\} \mid \mathcal{H}_t\right]\right] \\
&= \mathbb{E}\left[\Pr(A_t = a \mid \mathcal{H}_t) \frac{1}{1 + N_a(t-1)}\right].
\end{aligned}$$

Hence,

$$\mathrm{IB}_T \le d(K+1) \sum_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{A_t = a\} \frac{1}{1 + N_a(t-1)}\right].$$

It remains to bound the inner sum deterministically. Fix an action a. Let

$$\tau_1^a < \tau_2^a < \cdots < \tau_{N_a(T)}^a$$

be the times at which action a is played up to time T. On the m-th play of action a, that is, at time $t = \tau_m^a$, the action has already been played $m-1$ times, so

$$N_a(\tau_m^a - 1) = m - 1.$$

Therefore,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{1}\{A_t = a\} \frac{1}{1 + N_a(t-1)} &= \sum_{m=1}^{N_a(T)} \frac{1}{1 + N_a(\tau_m^a - 1)} \\
&= \sum_{m=1}^{N_a(T)} \frac{1}{m}.
\end{aligned}$$

If $N_a(T) = 0$, the sum is empty and equals zero. Otherwise,

$$\sum_{m=1}^{N_a(T)} \frac{1}{m} \le 1 + \log N_a(T) \le 1 + \log T.$$

Thus, for every action a,

$$\sum_{t=1}^{T} \mathbb{1}\{A_t = a\} \frac{1}{1 + N_a(t-1)} \le 1 + \log T.$$

Substituting this deterministic bound and summing over $a \in \mathcal{A}$ yields

$$\mathrm{IB}_T \le d(K+1)|\mathcal{A}|(1+\log T).$$

Hence,

$$\mathrm{IB}_T = O(dK|\mathcal{A}|\log T),$$

as claimed. □

Lemma A.5 (Logarithmic information growth for finite-action Gaussian reward models)**.** *Fix a finite action set $\mathcal{A}$ and reward dimension $d \in \mathbb{N}$. Consider the finite-action Gaussian reward model*

$$Y_t \mid (\boldsymbol{\theta}, A_t = a, \mathcal{H}_t) \sim \mathcal{N}(\mu_a, \Sigma), \qquad \boldsymbol{\theta} := (\{\mu_a\}_{a\in\mathcal{A}}, \Sigma),$$

where $\mu_a \in \mathbb{R}^d$ is an unknown action-specific mean and Σ is an unknown $d \times d$ symmetric positive definite shared covariance matrix. Suppose $\Sigma \sim \mathrm{IW}(\Psi_1, \nu_1)$, where Ψ_1 is a $d \times d$ symmetric positive definite matrix, and, conditional on Σ, the means are independent with

$$\mu_a \mid \Sigma \sim \mathcal{N}\left(m_{a,1}, \frac{1}{\kappa_{a,1}}\Sigma\right), \qquad a \in \mathcal{A},$$

where $m_{a,1} \in \mathbb{R}^d$ and $\kappa_{a,1} > 0$. Let $\mathcal{H}_t = (A_1, Y_1, \ldots, A_{t-1}, Y_{t-1})$, and assume the action-selection rule uses only $\mathcal{H}_t$ and fresh randomness independent of $\boldsymbol{\theta}$, so that $I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) = 0$ for each t. Then, for $\mathrm{IB}_T := I(\boldsymbol{\theta}; \mathcal{H}_{T+1})$,

$$\mathrm{IB}_T = O\big(d|\mathcal{A}|\log T + d^2 \log T\big).$$

In particular, for fixed d and $|\mathcal{A}|$, we have $\mathrm{IB}_T = O(\log T)$.

Proof. Before proving Lemma A.5, we record an auxiliary bound controlling the amount of information contained in one Gaussian observation about an inverse-Wishart covariance matrix. This lemma will be applied in Step 5, conditionally on the history $\mathcal{H}_t$ and the selected action $A_t = a$, with

$$Z = Y_t - m_{a,t}, \qquad c = 1 + \frac{1}{\kappa_{a,t}}, \qquad \Sigma \mid \mathcal{H}_t \sim \mathrm{IW}(\Psi_t, \nu_t).$$

Lemma A.6 (Covariance-information bound under inverse-Wishart mixing)**.** *Fix $d \in \mathbb{N}$, and define*

$$\widetilde{C}_d := \frac{d(3d+1)}{4}.$$

Suppose

$$\Sigma \sim \mathrm{IW}(\Psi, \nu), \qquad Z \mid \Sigma \sim \mathcal{N}(0, c\Sigma),$$

where Ψ is a $d \times d$ symmetric positive definite matrix, $c > 0$, and $\nu > d - 1$. Then $I(\Sigma; Z) < \infty$. Moreover, if $\nu > d + 3$, then

$$I(\Sigma; Z) \le \frac{\widetilde{C}_d}{\nu - d}.$$

Proof. Because mutual information is invariant under invertible deterministic transformations, it suffices to treat the canonical case

$$\Sigma \sim \mathrm{IW}(I_d, \nu), \qquad Z \mid \Sigma \sim \mathcal{N}(0, \Sigma).$$

Indeed, define

$$\widetilde{\Sigma} := \Psi^{-1/2}\Sigma\Psi^{-1/2}, \qquad \widetilde{Z} := c^{-1/2}\Psi^{-1/2}Z.$$

Then

$$I(\Sigma; Z) = I(\widetilde{\Sigma}; \widetilde{Z}),$$

while

$$\widetilde{\Sigma} \sim \mathrm{IW}(I_d, \nu), \qquad \widetilde{Z} \mid \widetilde{\Sigma} \sim \mathcal{N}(0, \widetilde{\Sigma}).$$

Hence assume from now on that

$$\Sigma \sim \mathrm{IW}(I_d, \nu), \qquad Z \mid \Sigma \sim \mathcal{N}(0, \Sigma).$$

Set

$$x := \frac{\nu - d + 1}{2}, \qquad \delta := \frac{d}{2}.$$

Since $\nu > d - 1$, we have $x > 0$, and

$$x + \delta = \frac{\nu + 1}{2}, \qquad \frac{\nu}{2} = x + \frac{d - 1}{2}.$$

We use

$$I(\Sigma; Z) = h(Z) - \mathbb{E}[h(Z \mid \Sigma)].$$

First,

$$h(Z \mid \Sigma = \sigma) = \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log |\sigma|.$$

Since

$$\Sigma^{-1} \sim \mathrm{Wishart}_d(\nu, I_d),$$

the standard Wishart identity gives

$$\mathbb{E}[\log |\Sigma|] = -\psi_d \left(\frac{\nu}{2} \right) - d \log 2,$$

where

$$\psi_d(a) := \sum_{i=1}^{d} \psi \left(a - \frac{i - 1}{2} \right).$$

Therefore,

$$\mathbb{E}[h(Z \mid \Sigma)] = \frac{d}{2} \log(\pi e) - \frac{1}{2} \psi_d \left(\frac{\nu}{2} \right).$$

Second, the marginal distribution of Z is multivariate Student-t with

$$r := \nu - d + 1 = 2x$$

degrees of freedom, location 0, and scale matrix I_d/r. The entropy of a d-variate Student-t distribution with r degrees of freedom and scale matrix Σ_t is

$$\begin{aligned} h(Z) = \log \Gamma \left(\frac{r}{2} \right) - \log \Gamma \left(\frac{r + d}{2} \right) + \frac{d}{2} \log(r\pi) \\ + \frac{1}{2} \log |\Sigma_t| + \frac{r + d}{2} \left[\psi \left(\frac{r + d}{2} \right) - \psi \left(\frac{r}{2} \right) \right]. \end{aligned}$$

Here,

$$\Sigma_t = \frac{I_d}{r},$$

so

$$\frac{d}{2} \log(r\pi) + \frac{1}{2} \log |I_d/r| = \frac{d}{2} \log \pi.$$

Thus,

$$h(Z) = \log \Gamma(x) - \log \Gamma(x + \delta) + \frac{d}{2} \log \pi + (x + \delta) \big(\psi(x + \delta) - \psi(x) \big).$$

Combining the two entropy formulas gives

$$\begin{aligned} I(\Sigma; Z) = \log \Gamma(x) - \log \Gamma(x + \delta) - \delta \\ + (x + \delta) \big(\psi(x + \delta) - \psi(x) \big) + \frac{1}{2} \sum_{j=0}^{d-1} \psi \left(x + \frac{j}{2} \right). \end{aligned}$$

Since $x > 0$, all Gamma and digamma terms in the preceding display are finite. Therefore,

$$I(\Sigma; Z) < \infty.$$

For later use, denote this finite quantity by

$$G_d(\nu) := \log\Gamma(x) - \log\Gamma(x+\delta) - \delta + (x+\delta)\big(\psi(x+\delta) - \psi(x)\big) + \frac{1}{2}\sum_{j=0}^{d-1}\psi\left(x + \frac{j}{2}\right), \qquad x = \frac{\nu - d + 1}{2}.$$

