

# Cost-Aware Exploration for Structured Bandits

Xinyu Liu<sup>1</sup>, Wei You<sup>1</sup>, and Chao Qin<sup>2</sup>

<sup>1</sup>Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, [xliufn@connect.ust.hk](mailto:xliufn@connect.ust.hk), [weiyou@ust.hk](mailto:weiyou@ust.hk)

<sup>2</sup>Stanford Graduate School of Business, Stanford University, [chaoqin@stanford.edu](mailto:chaoqin@stanford.edu)

February 4, 2026

## Abstract

We study pure exploration with heterogeneous per-measurement costs. An agent sequentially selects among  $K$  arms whose observation laws belong to (possibly structured) canonical exponential families, and each pull incurs an arm- and instance-dependent cost. The goal is to identify an instance-dependent answer (e.g., best arm, thresholding bandits, or Pareto set identification) while maximizing the rate at which the posterior probability of error decays *per unit spent budget*. We characterize the optimal cost-normalized posterior error exponent as the value of a maximin program that trades off statistical discrimination against the average cost, and show that no adaptive sampling rule can exceed this exponent. By working with the cost normalization, the exponent is characterized by a concave maximization problem. Motivated by this characterization, we develop a cost-aware pure exploration algorithm. The resulting dynamics is inherently nonsmooth and set-valued due to argmin operations, boundary effects, and normalization. We analyze the stochastic iterates through a continuous-time approximation based on differential inclusions and prove that, under mild regularity conditions, the algorithm attains the optimal cost-normalized posterior error exponent almost surely. Our results provide a general asymptotic optimality guarantee for cost-aware pure exploration beyond best-arm identification, covering broad exploration queries, exponential-family rewards, and structured bandit models, with substantially lower per-iteration computational overhead than optimize-then-track baselines such as Track-and-Stop.

## 1 Introduction

Pure exploration seeks to identify an instance-dependent answer (e.g., a best arm, a top- $m$  set, or the arms above a threshold) from sequential noisy measurements. Classical formulations measure efficiency in the *number of samples*, effectively treating each pull as unit cost. In many applications, however, the binding constraint is a heterogeneous *budget* (time, money, compute, or human effort), and arms can consume that budget at vastly different rates. Such cost heterogeneity arises naturally in clinical screening (tests with different prices and turnaround times), online experimentation

(traffic and operational costs), sensing and robotics (energy), multi-fidelity simulation (cheap proxies vs. expensive high fidelity), and data-centric workflows such as prompt selection or human labeling (token/annotation costs). Information-theoretically, this shifts the objective from information per sample to *information per unit cost*. As a result, a sample-optimal policy can be budget-inefficient, over-investing in expensive measurements even when cheaper ones provide comparable discriminative power. The mismatch is particularly pronounced in structured bandits such as linear bandits, where geometry couples arms: a cheaper arm can be *more* informative because it probes directions that best separate the true instance from its hardest alternatives.

A growing literature incorporates sampling costs into pure exploration. Qin and Russo (2024) bridges regret minimization and pure exploration in a cost-aware setting for unstructured bandits. Related directions include multi-fidelity feedback (Poiani et al. 2024), LLM prompt selection (Hu et al. 2025), and other cost-sensitive designs (Kanarios et al. 2024, He-Yueya et al. 2025). Several works also pursue asymptotic optimality: Qin and Russo (2024) combine IDS with best-arm identification; Kanarios et al. (2024), Wu et al. (2025) adapt Track-and-Stop; and Poiani et al. (2024) extends Ménard (2019) and analyze subgradient-type schemes for multi-fidelity trade-offs.

In this work, we extend the algorithm of Qin and You (2025) to incorporate heterogeneous costs and general structured bandit models. Doing so is technically nontrivial for three reasons. First, pure-exploration designs rely on repeated  $\arg \min$  operations (e.g., to detect the most confusing alternative), which become set-valued under ties; existing analyses often address this via carefully constructed subdifferential subspaces (Wang et al. 2021). Second, structured bandits frequently admit sparse optimal allocations, pushing the dynamics toward the simplex boundary where Chernoff information can be nonsmooth and Clarke-type generalized gradients are required. Third, the information-directed selection (IDS) rule (You et al. 2023) we build upon involves a normalization whose denominator may vanish, creating genuine singularities. To handle these pathologies and establish convergence, we develop a continuous-time approximation based on *differential inclusions*. This framework naturally accommodates set-valued detection/selection rules and boundary-induced nonsmoothness, enabling a Lyapunov-based analysis of the resulting learning dynamics.