We now establish the sharper bound under the additional condition $\nu > d + 3$. Under this condition, $x > 2$. Adding and subtracting $\delta\psi(x)$, we obtain

$$\begin{aligned} I(\Sigma; Z) = &\Big[\log\Gamma(x) - \log\Gamma(x+\delta) + \delta\psi(x)\Big] - \delta \\ &+ (x+\delta)\big(\psi(x+\delta) - \psi(x)\big) \\ &+ \frac{1}{2}\sum_{j=0}^{d-1}\left[\psi\left(x + \frac{j}{2}\right) - \psi(x)\right]. \end{aligned}$$

By Taylor's theorem,

$$\log\Gamma(x+\delta) = \log\Gamma(x) + \delta\psi(x) + \frac{\delta^2}{2}\psi'(\xi_1)$$

for some

$$\xi_1 \in (x, x+\delta).$$

Since

$$\psi'(\xi_1) > 0,$$

we have

$$\log\Gamma(x) - \log\Gamma(x+\delta) + \delta\psi(x) = -\frac{\delta^2}{2}\psi'(\xi_1) \le 0.$$

Therefore,

$$\begin{aligned} I(\Sigma; Z) \le &-\delta + (x+\delta)\big(\psi(x+\delta) - \psi(x)\big) \\ &+ \frac{1}{2}\sum_{j=0}^{d-1}\left[\psi\left(x + \frac{j}{2}\right) - \psi(x)\right]. \end{aligned}$$

Since the trigamma function ψ' is positive and decreasing on $(0, \infty)$,

$$\psi(x+\delta) - \psi(x) \le \delta\psi'(x),$$

and, for every $j = 0, \ldots, d-1$,

$$\psi\left(x + \frac{j}{2}\right) - \psi(x) = \int_x^{x+j/2}\psi'(u)\,du \le \frac{j}{2}\psi'(x).$$

Hence,

$$I(\Sigma; Z) \le -\delta + (x+\delta)\delta\psi'(x) + \frac{1}{2}\sum_{j=0}^{d-1}\frac{j}{2}\psi'(x).$$

Since

$$\sum_{j=0}^{d-1} j = \frac{d(d-1)}{2},$$

this becomes

$$I(\Sigma; Z) \le -\delta + (x+\delta)\delta\psi'(x) + \frac{d(d-1)}{8}\psi'(x).$$

Using the standard trigamma bound

$$\psi'(x) \le \frac{1}{x - \frac{1}{2}}, \qquad x > \frac{1}{2},$$

we obtain

$$I(\Sigma; Z) \le -\delta + \frac{(x+\delta)\delta}{x - \frac{1}{2}} + \frac{d(d-1)}{8(x - \frac{1}{2})}.$$

Now,

$$-\delta + \frac{(x+\delta)\delta}{x-\frac{1}{2}} = \delta\left(\frac{x+\delta}{x-\frac{1}{2}} - 1\right) = \delta\frac{\delta+\frac{1}{2}}{x-\frac{1}{2}}.$$

Therefore,

$$I(\Sigma; Z) \le \frac{\delta(\delta+\frac{1}{2})}{x-\frac{1}{2}} + \frac{d(d-1)}{8(x-\frac{1}{2})}.$$

Since

$$\delta = \frac{d}{2},$$

we have

$$\delta\left(\delta + \frac{1}{2}\right) = \frac{d(d+1)}{4}.$$

Thus,

$$\begin{aligned} I(\Sigma; Z) &\le \frac{d(d+1)}{4(x-\frac{1}{2})} + \frac{d(d-1)}{8(x-\frac{1}{2})} \\ &= \frac{d(3d+1)}{8(x-\frac{1}{2})}. \end{aligned}$$

Finally,

$$x - \frac{1}{2} = \frac{\nu - d}{2},$$

and hence

$$I(\Sigma; Z) \le \frac{d(3d+1)}{4(\nu-d)} = \frac{\widetilde{C}_d}{\nu-d}.$$

This proves the lemma A.6. □

We now return to the proof of Lemma A.5. We first describe the posterior updates for the Gaussian reward model with unknown action-specific means and an unknown shared covariance matrix.

Step 1: Posterior update.

Because the policy selects each A_t using only $\mathcal{H}_t$ and fresh randomness independent of $\boldsymbol{\theta}$, the joint distribution factors as

$$p(\boldsymbol{\theta}, \mathcal{H}_{T+1}) = p(\boldsymbol{\theta}) \prod_{t=1}^{T} p(A_t \mid \mathcal{H}_t) p(Y_t \mid \boldsymbol{\theta}, A_t, \mathcal{H}_t).$$

Since

$$p(A_t \mid \mathcal{H}_t)$$

does not depend on $\boldsymbol{\theta}$, these policy factors do not affect the posterior kernel. Therefore, for a realized history, the observation likelihood can be regrouped according to the actions selected.

For each $a \in \mathcal{A}$, let

$$\mathcal{I}_a(T) := \{t \in [T] : A_t = a\}$$

be the set of times at which action a is played, and let

$$N_a(T) := |\mathcal{I}_a(T)|.$$

If $N_a(T) > 0$, define

$$\bar{Y}_a(T) := \frac{1}{N_a(T)} \sum_{t \in \mathcal{I}_a(T)} Y_t$$

and

$$W_a(T) := \sum_{t \in \mathcal{I}_a(T)} \big(Y_t - \bar{Y}_a(T)\big)\big(Y_t - \bar{Y}_a(T)\big)^\top.$$

If $N_a(T) = 0$, set

$$W_a(T) := 0,$$

and choose any value for $\bar{Y}_a(T)$; its value is immaterial because all terms involving it below are multiplied by $N_a(T)$.

For the realized observations associated with action a, the likelihood contribution is

$$L_a(\mu_a, \Sigma) = (2\pi)^{-dN_a(T)/2} |\Sigma|^{-N_a(T)/2} \exp\left(-\frac{1}{2} \sum_{t \in \mathcal{I}_a(T)} (Y_t - \mu_a)^\top \Sigma^{-1} (Y_t - \mu_a) \right).$$

Using the standard quadratic decomposition,

$$\begin{aligned} &\sum_{t \in \mathcal{I}_a(T)} (Y_t - \mu_a)(Y_t - \mu_a)^\top \\ &= W_a(T) + N_a(T) \big(\bar{Y}_a(T) - \mu_a\big) \big(\bar{Y}_a(T) - \mu_a\big)^\top, \end{aligned}$$

we may rewrite the likelihood as

$$\begin{aligned} L_a(\mu_a, \Sigma) \propto\ & |\Sigma|^{-N_a(T)/2} \\ & \times \exp\left(-\frac{1}{2} \operatorname{tr}\big(\Sigma^{-1} W_a(T)\big) - \frac{N_a(T)}{2} (\bar{Y}_a(T) - \mu_a)^\top \Sigma^{-1} (\bar{Y}_a(T) - \mu_a) \right). \end{aligned}$$

Now multiply this by the conditional Gaussian prior

$$\mu_a \mid \Sigma \sim \mathcal{N}\left(m_{a,1}, \frac{1}{\kappa_{a,1}} \Sigma \right),$$

whose density is

$$p(\mu_a \mid \Sigma) = (2\pi)^{-d/2} \kappa_{a,1}^{d/2} |\Sigma|^{-1/2} \exp\left(-\frac{\kappa_{a,1}}{2} (\mu_a - m_{a,1})^\top \Sigma^{-1} (\mu_a - m_{a,1}) \right).$$

Combining the quadratic terms in μ_a, we obtain

$$\begin{aligned} & N_a(T)(\bar{Y}_a(T) - \mu_a)^\top \Sigma^{-1} (\bar{Y}_a(T) - \mu_a) \\ & \quad + \kappa_{a,1} (\mu_a - m_{a,1})^\top \Sigma^{-1} (\mu_a - m_{a,1}) \\ & = \big(\kappa_{a,1} + N_a(T)\big) (\mu_a - m_{a,T+1})^\top \Sigma^{-1} (\mu_a - m_{a,T+1}) \\ & \quad + \frac{\kappa_{a,1} N_a(T)}{\kappa_{a,1} + N_a(T)} (\bar{Y}_a(T) - m_{a,1})^\top \Sigma^{-1} (\bar{Y}_a(T) - m_{a,1}), \end{aligned}$$

where

$$m_{a,T+1} = \frac{\kappa_{a,1} m_{a,1} + N_a(T) \bar{Y}_a(T)}{\kappa_{a,1} + N_a(T)}.$$

Hence,

$$\mu_a \mid (\Sigma, \mathcal{H}_{T+1}) \sim \mathcal{N}\left(m_{a,T+1}, \frac{1}{\kappa_{a,T+1}} \Sigma \right), \qquad \kappa_{a,T+1} := \kappa_{a,1} + N_a(T).$$

Integrating out μ_a, the Gaussian integral contributes the factor

$$(2\pi)^{d/2} \kappa_{a,T+1}^{-d/2} |\Sigma|^{1/2},$$

which cancels the additional

$$|\Sigma|^{-1/2}$$

arising from the Gaussian prior density. Therefore, after marginalizing μ_a, the contribution of action a to the kernel in Σ is

$$\begin{aligned} & |\Sigma|^{-N_a(T)/2} \\ & \quad \times \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Sigma^{-1} \left[W_a(T) + \frac{\kappa_{a,1} N_a(T)}{\kappa_{a,1} + N_a(T)} (\bar{Y}_a(T) - m_{a,1})(\bar{Y}_a(T) - m_{a,1})^\top \right] \right) \right). \end{aligned}$$

The prior on Σ is inverse-Wishart:

$$\Sigma \sim \mathrm{IW}(\Psi_1, \nu_1),$$

whose density is proportional to

$$|\Sigma|^{-(\nu_1+d+1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\Psi_1 \Sigma^{-1})\right).$$

Multiplying over all actions $a \in \mathcal{A}$, the posterior density in Σ is proportional to

$$|\Sigma|^{-(\nu_1+d+1+T)/2} \exp\left(-\frac{1}{2}\operatorname{tr}\big(\Psi_{T+1} \Sigma^{-1}\big)\right),$$

where

$$\nu_{T+1} := \nu_1 + T$$

and

$$\Psi_{T+1} = \Psi_1 + \sum_{a \in \mathcal{A}} \left[W_a(T) + \frac{\kappa_{a,1} N_a(T)}{\kappa_{a,1} + N_a(T)} \big(\bar{Y}_a(T) - m_{a,1}\big)\big(\bar{Y}_a(T) - m_{a,1}\big)^\top \right].$$

This is exactly the kernel of an inverse-Wishart distribution, so

$$\Sigma \mid \mathcal{H}_{T+1} \sim \mathrm{IW}(\Psi_{T+1}, \nu_{T+1}).$$

Applying the same update to the first $t-1$ observations gives

$$\Sigma \mid \mathcal{H}_t \sim \mathrm{IW}(\Psi_t, \nu_t), \qquad \nu_t = \nu_1 + t - 1,$$

and

$$\mu_a \mid (\mathcal{H}_t, \Sigma) \sim \mathcal{N}\left(m_{a,t}, \frac{1}{\kappa_{a,t}} \Sigma\right), \qquad \kappa_{a,t} = \kappa_{a,1} + N_a(t-1).$$

Step 2: Chain rule for the information budget.

By definition,

$$\mathrm{IB}_T = I(\boldsymbol{\theta}; \mathcal{H}_{T+1}), \qquad \mathcal{H}_{T+1} = (A_1, Y_1, \ldots, A_T, Y_T).$$

Applying the chain rule for mutual information,

$$I(\boldsymbol{\theta}; \mathcal{H}_{T+1}) = \sum_{t=1}^{T} I\big(\boldsymbol{\theta}; (A_t, Y_t) \mid \mathcal{H}_t\big).$$

For each t,

$$I\big(\boldsymbol{\theta}; (A_t, Y_t) \mid \mathcal{H}_t\big) = I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) + I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t).$$

By assumption,

$$I(\boldsymbol{\theta}; A_t \mid \mathcal{H}_t) = 0.$$

Therefore,

$$\mathrm{IB}_T = \sum_{t=1}^{T} I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t).$$

Step 3: Per-round decomposition into mean and covariance information.

For each round t, since

$$\boldsymbol{\theta} = (\mu, \Sigma), \qquad \mu := (\mu_a)_{a \in \mathcal{A}},$$

the chain rule gives

$$I(\boldsymbol{\theta}; Y_t \mid \mathcal{H}_t, A_t) = I(\Sigma; Y_t \mid \mathcal{H}_t, A_t) + I(\mu; Y_t \mid \mathcal{H}_t, A_t, \Sigma).$$

Summing over t and using Step 2,

$$\mathrm{IB}_T = \sum_{t=1}^{T} I(\mu; Y_t \mid \mathcal{H}_t, A_t, \Sigma) + \sum_{t=1}^{T} I(\Sigma; Y_t \mid \mathcal{H}_t, A_t).$$

Step 4: Bound the mean-information term.

Fix a round t and condition on

$$(\mathcal{H}_t = h_t, A_t = a, \Sigma = \sigma).$$

By conjugacy,

$$\mu_a \mid (\mathcal{H}_t = h_t, \Sigma = \sigma) \sim \mathcal{N}\left(m_{a,t}, \frac{1}{\kappa_{a,t}}\sigma\right),$$

where

$$\kappa_{a,t} = \kappa_{a,1} + N_a(t-1).$$

Since A_t is conditionally independent of $\boldsymbol{\theta}$ given $\mathcal{H}_t$, conditioning additionally on $A_t = a$ does not change this posterior.

Moreover,

$$Y_t \mid (\mu_a, \mathcal{H}_t = h_t, A_t = a, \Sigma = \sigma) \sim \mathcal{N}(\mu_a, \sigma).$$

Conditional on

$$(\mathcal{H}_t = h_t, A_t = a, \Sigma = \sigma),$$

only μ_a enters the observation model, so

$$\mu_{-a} := (\mu_b)_{b \neq a}$$

drops out by conditional independence. Therefore,

$$I(\mu; Y_t \mid \mathcal{H}_t = h_t, A_t = a, \Sigma = \sigma) = I(\mu_a; Y_t \mid \mathcal{H}_t = h_t, A_t = a, \Sigma = \sigma).$$

Thus, conditional on these quantities, this is a Gaussian linear channel with prior covariance

$$\frac{\sigma}{\kappa_{a,t}}$$

and noise covariance σ. Therefore,

$$\begin{aligned} &I(\mu; Y_t \mid \mathcal{H}_t = h_t, A_t = a, \Sigma = \sigma) \\ &= \frac{1}{2}\log\det\left(I_d + \frac{1}{\kappa_{a,t}} I_d\right) \\ &= \frac{d}{2}\log\left(1 + \frac{1}{\kappa_{a,t}}\right). \end{aligned}$$

Averaging over

$$(\mathcal{H}_t, A_t, \Sigma)$$

gives

$$I(\mu; Y_t \mid \mathcal{H}_t, A_t, \Sigma) = \frac{d}{2}\mathbb{E}\left[\log\left(1 + \frac{1}{\kappa_{A_t,t}}\right)\right].$$

Summing over t,

$$\sum_{t=1}^{T} I(\mu; Y_t \mid \mathcal{H}_t, A_t, \Sigma) = \frac{d}{2}\mathbb{E}\left[\sum_{t=1}^{T}\log\left(1 + \frac{1}{\kappa_{A_t,t}}\right)\right].$$

Now group the sum by action. For a fixed action a, every time action a is played the value of $\kappa_{a,t}$ runs through

$$\kappa_{a,1}, \quad \kappa_{a,1}+1, \quad \ldots, \quad \kappa_{a,1} + N_a(T) - 1.$$

Therefore, pathwise,

$$\begin{aligned} \sum_{t:\, A_t = a} \log\left(1 + \frac{1}{\kappa_{A_t,t}}\right) &= \sum_{j=0}^{N_a(T)-1} \log\left(1 + \frac{1}{\kappa_{a,1}+j}\right) \\ &= \sum_{j=0}^{N_a(T)-1} \log\left(\frac{\kappa_{a,1}+j+1}{\kappa_{a,1}+j}\right) \\ &= \log\left(\frac{\kappa_{a,1} + N_a(T)}{\kappa_{a,1}}\right) \\ &= \log\left(1 + \frac{N_a(T)}{\kappa_{a,1}}\right). \end{aligned}$$

Summing over actions yields the pathwise identity

$$\sum_{t=1}^{T} \log\left(1+\frac{1}{\kappa_{A_t,t}}\right) = \sum_{a\in\mathcal{A}} \log\left(1+\frac{N_a(T)}{\kappa_{a,1}}\right).$$

Taking expectation gives

$$\sum_{t=1}^{T} I(\mu; Y_t \mid \mathcal{H}_t, A_t, \Sigma) = \frac{d}{2}\mathbb{E}\left[\sum_{a\in\mathcal{A}} \log\left(1+\frac{N_a(T)}{\kappa_{a,1}}\right)\right].$$

Step 5: Bound the covariance-information term.