**Contributions.** We now summarize our contributions and compare with relevant literature.

First, we propose a simple cost-aware pure-exploration algorithm by extending Qin and You (2025) to cost-sensitive and structured settings. Many asymptotically optimal methods follow an “optimize-then-track” blueprint (Kanarios et al. 2024, Wu et al. 2025), solving a maximin program and tracking its solution online. Track-and-Stop (Garivier and Kaufmann 2016) is the canonical example, but its per-round allocation solve can be costly beyond small unstructured instances. Related approaches reduce computation via online mirror ascent (Ménard 2019) or game-theoretic no-regret dynamics (Degenne et al. 2019, 2020). Frank–Wolfe Sampling (Wang et al. 2021) offers a broadly applicable conditional-gradient tracker, but still requires a per-round linear subproblem and tuning of an  $r$ -subdifferential characterization. In contrast, our algorithm reduces per-iteration overhead and completely remove tracking steps by directly mapping the current state to sampling probabilities through the Chernoff-information structure.

Second, we provide the first asymptotic optimality guarantee for IDS beyond Gaussian best-arm identification. Our results cover general pure-exploration queries under one-parameter exponential-family rewards and structured bandit models, addressing the open questions raised in Qin and You (2025), Jourdan (2024). Top-two algorithms, for instance, sample via a  $\beta$ -biased randomization between a leader and challenger (Russo 2020, Qin et al. 2017, Shang et al. 2020, Jourdan et al. 2022). Despite strong empirical performance, optimality can require nontrivial tuning of  $\beta$  (Russo 2020). IDS (You et al. 2023) replaces the fixed- $\beta$  rule with an information-gain criterion and yields asymptotic optimality for Gaussian BAI, while Qin and Russo (2024) proposes IDS for cost-aware BAI; until this work, guarantees beyond the Gaussian setting remained limited.

Third, we develop a proof technique tailored to the nonsmooth, set-valued nature of pure-exploration dynamics, offering a powerful toolkit for analyzing bandit algorithms that behave like nonsmooth first-order methods. We analyze the IDS updates through differential inclusions, leveraging nonsmooth chain rules (path differentiability) (Davis et al. 2020, Bolte and Pauwels 2021) and weak Sard-type properties (Benaïm et al. 2005) to construct Lyapunov arguments. Our analysis follows the stochastic-approximation perspective, which links discrete-time iterates to continuous-time limits via differential inclusions (Benaïm et al. 2005, Borkar 2008, Davis et al. 2020, Bolte and Pauwels 2021). While related tools are standard in nonsmooth optimization, they are not designed for the distinctive singularities induced by bandit sampling and IDS.

## 2 Problem Formulation

We study pure exploration with heterogeneous per-measurement costs. There are  $K$  arms, indexed by  $[K] \triangleq \{1, \dots, K\}$ . The unknown state of nature is represented by a parameter vector  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$ . Sampling arm  $i$  produces a noisy measurement whose distribution we denote by  $P_{\boldsymbol{\theta},i}$ . Measurements are assumed independent across rounds, and conditionally identically distributed given the selected arm and  $\boldsymbol{\theta}$ . We assume that for each arm  $i$ , the observation law belongs to a one-dimensional canonical exponential family. Specifically, the law  $P_{\boldsymbol{\theta},i}$  has density  $p_{\boldsymbol{\theta},i}(y) = b(y) \exp(\eta_i(\boldsymbol{\theta})T(y) - A(\eta_i(\boldsymbol{\theta})))$ , where  $\eta_i(\cdot)$  is the arm-specific natural-parameter map that allows for structured bandits.

**Example 1** (Best-arm identification in linear bandits). *Consider a linear bandit with unknown parameter  $\boldsymbol{\theta} \in \mathbb{R}^d$  and design matrix  $X \in \mathbb{R}^{K \times d}$ , whose  $i$ -th row is the feature vector  $\mathbf{x}_i^\top$ . The mean reward of arm  $i$  is  $m_i \triangleq \mathbf{x}_i^\top \boldsymbol{\theta}$ . Pulling arm  $i$  yields an observation  $Y \sim \mathcal{N}(m_i, \sigma_i^2)$ , and observations are independent across pulls (and across arms). A canonical objective is best-arm identification: determine the (assumed unique) optimal arm  $I^* \in \arg \max_{i \in [K]} m_i$ .*

**Assumption 1.** *The parameter set  $\Theta \subset \mathbb{R}^d$  is compact with nonempty interior, each  $\eta_i$  is continuous on  $\Theta$ , and  $A$  is twice continuously differentiable on an open neighborhood of  $\eta_i(\Theta)$  for each  $i$ . Moreover, there is a finite constant  $L_A < \infty$  such that  $\sup_{i \in [K]} \sup_{\boldsymbol{\theta} \in \Theta} |A'(\eta_i(\boldsymbol{\theta}))| \leq L_A$ .*

At each round  $n \geq 1$ , an  $\mathcal{H}_{n-1}$ -measurable sampling rule selects an arm  $I_n \in [K]$  and observes  $Y_n \sim P_{\boldsymbol{\theta},I_n}$ , where  $\mathcal{H}_n \triangleq \sigma(I_1, Y_1, \dots, I_n, Y_n)$  is the interaction history up to round  $n$ .