Fix t, a realized history $\mathcal{H}_t = h_t$, and an action $a \in \mathcal{A}$. By Step 1 applied at horizon $t-1$,

$$\Sigma \mid \mathcal{H}_t = h_t \sim \mathrm{IW}(\Psi_t, \nu_t), \qquad \nu_t = \nu_1 + t - 1.$$

Because A_t is conditionally independent of $\boldsymbol{\theta}$ given $\mathcal{H}_t$, conditioning additionally on $A_t = a$ does not change this posterior:

$$\Sigma \mid (\mathcal{H}_t = h_t, A_t = a) \sim \mathrm{IW}(\Psi_t, \nu_t).$$

Also, by conjugacy,

$$\mu_a \mid (\mathcal{H}_t = h_t, \Sigma) \sim \mathcal{N}\left(m_{a,t}, \frac{1}{\kappa_{a,t}}\Sigma\right).$$

Since

$$Y_t \mid (\mu_a, \mathcal{H}_t = h_t, A_t = a, \Sigma) \sim \mathcal{N}(\mu_a, \Sigma),$$

integrating out μ_a gives

$$Y_t \mid (\mathcal{H}_t = h_t, A_t = a, \Sigma) \sim \mathcal{N}\left(m_{a,t}, \left(1+\frac{1}{\kappa_{a,t}}\right)\Sigma\right).$$

Because $m_{a,t}$ is fixed conditional on $\mathcal{H}_t = h_t$, translation invariance of mutual information yields

$$I(\Sigma; Y_t \mid \mathcal{H}_t = h_t, A_t = a) = I\left(\Sigma; Y_t - m_{a,t} \mid \mathcal{H}_t = h_t, A_t = a\right).$$

Conditional on $(\mathcal{H}_t = h_t, A_t = a)$, the pair therefore has the form

$$\Sigma \sim \mathrm{IW}(\Psi_t, \nu_t), \qquad Z \mid \Sigma \sim \mathcal{N}(0, c_t\Sigma), \qquad c_t := 1 + \frac{1}{\kappa_{a,t}}.$$

Because $\mathrm{IW}(\Psi_1, \nu_1)$ is a proper inverse-Wishart prior, $\nu_1 > d-1$. Define

$$t_0 := \min\left\{t \geq 1 : \nu_1 + t - 1 > d + 3\right\}.$$

Then t_0 is finite (indeed, $t_0 \leq 5$). For $t < t_0$, the first part of Lemma A.6 gives, pointwise in (h_t, a),

$$I(\Sigma; Y_t \mid \mathcal{H}_t = h_t, A_t = a) = G_d(\nu_t) < \infty.$$

For $t \geq t_0$, the second part of Lemma A.6 gives

$$I(\Sigma; Y_t \mid \mathcal{H}_t = h_t, A_t = a) \leq \frac{\widetilde{C}_d}{\nu_t - d}.$$

Both bounds are independent of the realized history h_t and action a. Hence, defining the finite prior-dependent constant

$$C_{\mathrm{early}} := \sum_{t=1}^{t_0-1} G_d(\nu_t),$$

we obtain, after averaging over $(\mathcal{H}_t, A_t)$ and summing over time,

$$\sum_{t=1}^{T} I(\Sigma; Y_t \mid \mathcal{H}_t, A_t) \leq C_{\mathrm{early}} + \widetilde{C}_d \sum_{t=t_0}^{T} \frac{1}{\nu_1 + t - d - 1},$$

where the second sum is understood to be zero when $T < t_0$.

Let

$$b_0 := \nu_1 + t_0 - d - 1 > 3.$$

If $T \geq t_0$, then

$$\begin{aligned}\sum_{t=t_0}^{T} \frac{1}{\nu_1 + t - d - 1} &= \sum_{s=0}^{T-t_0} \frac{1}{b_0 + s} \\ &\leq \frac{1}{b_0} + \int_{b_0}^{b_0+T-t_0} \frac{dx}{x} \\ &= \frac{1}{b_0} + \log\left(1 + \frac{T - t_0}{b_0}\right) \\ &\leq \frac{1}{b_0} + \log(1 + T).\end{aligned}$$

Thus, for every $T \geq 1$,

$$\sum_{t=1}^{T} I(\Sigma; Y_t \mid \mathcal{H}_t, A_t) \leq C_{\text{early}} + \widetilde{C}_d \left(\frac{1}{b_0} + \log(1 + T)\right).$$

Step 6: Combine the bounds and derive the final rate.

Combining Steps 3–5 gives

$$\begin{aligned}\mathrm{IB}_T \leq \frac{d}{2}\mathbb{E}\left[\sum_{a\in\mathcal{A}} \log\left(1 + \frac{N_a(T)}{\kappa_{a,1}}\right)\right] \\ + C_{\text{early}} + \widetilde{C}_d \left(\frac{1}{b_0} + \log(1 + T)\right).\end{aligned}$$

We now bound the mean contribution. For every $a \in \mathcal{A}$,

$$0 \leq N_a(T) \leq T.$$

Since $\kappa_{a,1} > 0$ is fixed by the prior,

$$\log\left(1 + \frac{N_a(T)}{\kappa_{a,1}}\right) \leq \log\left(1 + \frac{T}{\kappa_{a,1}}\right).$$

Moreover,

$$\begin{aligned}\log\left(1 + \frac{T}{\kappa_{a,1}}\right) &\leq \log\left((1 + T)\left(1 + \frac{1}{\kappa_{a,1}}\right)\right) \\ &= \log(1 + T) + \log\left(1 + \frac{1}{\kappa_{a,1}}\right).\end{aligned}$$

Therefore,

$$\sum_{a\in\mathcal{A}} \log\left(1 + \frac{N_a(T)}{\kappa_{a,1}}\right) \leq |\mathcal{A}| \log(1 + T) + \sum_{a\in\mathcal{A}} \log\left(1 + \frac{1}{\kappa_{a,1}}\right).$$

Taking expectations preserves the same bound. Hence,

$$\begin{aligned}\mathrm{IB}_T \leq \frac{d}{2}|\mathcal{A}| \log(1 + T) + \frac{d}{2} \sum_{a\in\mathcal{A}} \log\left(1 + \frac{1}{\kappa_{a,1}}\right) \\ + \widetilde{C}_d \log(1 + T) + C_{\text{early}} + \frac{\widetilde{C}_d}{b_0}.\end{aligned}$$

Equivalently,

$$\mathrm{IB}_T \le \left(\frac{d}{2}|\mathcal{A}| + \widetilde{C}_d\right)\log(1+T) + C_{\mathrm{prior}},$$

where

$$C_{\mathrm{prior}} := \frac{d}{2}\sum_{a\in\mathcal{A}}\log\left(1+\frac{1}{\kappa_{a,1}}\right) + C_{\mathrm{early}} + \frac{\widetilde{C}_d}{b_0}.$$

The quantity C_{prior} is finite and depends only on the fixed prior hyperparameters and model dimension, not on T. Finally, since

$$\widetilde{C}_d = \frac{d(3d+1)}{4} = O(d^2),$$

we obtain

$$\mathrm{IB}_T = O\left(d|\mathcal{A}|\log(1+T) + d^2\log(1+T)\right).$$

Since

$$\log(1+T) = O(\log T),$$

this is equivalently

$$\mathrm{IB}_T = O\left(d|\mathcal{A}|\log T + d^2\log T\right),$$

as claimed. □

A.17 Gaussian Rewards, Clipping, and Tail Localization

Theorems 4.1 and 4.2 are stated under a bounded-reward assumption. We now show how their gap-based regret guarantees extend, through clipping, to Gaussian and more generally uniformly sub-Gaussian reward models. The argument has three components. First, clipping produces a bounded observation model to which Theorems 4.1 and 4.2 apply directly. Second, because the clipped observations are deterministic functions of the original observations, the data-processing inequality upper bounds the information budget of the clipped process by that of the corresponding untruncated experiment. Third, the Pareto gaps induced by the clipped and original mean rewards differ by a uniformly controlled clipping-bias term.

Assume that there exist finite constants $\sigma > 0$ and $\mu_{\max} > 0$ such that, for every $\boldsymbol{\theta} \in \mathrm{supp}(p_1)$, $a \in \mathcal{A}$, and $j \in [d]$,

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\exp\big(\lambda\big(Y_{t,a,j} - \mu_{\boldsymbol{\theta},j}(a)\big)\big)\right] \le \exp\left(\frac{\sigma^2\lambda^2}{2}\right) \qquad \text{for all } \lambda \in \mathbb{R},$$

and

$$|\mu_{\boldsymbol{\theta},j}(a)| \le \mu_{\max}.$$

Thus, every reward coordinate is uniformly σ-sub-Gaussian around a uniformly bounded mean. In particular, for Gaussian rewards the sub-Gaussian condition holds whenever the corresponding conditional coordinate variances are uniformly bounded by σ^2.