To capture cost heterogeneity across arms, we assume that sampling arm  $i$  under the true instance  $\boldsymbol{\theta}$  incurs a (per-sample) cost  $C_i(\boldsymbol{\theta})$ .

**Assumption 2.** *The sampling costs are known continuous functions. Furthermore, there exist constants  $0 < c_{\min} \leq c_{\max} < \infty$  such that for all arms  $i$  and all  $\boldsymbol{\theta} \in \Theta$ ,  $c_{\min} \leq C_i(\boldsymbol{\theta}) \leq c_{\max}$ .*

Let  $\Delta_K \triangleq \{\mathbf{q} \in \mathbb{R}_{\geq 0}^K : \sum_{i=1}^K q_i = 1\}$  denote the simplex in  $\mathbb{R}^K$ . Define the sample allocation  $\mathbf{p}_n \triangleq (p_{n,1}, \dots, p_{n,K}) \in \Delta_K$ , where  $p_{n,i} \triangleq N_{n,i}/n$  and  $N_{n,i} \triangleq \sum_{\ell=1}^n \mathbf{1}\{I_{\ell} = i\}$ . For any  $\mathbf{p} \in \Delta_K$ , denote  $\bar{C}_{\boldsymbol{\theta}}(\mathbf{p}) = \sum_{i=1}^K p_i C_i(\boldsymbol{\theta})$  as the average cost per sample. Then, the budget spent after  $n$  rounds is  $B_n \triangleq \sum_{i=1}^K N_{n,i} C_i(\boldsymbol{\theta}) = n \bar{C}_{\boldsymbol{\theta}}(\mathbf{p}_n)$ .

**Posterior Error Exponent Per Unit Cost.** The goal is to correctly identify an instance-dependent answer  $\mathcal{I}(\boldsymbol{\theta})$  based on these noisy measurements. For a given instance  $\boldsymbol{\theta}$ , define the *alternative set* as the collection of instances yielding a different answer,

$$\text{Alt}(\boldsymbol{\theta}) \triangleq \{\boldsymbol{\vartheta} \in \Theta : \mathcal{I}(\boldsymbol{\vartheta}) \neq \mathcal{I}(\boldsymbol{\theta})\},$$

as in Garivier and Kaufmann (2016), Wang et al. (2021), Qin and You (2025).

**Assumption 3.** *The true instance  $\boldsymbol{\theta} \in \Theta$  has a unique answer  $\mathcal{I}(\boldsymbol{\theta})$ . Moreover, the answer is identifiable from the family of observation laws in the sense that for any  $\boldsymbol{\vartheta} \in \Theta$ ,*

$$P_{\boldsymbol{\theta},i} = P_{\boldsymbol{\vartheta},i} \quad \forall i \in [K] \implies \mathcal{I}(\boldsymbol{\vartheta}) = \mathcal{I}(\boldsymbol{\theta}).$$

Furthermore,  $\text{Alt}(\boldsymbol{\theta})$  is non-empty and  $\text{dist}(\boldsymbol{\theta}, \text{Alt}(\boldsymbol{\theta})) > 0$ .

The form of the alternative set depends on the query to be answered. This assumption is automatically satisfied if  $\text{Alt}(\boldsymbol{\theta})$  is open, e.g., in best-arm identification. Moreover, it also covers cases where  $\text{Alt}(\boldsymbol{\theta})$  is not necessarily open, e.g., in thresholding bandit problem (Chen et al. 2014, Locatelli et al. 2016), but we require that the true parameter is not on the boundary.

Let  $\Pi_0$  be a prior on  $\Theta$ , and let  $\Pi_n(\cdot) \triangleq \Pi(\cdot \mid \mathcal{H}_n)$  denote the posterior distribution after  $n$  rounds. Following Russo (2020), our goal is to design a sampling rule that drives rapid decay of the posterior mass  $\Pi_n(\text{Alt}(\boldsymbol{\theta}))$ —the posterior probability of identifying an incorrect answer—while accounting for heterogeneous sampling costs across arms.

As discussed in Russo (2020), Qin et al. (2017), under a well-designed sampling rule the posterior mass  $\Pi_n(\text{Alt}(\boldsymbol{\theta}))$  typically decays exponentially in the *number of total samples*. With cost considerations, we instead normalize by the cumulative budget spent  $B_n$  and study the (cost-normalized) posterior error exponent  $-\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta}))$ . Accordingly, our objective is to characterize and attain the optimal cost-aware large-deviation rate of posterior concentration.

**Cost-normalized discriminative information rate.** The large-deviation decay of  $\Pi_n(\text{Alt}(\boldsymbol{\theta}))$  under adaptive sampling is characterized by Russo (2020, Proposition 5) for the unstructured bandit in the non-cost-aware setting. Extension to our setting here is relatively straightforward, as we

detail in Appendix A. In particular, we show that, up to subexponential factors, the posterior mass of any open set  $\tilde{\Theta} \subset \Theta$  decays exponentially at rate  $\inf_{\vartheta \in \tilde{\Theta}} \Gamma(\mathbf{p}_n; \vartheta)$ , where  $\Gamma(\mathbf{p}_n; \vartheta) \triangleq \sum_{i=1}^K p_i \text{KL}(P_{\theta,i} \| P_{\vartheta,i})$  is the discrimination information rate under allocation  $\mathbf{p} \in \Delta_K$  and alternative instance  $\vartheta \in \Theta$ . In the cost-aware setting, we normalize by the cumulative budget spent  $B_n$  and obtain the heuristic approximation

$$-\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) = \frac{n}{B_n} \left( -\frac{1}{n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) \right) \approx \inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \frac{\Gamma(\mathbf{p}_n; \vartheta)}{\bar{C}_{\boldsymbol{\theta}}(\mathbf{p}_n)}.$$

See Theorem 1 and Proposition 3 for rigorous statements.

Motivated by this posterior-exponent characterization, we define a cost-normalized discrimination rate for a fixed sample allocation  $\mathbf{p} \in \Delta_K$  and any alternative instance  $\vartheta \in \Theta$

$$\Gamma^c(\mathbf{p}; \vartheta) \triangleq \frac{\sum_{i=1}^K p_i \text{KL}(P_{\theta,i} \| P_{\vartheta,i})}{\sum_{i=1}^K p_i C_i(\boldsymbol{\theta})} = \frac{\Gamma(\mathbf{p}; \vartheta)}{\bar{C}_{\boldsymbol{\theta}}(\mathbf{p})}. \quad (1)$$

This quantity measures the expected amount of discriminative information *per unit cost* against the alternative  $\vartheta$  when samples are allocated according to  $\mathbf{p}$ . A natural design criterion is to select an allocation that maximizes the discrimination rate against the worst-case alternative, leading to

$$\Gamma^* \triangleq \max_{\mathbf{p} \in \Delta_K} \inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \Gamma^c(\mathbf{p}; \vartheta). \quad (2)$$

**Sufficient condition for optimality.** We characterize the maximal achievable cost-normalized large-deviation rate of posterior concentration and provide a convenient sufficient condition under which an algorithm attains the optimal exponent. As in Russo (2020), we impose the following mild boundedness condition on the prior.

**Assumption 4.** *The prior  $\Pi_0$  admits a density  $\pi_0$  with respect to Lebesgue measure on  $\Theta$  such that*

$$0 < \inf_{\boldsymbol{\theta} \in \Theta} \pi_0(\boldsymbol{\theta}) \leq \sup_{\boldsymbol{\theta} \in \Theta} \pi_0(\boldsymbol{\theta}) < \infty.$$

**Theorem 1.** *Assume Assumptions 1–4. Then, for any adaptive sampling rule,*

$$\limsup_{n \rightarrow \infty} -\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) \leq \Gamma^* \quad \mathbb{P}_{\boldsymbol{\theta}}\text{-a.s.}$$

Moreover, suppose an algorithm produces sample allocation  $\mathbf{p}_n$  satisfying

$$\lim_{n \rightarrow \infty} \inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \Gamma^c(\mathbf{p}_n; \vartheta) = \Gamma^* \quad \mathbb{P}_{\boldsymbol{\theta}}\text{-a.s.} \quad (3)$$

Then,

$$\lim_{n \rightarrow \infty} -\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) = \Gamma^* \quad \mathbb{P}_{\boldsymbol{\theta}}\text{-a.s.}$$

Theorem 1 is convenient in that it reduces asymptotic optimality to an almost sure *value convergence* property: it suffices to show that the algorithm drives the worst-case cost-normalized discrimination rate  $\inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \Gamma^c(\mathbf{p}_n; \vartheta)$  to  $\Gamma^*$ .