For $B > \mu_{\max}$, define the coordinate-wise clipping map

$$c_B(x) := \max\{-B, \min\{x, B\}\},$$

and let

$$\widetilde{Y}_{t,a,j} := c_B(Y_{t,a,j}), \qquad \widetilde{\mu}_{\boldsymbol{\theta},j}(a) := \mathbb{E}_{\boldsymbol{\theta}}\left[\widetilde{Y}_{t,a,j}\right].$$

The clipping bias is uniformly controlled. Indeed,

$$
\begin{aligned}
|\mu_{\boldsymbol{\theta},j}(a) - \widetilde{\mu}_{\boldsymbol{\theta},j}(a)| &= |\mathbb{E}_{\boldsymbol{\theta}}\left[Y_{t,a,j} - c_B(Y_{t,a,j})\right]| \\
&\le \mathbb{E}_{\boldsymbol{\theta}}\left[|Y_{t,a,j} - c_B(Y_{t,a,j})|\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\left(|Y_{t,a,j}| - B\right)_+\right] \\
&\le \mathbb{E}_{\boldsymbol{\theta}}\left[\left(|Y_{t,a,j} - \mu_{\boldsymbol{\theta},j}(a)| - (B - \mu_{\max})\right)_+\right] \\
&= \int_{B-\mu_{\max}}^{\infty} \mathbb{P}_{\boldsymbol{\theta}}\left(|Y_{t,a,j} - \mu_{\boldsymbol{\theta},j}(a)| > u\right)\, du \\
&\le 2\int_{B-\mu_{\max}}^{\infty} \exp\!\left(-\frac{u^2}{2\sigma^2}\right) du \\
&\le \frac{2\sigma^2}{B-\mu_{\max}} \exp\!\left(-\frac{(B-\mu_{\max})^2}{2\sigma^2}\right).
\end{aligned}
$$

Define

$$
\varepsilon_B := \frac{2\sigma^2}{B-\mu_{\max}} \exp\!\left(-\frac{(B-\mu_{\max})^2}{2\sigma^2}\right).
$$

Then

$$
|\mu_{\boldsymbol{\theta},j}(a) - \widetilde{\mu}_{\boldsymbol{\theta},j}(a)| \le \varepsilon_B
$$

uniformly over $\boldsymbol{\theta} \in \mathrm{supp}(p_1)$, $a \in \mathcal{A}$, and $j \in [d]$.

Define the Pareto gap associated with the clipped mean rewards by

$$
\widetilde{\Delta}_{\boldsymbol{\theta},B}(a) := \left(\max_{b\in\mathcal{A}} \min_{j\in[d]} \left[\widetilde{\mu}_{\boldsymbol{\theta},j}(b) - \widetilde{\mu}_{\boldsymbol{\theta},j}(a)\right]\right)_+.
$$

For every $a, b \in \mathcal{A}$ and $j \in [d]$,

$$
\begin{aligned}
&|(\mu_{\boldsymbol{\theta},j}(b) - \mu_{\boldsymbol{\theta},j}(a)) - (\widetilde{\mu}_{\boldsymbol{\theta},j}(b) - \widetilde{\mu}_{\boldsymbol{\theta},j}(a))| \\
&\quad \le |\mu_{\boldsymbol{\theta},j}(b) - \widetilde{\mu}_{\boldsymbol{\theta},j}(b)| + |\mu_{\boldsymbol{\theta},j}(a) - \widetilde{\mu}_{\boldsymbol{\theta},j}(a)| \\
&\quad \le 2\varepsilon_B.
\end{aligned}
$$

Since the operations $\min$, $\max$, and $x \mapsto x_+$ are 1-Lipschitz with respect to the sup norm,

$$
\left|\Delta_{\boldsymbol{\theta}}(a) - \widetilde{\Delta}_{\boldsymbol{\theta},B}(a)\right| \le 2\varepsilon_B \qquad \text{for every } \boldsymbol{\theta} \in \mathrm{supp}(p_1) \text{ and } a \in \mathcal{A}.
$$

We now consider clipped versions of Algorithms 3 and 4. At round t, the environment generates the original reward vector $\mathbf{Y}_{t,A_t}$, while the learner observes only

$$
\widetilde{\mathbf{Y}}_{t,A_t} = c_B(\mathbf{Y}_{t,A_t}),
$$

where clipping is applied coordinate-wise. The learner updates its posterior using the induced clipped observation law and computes the corresponding Thompson-sampling or IDS quantities using the clipped mean vectors

$$
\widetilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(a) = \mathbb{E}_{\boldsymbol{\theta}}\left[\widetilde{\mathbf{Y}}_{t,a}\right].
$$

Equivalently, the clipped observation law is the pushforward of the original observation law under coordinate-wise clipping.

Every coordinate of $\widetilde{\mathbf{Y}}_{t,a}$ lies in $[-B, B]$. Adding the same constant B to every coordinate produces rewards in $[0, 2B]$. This common translation leaves all pairwise mean differences, Pareto-dominance relations, Pareto-optimal sets, and Pareto gaps unchanged. Hence the clipped model satisfies the bounded-reward assumptions of Theorems 4.1 and 4.2.

Let

$$\widetilde{\mathcal{S}}_B^*(\boldsymbol{\theta})$$

denote the Pareto-optimal set induced by the clipped mean rewards, and define

$$M_B := \sup_{\boldsymbol{\theta} \in \mathrm{supp}(p_1)} \left| \widetilde{\mathcal{S}}_B^*(\boldsymbol{\theta}) \right| \leq |\mathcal{A}|.$$

Also define the clipped interaction history

$$\widetilde{\mathcal{H}}_{T+1} := \left(A_1, \widetilde{\mathbf{Y}}_{1,A_1}, \ldots, A_T, \widetilde{\mathbf{Y}}_{T,A_T} \right)$$

and its parameter-information budget

$$\widetilde{IB}_T := I\left(\boldsymbol{\theta}; \widetilde{\mathcal{H}}_{T+1}\right).$$

Tracking explicitly the dependence on the reward range in the proofs of Theorems 4.1 and 4.2, and using the reward range $2B$, gives for either clipped Thompson Sampling or clipped IDS

$$\mathbb{E}\left[\sum_{t=1}^{T} \widetilde{\Delta}_{\boldsymbol{\theta},B}(A_t)\right] = O\left(B\sqrt{|\mathcal{A}|\, M_B\, \widetilde{IB}_T\, T}\right).$$

We next relate the information budget of the clipped process to that of the corresponding untruncated experiment. Couple the two processes by first generating $\mathbf{Y}_{t,A_t}$ and then revealing only $c_B(\mathbf{Y}_{t,A_t})$ to the clipped learner. Define the corresponding untruncated history along the same action sequence by

$$\mathcal{H}_{T+1}^{Y} := (A_1, \mathbf{Y}_{1,A_1}, \ldots, A_T, \mathbf{Y}_{T,A_T}).$$

Because the clipped learner selects A_t using only the past clipped history and fresh internal randomness, and because the clipped history is a deterministic function of the corresponding untruncated history, the clipped policy is also a valid adaptive policy for the original observation model.

Moreover, for a deterministic coordinate-wise clipping map f_B,

$$\widetilde{\mathcal{H}}_{T+1} = f_B\left(\mathcal{H}_{T+1}^{Y}\right).$$

Consequently,

$$\boldsymbol{\theta} \longrightarrow \mathcal{H}_{T+1}^{Y} \longrightarrow \widetilde{\mathcal{H}}_{T+1}$$

forms a Markov chain. By the data-processing inequality,

$$\widetilde{IB}_T = I\left(\boldsymbol{\theta}; \widetilde{\mathcal{H}}_{T+1}\right) \leq I\left(\boldsymbol{\theta}; \mathcal{H}_{T+1}^{Y}\right).$$

Thus, clipping cannot increase the amount of information that the interaction reveals about the unknown parameter.

In particular, whenever the original Gaussian experiment satisfies the information-budget bound established in Proposition 4.1,

$$I\big(\boldsymbol{\theta}; \mathcal{H}_{T+1}^{Y}\big) = O\big(d|\mathcal{A}| \log T + d^2 \log T\big)\,,$$

the clipped process inherits the same asymptotic upper bound:

$$\widetilde{IB}_T = O\big(d|\mathcal{A}| \log T + d^2 \log T\big)\,.$$

The adaptive-policy condition required by the Gaussian information-budget calculation is preserved here: the clipped learner uses only the clipped past and fresh internal randomness, and the clipped past is itself a deterministic function of the original past.

Substituting this information-budget bound into the bounded-reward regret guarantee yields

$$\mathbb{E}\left[\sum_{t=1}^{T} \widetilde{\Delta}_{\boldsymbol{\theta},B}(A_t)\right] = O\Big(B\sqrt{|\mathcal{A}|\, M_B\, (d|\mathcal{A}| + d^2)\, T \log T}\Big)\,.$$

It remains to translate regret in the clipped reward geometry into regret measured using the original mean rewards. From the uniform gap approximation,

$$\Delta_{\boldsymbol{\theta}}(A_t) \le \widetilde{\Delta}_{\boldsymbol{\theta},B}(A_t) + 2\varepsilon_B.$$

Therefore, defining

$$R_T := \sum_{t=1}^{T} \Delta_{\boldsymbol{\theta}}(A_t),$$

we obtain

$$\begin{aligned}\mathbb{E}[R_T] &\le \mathbb{E}\left[\sum_{t=1}^{T} \widetilde{\Delta}_{\boldsymbol{\theta},B}(A_t)\right] + 2T\varepsilon_B \\ &= O\Big(B\sqrt{|\mathcal{A}|\, M_B\, (d|\mathcal{A}| + d^2)\, T \log T}\Big) + 2T\varepsilon_B.\end{aligned}$$

Choose

$$B_T := \mu_{\max} + 2\sigma\sqrt{\log(2T)}.$$

Then

$$\frac{(B_T - \mu_{\max})^2}{2\sigma^2} = 2\log(2T),$$

so

$$\exp\left(-\frac{(B_T - \mu_{\max})^2}{2\sigma^2}\right) = (2T)^{-2}.$$

Consequently,

$$\begin{aligned}T\varepsilon_{B_T} &= T\frac{2\sigma^2}{B_T - \mu_{\max}}(2T)^{-2} \\ &= \frac{\sigma}{4T\sqrt{\log(2T)}} = o(1),\end{aligned}$$

and therefore

$$2T\varepsilon_{B_T} = o(1).$$

Moreover,

$$B_T = O\Big(\mu_{\max} + \sigma\sqrt{\log T}\Big).$$

Substituting $B = B_T$ gives

$$\mathbb{E}[R_T] = O\Big((\mu_{\max} + \sigma\sqrt{\log T})\sqrt{|\mathcal{A}|\, M_{B_T}\,(d|\mathcal{A}| + d^2)\, T \log T}\Big) + o(1).$$

Since

$$M_{B_T} \le |\mathcal{A}|,$$

for fixed d, $|\mathcal{A}|$, $\mu_{\max}$, and σ,

$$\mathbb{E}[R_T] = \widetilde{O}(\sqrt{T}).$$

More generally, suppose that a uniformly sub-Gaussian reward model with uniformly bounded means satisfies

$$I\big(\boldsymbol{\theta}; \mathcal{H}_{T+1}^{Y}\big) \le \mathcal{I}_T$$

for every adaptive policy of the form considered above. The same data-processing argument gives

$$\widetilde{IB}_T \le \mathcal{I}_T,$$

and hence

$$\mathbb{E}[R_T] = O\Big((\mu_{\max} + \sigma\sqrt{\log T})\sqrt{|\mathcal{A}|\, M_{B_T}\, \mathcal{I}_T\, T}\Big) + o(1).$$

Thus, whenever

$$\mathcal{I}_T = O(\log T)$$

for fixed model dimension,

$$\mathbb{E}[R_T] = \widetilde{O}(\sqrt{T}).$$

The two ingredients in this extension are logically separate. Proposition 4.1 supplies a logarithmic parameter-information budget for the finite-action Gaussian experiment considered there, while the clipping transfer requires the uniform mean and sub-Gaussian tail conditions imposed at the beginning of this subsection. Whenever both ingredients hold, clipping converts the original unbounded observation model into a bounded model to which Theorems 4.1 and 4.2 apply, the data-processing inequality transfers the logarithmic information-budget bound to the clipped process, and the discrepancy between the clipped and original Pareto gaps contributes only the $o(1)$ clipping-bias term above.

A.18 Computation of the IDS Information Ratios

We describe how the IDS quantities used for Pareto-violation and gap-based regret are computed in the experiments. The two formulations share the same information-ratio optimization, but differ in both the one-step loss and the information target. We use $N_{\mathrm{MC}} = 100$ posterior samples per round for quantities that depend nonlinearly on the unknown Pareto geometry. For gap-based IDS, the parameter-level information gain can be estimated using a general posterior-sampling procedure; in the Beta–Bernoulli model and in a Gaussian model with known observation covariance, it also admits useful closed-form expressions.

Pareto-violation IDS. For Algorithm 2, the one-step loss and information gain are

$$\Delta_t^{\mathrm{viol}}(a) := 1 - \mathbb{P}(a \in \mathcal{S}^* \mid \mathcal{H}_t)\,, \qquad I_t^{\mathrm{viol}}(a) := I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)\,.$$

At round t, we draw

$$\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(N_{\mathrm{MC}})} \overset{\text{i.i.d.}}{\sim} p_t(\boldsymbol{\theta}),$$

and compute the sampled Pareto sets

$$S^{(m)} := \mathcal{S}^*\left(\boldsymbol{\theta}^{(m)}\right).$$

The posterior violation probability is approximated by

$$\widehat{\Delta}_t^{\mathrm{viol}}(a) = 1 - \frac{1}{N_{\mathrm{MC}}} \sum_{m=1}^{N_{\mathrm{MC}}} \mathbb{1}\{a \in S^{(m)}\}.$$

For the Bernoulli experiments, let $\mathcal{Y} = \{0,1\}^d$. Conditional on posterior sample $\boldsymbol{\theta}^{(m)}$, the probability of $\mathbf{y} \in \mathcal{Y}$ under action a is

$$p_m^a(\mathbf{y}) := \prod_{j=1}^{d} \left(\theta_{a,j}^{(m)}\right)^{y_j} \left(1 - \theta_{a,j}^{(m)}\right)^{1-y_j}.$$

Let $\mathcal{U}_t$ denote the distinct Pareto sets among $S^{(1)}, \ldots, S^{(N_{\mathrm{MC}})}$, and define

$$\widehat{p}_t(s) = \frac{1}{N_{\mathrm{MC}}} \sum_{m=1}^{N_{\mathrm{MC}}} \mathbb{1}\{S^{(m)} = s\},$$

$$\widehat{p}_t^a(\mathbf{y} \mid s) = \frac{\sum_{m: S^{(m)} = s} p_m^a(\mathbf{y})}{\sum_{m=1}^{N_{\mathrm{MC}}} \mathbb{1}\{S^{(m)} = s\}},$$

and

$$\widehat{p}_t^a(\mathbf{y}) = \frac{1}{N_{\mathrm{MC}}} \sum_{m=1}^{N_{\mathrm{MC}}} p_m^a(\mathbf{y}).$$

The Pareto-set information gain is approximated by

$$\widehat{I}_t^{\mathrm{viol}}(a) = \sum_{s \in \mathcal{U}_t} \widehat{p}_t(s) \sum_{\mathbf{y} \in \mathcal{Y}} \widehat{p}_t^a(\mathbf{y} \mid s) \log \frac{\widehat{p}_t^a(\mathbf{y} \mid s)}{\widehat{p}_t^a(\mathbf{y})}.$$

Thus, although the conditional Bernoulli reward probabilities are available explicitly, the information gain about $\mathcal{S}^*$ is evaluated through posterior samples. This is necessary because the Pareto set is a nonlinear function of the complete reward geometry, so Beta–Bernoulli conjugacy does not in general yield a comparable closed-form expression for $I(\mathcal{S}^*; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)$.

The same posterior-sampling construction extends to continuous observation models. For Gaussian rewards, let $p_m^a(\mathbf{y})$ denote the Gaussian observation density under posterior sample $\boldsymbol{\theta}^{(m)}$, and define

$$\widehat{p}_t^a(\mathbf{y}) = \frac{1}{N_{\mathrm{MC}}} \sum_{m=1}^{N_{\mathrm{MC}}} p_m^a(\mathbf{y})$$

and, for each $s \in \mathcal{U}_t$,

$$\widehat{p}_t^{\,a}(\mathbf{y} \mid s) = \frac{\sum_{m:S^{(m)}=s} p_m^a(\mathbf{y})}{N_s}, \qquad N_s := \sum_{m=1}^{N_{\mathrm{MC}}} \mathbb{1}\{S^{(m)} = s\}.$$

The remaining continuous expectation may be approximated by drawing

$$\mathbf{Y}_{t,a}^{(m,r)} \sim p_m^a, \qquad r = 1, \dots, R,$$

which gives the plug-in Monte Carlo approximation

$$\widehat{I}_t^{\mathrm{viol}}(a) = \sum_{s \in \mathcal{U}_t} \widehat{p}_t(s) \frac{1}{N_s R} \sum_{m:S^{(m)}=s} \sum_{r=1}^{R} \log \frac{\widehat{p}_t^{\,a}(\mathbf{Y}_{t,a}^{(m,r)} \mid s)}{\widehat{p}_t^{\,a}(\mathbf{Y}_{t,a}^{(m,r)})}.$$

Hence Pareto-violation IDS can be implemented through posterior sampling for both discrete and continuous reward models while retaining $\mathcal{S}^*$ as the information target.