### 3 The Algorithm

In this section, we adapt the Pitfall-Adapted Nomination (PAN) algorithm of Qin and You (2025) to our cost-aware setting.

#### 3.1 Decomposition of the Alternative Set

We impose the following mild structural condition on the alternative set, commonly adopted in the pure exploration literature, e.g., in Wang et al. (2021).

**Assumption 5.** *The alternative set  $\text{Alt}(\boldsymbol{\theta})$  is a finite union of convex sets. That is, there exists a finite index set  $\mathcal{X}(\boldsymbol{\theta})$  and convex sets  $\{\text{Alt}_x(\boldsymbol{\theta}) : x \in \mathcal{X}(\boldsymbol{\theta})\}$  such that  $\text{Alt}(\boldsymbol{\theta}) = \bigcup_{x \in \mathcal{X}(\boldsymbol{\theta})} \text{Alt}_x(\boldsymbol{\theta})$ .*

The set  $\mathcal{X}(\boldsymbol{\theta})$  captures the fundamental types of “confusing scenarios” under which an alternative instance produces an answer different from  $\mathcal{I}(\boldsymbol{\theta})$ ; see Qin and You (2025, Appendix A) for comprehensive examples.

For each confusing scenario  $x \in \mathcal{X}(\boldsymbol{\theta})$  and any sample allocation  $\mathbf{p}$ , the following quantity captures the cost-normalized discriminative information available to rule it out:

$$D_x(\mathbf{p}; \boldsymbol{\theta}) \triangleq \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta})} \Gamma^c(\mathbf{p}; \boldsymbol{\vartheta}) = \frac{1}{\bar{C}_{\boldsymbol{\theta}}(\mathbf{p})} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta})} \sum_{i=1}^K p_i \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i}),$$

This allows us to rewrite (2) as  $\Gamma^* = \max_{\mathbf{p} \in \Delta_K} \min_{x \in \mathcal{X}(\boldsymbol{\theta})} D_x(\mathbf{p}; \boldsymbol{\theta})$ .

#### 3.2 Cost-Aware PAN Algorithm

The PAN algorithm of Qin and You (2025), originally developed for non-cost-aware exploration in unstructured bandits, proceeds in three steps. At each round  $n$ , it performs: (i) *estimation*, producing an estimator  $\boldsymbol{\theta}_n$ ; (ii) *detection*, identifying the most confusing scenario (the *pitfall*) in  $\mathcal{X}(\boldsymbol{\theta}_n)$ ; and (iii) *selection*, allocating measurements across arms. We extend PAN to incorporate two additional features, cost-aware exploration and structured bandits, as summarized in Algorithm 1.<sup>1</sup>

**Estimation rule.** The problem instance  $\boldsymbol{\theta}$  is unknown and must be learned from noisy bandit feedback. We assume access to an estimation oracle that, at each round, maps the current history to an instance estimate. In unstructured bandits, the empirical mean vector provides such an oracle, while in linear bandits a natural choice is the (regularized) least-squares estimator.

**Assumption 6.** *For each round  $n$ , given the history  $\mathcal{H}_n$ , an estimation routine returns an estimate  $\boldsymbol{\theta}_n = \boldsymbol{\theta}_n(\mathcal{H}_n) \in \Theta$  of the true instance  $\boldsymbol{\theta}$ .*

---

<sup>1</sup>The algorithm is *anytime* in the sense that the sampling rule itself does not depend on  $T$ .

**Detection rule.** The detection rule identifies the currently most confusing alternative scenario  $x \in \mathcal{X}(\boldsymbol{\theta}_n)$ . Recall that  $D_x(\mathbf{p}_n; \boldsymbol{\theta}_n)$  denotes the plug-in estimate of the cost-normalized discriminative information available to rule out scenario  $x$  under the current sample allocations  $\mathbf{p}_n$ . Accordingly, we select the scenario with the smallest value of  $D_x(\mathbf{p}_n; \boldsymbol{\theta}_n)$ , i.e.,

$$x_n \in \arg \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x(\mathbf{p}_n; \boldsymbol{\theta}_n), \quad \text{breaking ties arbitrarily.} \quad (4)$$

Although  $D_x(\mathbf{p}; \boldsymbol{\theta})$  is defined via a cost normalization, the argmin in (4) is in fact independent of the cost functions  $C_i(\cdot)$ . Indeed, for any fixed  $(\mathbf{p}, \boldsymbol{\theta})$  we have  $\bar{C}_{\boldsymbol{\theta}}(\mathbf{p}) > 0$  by Assumption 2 and the factor  $1/\bar{C}_{\boldsymbol{\theta}}(\mathbf{p})$  does not depend on  $x$ . Therefore,

$$\arg \min_{x \in \mathcal{X}(\boldsymbol{\theta})} D_x(\mathbf{p}; \boldsymbol{\theta}) = \arg \min_{x \in \mathcal{X}(\boldsymbol{\theta})} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta})} \sum_{i=1}^K p_i \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i}). \quad (5)$$