Gap-based IDS. For Algorithm 4, the corresponding quantities are

$$\Delta_t^{\mathrm{gap}}(a) := \mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t], \qquad I_t^{\mathrm{gap}}(a) := I(\boldsymbol{\theta}; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a).$$

The posterior expected gap is nonlinear in the complete collection of mean reward vectors. In particular,

$$\mathbb{E}[\Delta_{\boldsymbol{\theta}}(a) \mid \mathcal{H}_t] \neq \Delta_{\mathbb{E}[\boldsymbol{\theta} \mid \mathcal{H}_t]}(a)$$

in general. Using the same $N_{\mathrm{MC}} = 100$ posterior samples, we compute

$$\Delta_{\boldsymbol{\theta}^{(m)}}(a) = \left(\max_{b \in \mathcal{A}} \min_{j \in [d]} \left[\mu_{\boldsymbol{\theta}^{(m)},j}(b) - \mu_{\boldsymbol{\theta}^{(m)},j}(a) \right] \right)_+$$

and approximate

$$\widehat{\Delta}_t^{\mathrm{gap}}(a) = \frac{1}{N_{\mathrm{MC}}} \sum_{m=1}^{N_{\mathrm{MC}}} \Delta_{\boldsymbol{\theta}^{(m)}}(a).$$

General Monte Carlo parameter-information estimator. The parameter-level information gain can also be approximated directly from posterior samples without requiring a model-specific closed form. For a fixed action a, let

$$p_m^a(\mathbf{y}) := p\Big(\mathbf{y} \mid A_t = a, \boldsymbol{\theta}^{(m)}\Big), \qquad m = 1, \dots, N_{\mathrm{MC}},$$

and approximate the posterior predictive law by

$$\widehat{p}_t^{\,a}(\mathbf{y}) = \frac{1}{N_{\mathrm{MC}}} \sum_{\ell=1}^{N_{\mathrm{MC}}} p_\ell^a(\mathbf{y}).$$

Using

$$I_t^{\mathrm{gap}}(a) = \mathbb{E}_{\boldsymbol{\theta} \mid \mathcal{H}_t}[D_{\mathrm{KL}}(p(\mathbf{Y}_{t,a} \mid \boldsymbol{\theta}, A_t = a) \,\|\, p(\mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a))],$$

we obtain the plug-in posterior-sampling approximation

$$\widehat{I}_{t,\mathrm{MC}}^{\mathrm{gap}}(a) = \frac{1}{N_{\mathrm{MC}}} \sum_{m=1}^{N_{\mathrm{MC}}} D_{\mathrm{KL}}(p_m^a \,\|\, \widehat{p}_t^{\,a}).$$

For discrete Bernoulli reward vectors, $\mathcal{Y} = \{0,1\}^d$, the KL divergence can be evaluated explicitly:

$$\widehat{I}_{t,\mathrm{MC}}^{\mathrm{gap}}(a) = \frac{1}{N_{\mathrm{MC}}} \sum_{m=1}^{N_{\mathrm{MC}}} \sum_{\mathbf{y} \in \mathcal{Y}} p_m^a(\mathbf{y}) \log \frac{p_m^a(\mathbf{y})}{\widehat{p}_t^{\,a}(\mathbf{y})}.$$

Thus no sampling of the Bernoulli observation itself is required: posterior parameters are sampled, while the sum over all 2^d possible reward vectors is evaluated explicitly.

For continuous observation models, the KL expectation may instead be approximated using an additional observation-level Monte Carlo step. For each posterior parameter sample, draw

$$\mathbf{Y}_{t,a}^{(m,r)} \sim p_m^a, \qquad r = 1, \dots, R,$$

and compute

$$\widehat{I}_{t,\mathrm{MC}}^{\mathrm{gap}}(a) = \frac{1}{N_{\mathrm{MC}} R} \sum_{m=1}^{N_{\mathrm{MC}}} \sum_{r=1}^{R} \log \frac{p_m^a(\mathbf{Y}_{t,a}^{(m,r)})}{\widehat{p}_t^a(\mathbf{Y}_{t,a}^{(m,r)})}.$$

This gives a general implementation of gap-based IDS even when no closed-form parameter-information expression is available.

Closed-form parameter information for Beta–Bernoulli rewards. For Bernoulli rewards, suppose the posterior factorizes across action–objective pairs:

$$Y_{t,a,j} \mid \theta_{a,j} \sim \mathrm{Bernoulli}(\theta_{a,j}), \qquad \theta_{a,j} \mid \mathcal{H}_t \sim \mathrm{Beta}(\alpha_{t,a,j}, \beta_{t,a,j})\,.$$

Conditional on $A_t = a$, the observation depends on the unknown parameter only through

$$\boldsymbol{\theta}_a = (\theta_{a,1}, \dots, \theta_{a,d}),$$

and, by factorization across objectives,

$$I_t^{\mathrm{gap}}(a) = \sum_{j=1}^{d} I(\theta_{a,j}; Y_{t,a,j} \mid \mathcal{H}_t, A_t = a)\,.$$

Define

$$\bar{\theta}_{t,a,j} = \frac{\alpha_{t,a,j}}{\alpha_{t,a,j} + \beta_{t,a,j}}, \qquad h(p) = -p \log p - (1-p)\log(1-p).$$

The posterior predictive distribution is

$$Y_{t,a,j} \mid \mathcal{H}_t, A_t = a \sim \mathrm{Bernoulli}(\bar{\theta}_{t,a,j}),$$

so

$$I(\theta_{a,j}; Y_{t,a,j} \mid \mathcal{H}_t, A_t = a) = h(\bar{\theta}_{t,a,j}) - \mathbb{E}[h(\theta_{a,j}) \mid \mathcal{H}_t]\,.$$

For

$$\theta \sim \mathrm{Beta}(\alpha, \beta),$$

the standard Beta identities give

$$\mathbb{E}[\theta \log \theta] = \frac{\alpha}{\alpha+\beta}\left[\psi(\alpha+1) - \psi(\alpha+\beta+1)\right],$$

and

$$\mathbb{E}[(1-\theta)\log(1-\theta)] = \frac{\beta}{\alpha+\beta}\left[\psi(\beta+1) - \psi(\alpha+\beta+1)\right],$$

where $\psi(\cdot)$ denotes the digamma function. Hence

$$\begin{aligned}\mathbb{E}[h(\theta_{a,j}) \mid \mathcal{H}_t] = &-\frac{\alpha_{t,a,j}}{\alpha_{t,a,j}+\beta_{t,a,j}}\left[\psi(\alpha_{t,a,j}+1) - \psi(\alpha_{t,a,j}+\beta_{t,a,j}+1)\right] \\ &-\frac{\beta_{t,a,j}}{\alpha_{t,a,j}+\beta_{t,a,j}}\left[\psi(\beta_{t,a,j}+1) - \psi(\alpha_{t,a,j}+\beta_{t,a,j}+1)\right],\end{aligned}$$

and therefore

$$I_t^{\mathrm{gap}}(a) = \sum_{j=1}^{d}\left[h\left(\frac{\alpha_{t,a,j}}{\alpha_{t,a,j}+\beta_{t,a,j}}\right) - \mathbb{E}[h(\theta_{a,j}) \mid \mathcal{H}_t]\right].$$

Thus, in the Beta–Bernoulli model the gap-based parameter-information term also admits the exact expression above. In our implementation we use the general posterior-sampling estimator, while the closed form provides an equivalent model-specific alternative that avoids Monte Carlo approximation of the information term. The posterior expected Pareto gap itself is still estimated from posterior samples.

Closed-form parameter information for a known-covariance Gaussian special case. For completeness, suppose the Gaussian observation covariance is known:

$$\mathbf{Y}_{t,a} \mid \boldsymbol{\mu}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \Sigma_a), \qquad \boldsymbol{\mu}_a \mid \mathcal{H}_t \sim \mathcal{N}(\mathbf{m}_{t,a}, V_{t,a}),$$

where $\Sigma_a \succ 0$. Integrating out $\boldsymbol{\mu}_a$ gives

$$\mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a \sim \mathcal{N}(\mathbf{m}_{t,a}, \Sigma_a + V_{t,a}).$$

Since the observation under action a depends on the unknown parameter only through $\boldsymbol{\mu}_a$,

$$I_t^{\mathrm{gap}}(a) = I(\boldsymbol{\mu}_a; \mathbf{Y}_{t,a} \mid \mathcal{H}_t, A_t = a)\,.$$

Using

$$I(X;Y) = H(Y) - H(Y \mid X)$$

and the entropy formula for a multivariate Gaussian,

$$\begin{aligned} I_t^{\mathrm{gap}}(a) &= \frac{1}{2} \log \frac{\det(\Sigma_a + V_{t,a})}{\det(\Sigma_a)} \\ &= \frac{1}{2} \log\det\left(I_d + \Sigma_a^{-1/2} V_{t,a} \Sigma_a^{-1/2}\right) \\ &= \frac{1}{2} \log\det\left(I_d + \Sigma_a^{-1} V_{t,a}\right), \end{aligned}$$

where the last equality follows from $\det(I + AB) = \det(I + BA)$.