This is expected: the detection step is purely a statistical discrimination task and is therefore independent of the cost of information collection. Similarly, for a given  $x$ , sample allocations  $\mathbf{p}$  and any problem instance  $\boldsymbol{\theta} \in \Theta$ , the hardest instance<sup>2</sup> within  $\text{Alt}_x$  is also independent of the cost:

$$\boldsymbol{\vartheta}_x = \boldsymbol{\vartheta}_x(\mathbf{p}, \boldsymbol{\theta}) \in \arg \min_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta})} \Gamma^c(\mathbf{p}; \boldsymbol{\vartheta}) = \arg \min_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta})} \Gamma(\mathbf{p}; \boldsymbol{\vartheta}).$$

**Selection rule.** We generalize the Information-Directed Selection (IDS) rule of Qin and You (2025) into the cost-aware IDS sampling probabilities

$$H_i^x(\mathbf{p}; \boldsymbol{\theta}) \triangleq \frac{p_i \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta}_x,i}) / C_i(\boldsymbol{\theta})}{\sum_{j=1}^K p_j \text{KL}(P_{\boldsymbol{\theta},j} \| P_{\boldsymbol{\vartheta}_x,j}) / C_j(\boldsymbol{\theta})}, \quad i \in [K]. \quad (6)$$

Let  $\mathbf{H}^x(\mathbf{p}; \boldsymbol{\theta}) \triangleq (H_1^x(\mathbf{p}; \boldsymbol{\theta}), \dots, H_K^x(\mathbf{p}; \boldsymbol{\theta})) \in \Delta_K$ .

**Forced exploration.** To guarantee sufficient exploration, we include forced-exploration blocks of length  $K$  that begin at times  $n = Km^2$  for integers  $m \geq 0$ ; over the subsequent  $K$  rounds, each arm is pulled exactly once. Between forced blocks, PAN follows the cost-aware IDS rule (6). Accordingly, the sample allocation  $\mathbf{p}_n \in \Delta_K$  evolves as

$$\mathbf{p}_{n+1} = \mathbf{p}_n + \frac{1}{n+1} (\mathbf{e}_{I_{n+1}} - \mathbf{p}_n), \quad I_{n+1} \begin{cases} \sim \mathbf{H}^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n), & \text{if } \sqrt{[n/K]} \notin \mathbb{Z}, \\ = 1 + (n \bmod K), & \text{if } \sqrt{[n/K]} \in \mathbb{Z}. \end{cases} \quad (7)$$

Our main result is the optimality of Algorithm 1.

**Theorem 2.** *Under Assumptions 1–7, Algorithm 1 satisfies*

$$\lim_{n \rightarrow \infty} -\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) = \Gamma^* \quad \mathbb{P}_{\boldsymbol{\theta}}\text{-a.s.}$$

---

<sup>2</sup>We establish uniqueness of such  $\boldsymbol{\vartheta}_x$  under mild conditions in Lemma 1.

---

**Algorithm 1** Cost-Aware PAN Algorithm

---

**Require:** Horizon  $T \in \mathbb{N}$ , estimation routine  $\text{est}$ 

- 1: **Initialize:**  $N_i(0) \leftarrow 0$  for all  $i \in [K]$ ; initialize an estimate  $\boldsymbol{\theta}_0$ ; set  $\mathbf{p}_0 \leftarrow (1/K)\mathbf{1}$
- 2: **for**  $n = 0, 1, \dots, T-1$  **do**
- 3:      $\boldsymbol{\theta}_n \leftarrow \text{est}(\mathcal{H}_n)$  ▷ Estimation
- 4:     **if**  $\sqrt{\lfloor n/K \rfloor} \in \mathbb{Z}$  **then**
- 5:          $I_{n+1} \leftarrow 1 + (n \bmod K)$  ▷ Forced exploration
- 6:     **else**
- 7:          $x_n \in \arg \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x(\mathbf{p}_n; \boldsymbol{\theta}_n)$ , breaking ties arbitrarily ▷ Detection
- 8:         Draw  $I_{n+1} \sim \mathbf{H}^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n)$  using (6) ▷ Selection
- 9:     **end if**
- 10:    Pull arm  $I_{n+1}$ , observe  $Y_{n+1} \sim P_{\boldsymbol{\theta}, I_{n+1}}$ , update  $\mathbf{p}_n$  and  $\mathcal{H}_{n+1} = \sigma(\mathcal{H}_n, I_{n+1}, Y_{n+1})$
- 11: **end for**
- 12: **return**  $\mathcal{I}_T = \mathcal{I}(\boldsymbol{\theta}_T)$  where  $\boldsymbol{\theta}_T \leftarrow \text{est}(\mathcal{H}_T)$

---

## 4 Dynamics of the Cost Allocations

The optimal cost-normalized exponent (2) is defined over sample allocations  $\mathbf{p} \in \Delta_K$  through the fractions, which is not concave in  $\mathbf{p}$ . For the dynamical analysis of Algorithm 1, it is therefore convenient to reparameterize the allocation by *cost allocations*, which linearize the objective and yield a concave maximin problem.

### 4.1 Cost-Allocation Iterations

Fix the true instance  $\boldsymbol{\theta}$ . For any  $\mathbf{p} \in \Delta_K$ , define the associated *cost allocation*

$$w_i = w_i(\mathbf{p}) \triangleq \frac{p_i C_i(\boldsymbol{\theta})}{\bar{C}_{\boldsymbol{\theta}}(\mathbf{p})}, \quad \bar{C}_{\boldsymbol{\theta}}(\mathbf{p}) \triangleq \sum_{j=1}^K p_j C_j(\boldsymbol{\theta}), \quad i \in [K]. \quad (8)$$

Thus  $w_i$  is the fraction of the average per-sample cost attributed to arm  $i$  under  $\mathbf{p}$ . By Assumption 2, the map  $\mathbf{p} \mapsto \mathbf{w}(\mathbf{p})$  is a bijection on  $\Delta_K$ .<sup>3</sup> Under (8), the cost-normalized information rate (1) becomes linear in  $\mathbf{w}$ :

$$\tilde{\Gamma}^c(\mathbf{w}; \boldsymbol{\vartheta}) \triangleq \sum_{i=1}^K w_i \frac{\text{KL}(P_{\boldsymbol{\theta}, i} \| P_{\boldsymbol{\vartheta}, i})}{C_i(\boldsymbol{\theta})}. \quad (9)$$

Indeed,  $\Gamma^c(\mathbf{p}; \boldsymbol{\vartheta}) = \tilde{\Gamma}^c(\mathbf{w}(\mathbf{p}); \boldsymbol{\vartheta})$  for all  $\mathbf{p} \in \Delta_K$  and  $\boldsymbol{\vartheta} \in \Theta$ . Consequently, the optimal exponent admits the equivalent concave maximization  $\Gamma^* = \max_{\mathbf{w} \in \Delta_K} \inf_{\boldsymbol{\vartheta} \in \text{Alt}(\boldsymbol{\theta})} \tilde{\Gamma}^c(\mathbf{w}; \boldsymbol{\vartheta})$ . Under the decomposition in Assumption 5,

$$\Gamma^* = \max_{\mathbf{w} \in \Delta_K} \min_{x \in \mathcal{X}(\boldsymbol{\theta})} D_x^w(\mathbf{w}; \boldsymbol{\theta}), \quad \text{where} \quad D_x^w(\mathbf{w}; \boldsymbol{\theta}) \triangleq \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta})} \tilde{\Gamma}^c(\mathbf{w}; \boldsymbol{\vartheta}). \quad (10)$$

The next lemma collects structural properties used throughout the sequel.

---

<sup>3</sup>The cost allocations are introduced for the analysis purpose and is not used by the algorithm. Consequently, we allow it to access the true costs to simplify exposition.

**Lemma 1.** Fix  $\boldsymbol{\theta}$  and  $x \in \mathcal{X}(\boldsymbol{\theta})$ . For  $\mathbf{w} \in \mathbb{R}_{\geq 0}^K$ , the map  $\mathbf{w} \mapsto D_x^w(\mathbf{w}; \boldsymbol{\theta})$  is nonnegative, continuous, concave and coordinatewise nondecreasing on  $\Delta_K$ , and degree-one homogeneous:  $D_x^w(\lambda \mathbf{w}; \boldsymbol{\theta}) = \lambda D_x^w(\mathbf{w}; \boldsymbol{\theta})$  for all  $\lambda > 0$ . For  $\mathbf{w} \in \mathbb{R}_{>0}^K$ , the minimizer  $\boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta}) \triangleq \arg \min_{\boldsymbol{\vartheta} \in \text{cl}(\text{Alt}_x(\boldsymbol{\theta}))} \tilde{\Gamma}^c(\mathbf{w}, \boldsymbol{\vartheta})$  is unique,  $D_x^w(\cdot; \boldsymbol{\theta})$  is continuously differentiable and  $[\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})]_i = \text{KL}(P_{\boldsymbol{\theta}, i} \| P_{\boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta}), i}) / C_i(\boldsymbol{\theta})$ . Moreover, there exists  $d_x(\boldsymbol{\theta}) > 0$  such that  $\sum_{i=1}^K [\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})]_i \geq d_x(\boldsymbol{\theta})$  for all  $\mathbf{w} \in \text{int}(\Delta_K)$ , and hence  $D_x^w(\mathbf{w}; \boldsymbol{\theta}) = \langle \mathbf{w}, \nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta}) \rangle > 0$  on  $\text{int}(\Delta_K)$ .