If the reward coordinates are conditionally independent, with

$$Y_{t,a,j} \mid \mu_{a,j} \sim \mathcal{N}(\mu_{a,j}, \sigma_{a,j}^2), \qquad \mu_{a,j} \mid \mathcal{H}_t \sim \mathcal{N}(m_{t,a,j}, \tau_{t,a,j}^2),$$

then

$$V_{t,a} = \mathrm{diag}(\tau_{t,a,1}^2, \ldots, \tau_{t,a,d}^2), \qquad \Sigma_a = \mathrm{diag}(\sigma_{a,1}^2, \ldots, \sigma_{a,d}^2),$$

and hence

$$I_t^{\mathrm{gap}}(a) = \frac{1}{2} \sum_{j=1}^{d} \log\left(1 + \frac{\tau_{t,a,j}^2}{\sigma_{a,j}^2}\right).$$

Under the Gaussian prior

$$\mu_{a,j} \sim \mathcal{N}(m_{0,a,j}, \tau_{0,a,j}^2),$$

if action a has been selected $N_{t,a}$ times before round t, then conjugate Gaussian updating gives

$$\tau_{t,a,j}^2 = \left(\frac{1}{\tau_{0,a,j}^2} + \frac{N_{t,a}}{\sigma_{a,j}^2}\right)^{-1}.$$

Thus, when the observation covariance is known, the Gaussian parameter-information term can be evaluated exactly from the current Gaussian posterior. For Gaussian models with unknown observation variance, including the more general setting considered in the additional Gaussian experiments, we instead use the posterior-sampling construction above; the known-covariance expression is included as a closed-form special case.

Optimization of the empirical IDS ratios. For either regret criterion, let $\widehat{\Delta}_t(a)$ and $\widehat{I}_t(a)$ denote the corresponding estimated or closed-form quantities above. For $\pi \in \Pi(\mathcal{A})$, define

$$\widehat{\Delta}_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a) \widehat{\Delta}_t(a), \qquad \widehat{I}_t(\pi) = \sum_{a \in \mathcal{A}} \pi(a) \widehat{I}_t(a),$$

and

$$\widehat{\psi}_t(\pi) = \frac{\widehat{\Delta}_t(\pi)^2}{\widehat{I}_t(\pi)}.$$

We use the extended-value convention

$$\frac{0}{0} = 0, \qquad \frac{x}{0} = +\infty \quad \text{for } x > 0.$$

As in standard IDS implementations, an optimizer can be chosen with support on at most two actions. We therefore search over all one- and two-action mixtures. For each pair $i, j \in \mathcal{A}$, let

$$\pi_q = qe_i + (1-q)e_j, \qquad q \in [0,1],$$

and write

$$\delta_i = \widehat{\Delta}_t(i), \qquad \delta_j = \widehat{\Delta}_t(j), \qquad \iota_i = \widehat{I}_t(i), \qquad \iota_j = \widehat{I}_t(j).$$

The corresponding one-dimensional objective is

$$f_{ij}(q) = \frac{(q\delta_i + (1-q)\delta_j)^2}{q\iota_i + (1-q)\iota_j}.$$

Writing

$$d_\delta = \delta_i - \delta_j, \qquad d_\iota = \iota_i - \iota_j,$$

we have, whenever the denominator is positive,

$$f_{ij}(q) = \frac{(\delta_j + qd_\delta)^2}{\iota_j + qd_\iota}.$$

Differentiating gives

$$f'_{ij}(q) = \frac{(\delta_j + qd_\delta)\,[2d_\delta(\iota_j + qd_\iota) - d_\iota(\delta_j + qd_\delta)]}{(\iota_j + qd_\iota)^2}.$$

Apart from the degenerate case in which the numerator gap term vanishes, an interior stationary point therefore satisfies

$$2d_\delta(\iota_j + qd_\iota) = d_\iota(\delta_j + qd_\delta).$$

Whenever

$$d_\delta d_\iota \neq 0,$$

the corresponding candidate is

$$q^\star_{ij} = \frac{d_\iota\delta_j - 2d_\delta\iota_j}{d_\delta d_\iota} = \frac{(\iota_i - \iota_j)\delta_j - 2(\delta_i - \delta_j)\iota_j}{(\delta_i - \delta_j)(\iota_i - \iota_j)}.$$

For each pair, we evaluate f_{ij} at the endpoints

$$q = 0, \qquad q = 1,$$

and, whenever

$$q^\star_{ij} \in (0,1)$$

and

$$q^\star_{ij}\iota_i + (1 - q^\star_{ij})\iota_j > 0,$$

also at $q = q^\star_{ij}$. If $d_\delta = 0$ or $d_\iota = 0$, no additional nondegenerate interior candidate is required; an optimum is attained at an endpoint unless the objective is constant over the pair. Cases with zero information denominator are handled by the extended-value convention above.

Finally, the pair and mixing weight attaining the smallest empirical information ratio define

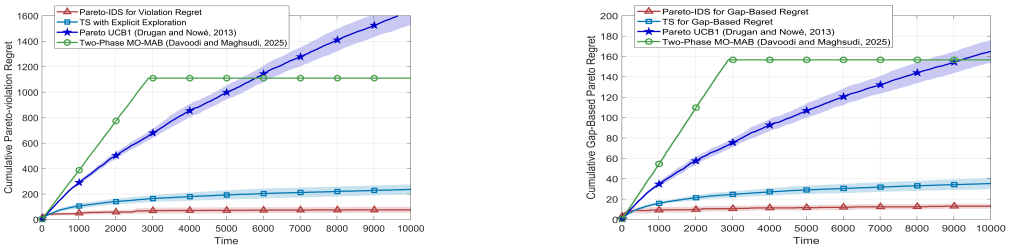
$$\widehat{\pi}^{\mathrm{IDS}}_t,$$

and the action A_t is sampled according to this distribution.

A.19 Additional Gaussian Reward Experiments

In addition to the Bernoulli experiments in Section 5, we evaluate the proposed methods on a continuous-reward Gaussian model. This experiment serves as a robustness check showing that the qualitative behavior of the algorithms is not specific to Bernoulli observations.

We use the same experimental setup as in the main paper, but replace Bernoulli rewards with Gaussian rewards. Specifically, we consider $|\mathcal{A}| = 10$ actions and $d = 4$ objectives. For each action $a \in \mathcal{A}$ and objective $j \in [d]$, the mean reward $\mu_j(a)$ is sampled independently from $\mathrm{Unif}(0.05, 0.95)$, and the



(a) Pareto-violation regret

(b) Gap-based Pareto regret

Figure 3: Additional Gaussian reward experiments with $|\mathcal{A}| = 10$ actions and $d = 4$ objectives. For each action and objective, the mean is sampled from $\mathrm{Unif}(0.05, 0.95)$ and the variance from $\mathrm{Unif}(1, 4)$. Rewards are generated independently across actions and objectives.

variance $\sigma_j^2(a)$ is sampled independently from $\mathrm{Unif}(1, 4)$. Rewards are then generated independently across actions and objectives according to

$$Y_{t,j}(a) \sim \mathcal{N}(\mu_j(a), \sigma_j^2(a)).$$

Thus, both the mean and variance are action- and objective-specific, and both are unknown to the learner. The learner maintains the corresponding Gaussian posterior over the unknown reward parameters and uses this posterior in the TS and IDS computations.

We evaluate both regret criteria considered in the paper. For Pareto-violation regret, the loss at round t is $\mathbb{1}\{A_t \notin \mathcal{S}^*\}$, where $\mathcal{S}^*$ is the Pareto set induced by the true mean reward vectors. For gap-based Pareto regret, we use the same Pareto-gap definition as in the main experiments,

$$\Delta(a) = \left(\sup_{b \in \mathcal{A}} \min_{j \in [d]} \big(\mu_j(b) - \mu_j(a)\big) \right)_+ .$$

We compare the Gaussian versions of TS-Pareto and Pareto-IDS against the same baselines as in Section 5. For each method, we report the mean cumulative regret with shaded bands corresponding to one standard deviation across independent trials.

The results are shown in Figure 3. The qualitative ordering is consistent with the Bernoulli experiments: Pareto-IDS achieves the lowest cumulative regret, followed by TS-Pareto, while Pareto-UCB1 and the two-phase MO-MAB baseline incur larger regret. This supports the robustness of the main empirical message beyond Bernoulli rewards. In particular, IDS remains effective because it directly trades off posterior expected loss against information gain, while TS-Pareto benefits from posterior sampling but is less targeted than IDS.

A.20 Experimental Compute Resources

All experiments were run on a server equipped with an AMD Ryzen Threadripper PRO 5995WX CPU, which has 64 cores and 128 threads, a base clock speed of 2.7 GHz, a maximum boost clock speed of 4.5 GHz, and 128 GB of RAM.