We impose a gradient bound that is satisfied by most pure-exploration tasks; see Wang et al. (2021, Lemma 1).

**Assumption 7.** For all  $x \in \mathcal{X}(\boldsymbol{\theta})$  and all differentiability points  $\mathbf{w}$ , the gradient  $\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})$  exists and is uniformly bounded: there is  $M = M(\boldsymbol{\theta}) < \infty$  such that  $\|\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})\|_{\infty} \leq M$ .

**Cost allocations along the trajectory.** To analyze the asymptotic behavior of the cost allocations along the trajectory, we denote by  $\mathbf{w}_n$  the cost allocation associated with  $\mathbf{p}_n$  and *true*  $\boldsymbol{\theta}$ , i.e.,

$$w_{n,i} \triangleq \frac{p_{n,i} C_i(\boldsymbol{\theta})}{\bar{C}_{\boldsymbol{\theta}}(\mathbf{p}_n)} = \frac{N_{n,i} C_i(\boldsymbol{\theta})}{B_n}, \quad i \in [K].$$

From the sampling process,  $\mathbf{w}_n$  follows the exact update

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \alpha_{n+1} (\mathbf{e}_{I_{n+1}} - \mathbf{w}_n), \quad \text{where } \alpha_{n+1} \triangleq \frac{C_{I_{n+1}}(\boldsymbol{\theta})}{B_{n+1}}. \quad (11)$$

Under Assumption 2, the stepsize satisfies  $\frac{c_{\min}}{c_{\max}(n+1)} \leq \alpha_{n+1} \leq \frac{c_{\max}}{c_{\min}(n+1)}$ . This recursion is the budget analogue of the frequentist update  $\mathbf{p}_{n+1} = \mathbf{p}_n + \frac{1}{n+1} (\mathbf{e}_{I_{n+1}} - \mathbf{p}_n)$ .

In the  $\mathbf{w}$ -coordinate system, the detection and selection rules can be naturally related to their counterparts in the  $\mathbf{p}$ -coordinates through the cost-weighted transformation.

**Detection rule in  $\mathbf{w}$ -coordinates.** The detection step (4) in  $\mathbf{w}$ -coordinate becomes

$$\begin{aligned} x_n \in \arg \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x(\mathbf{p}_n; \boldsymbol{\theta}_n) &= \arg \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} \frac{1}{\bar{C}_{\boldsymbol{\theta}_n}(\mathbf{p}_n)} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta}_n)} \sum_{i=1}^K p_{n,i} \text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}, i}) \\ &= \arg \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} \frac{1}{\bar{C}_{\boldsymbol{\theta}}(\mathbf{p}_n)} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta}_n)} \sum_{i=1}^K p_{n,i} \text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}, i}) \\ &= \arg \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta}_n)} \sum_{i=1}^K w_{n,i} \frac{\text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}, i})}{C_i(\boldsymbol{\theta})}. \end{aligned} \quad (12)$$

Although the explicit expression for the detection rule in  $\mathbf{w}$ -coordinates involves the cost terms  $C_i(\boldsymbol{\theta})$ , as implied by (5), the resulting most confusing answer  $x_n$  is actually independent of the cost. This stems from the fact that the detection step is fundamentally a statistical decision problem, aimed at identifying the most confusing alternative hypothesis based on how distinguishable the current estimate is from each alternative. The costs influence how we allocate resources across the arms, but they do not change the underlying task of distinguishing between hypotheses.

**Selection rule in  $w$ -coordinates.** IDS admits an elegant gradient representation in the  $w$ -coordinates. For a fixed  $x \in \mathcal{X}(\boldsymbol{\theta})$  and at a point  $\mathbf{w}$  where  $D_x^w(\cdot; \boldsymbol{\theta})$  is differentiable and positive, we define

$$\mathbf{h}^x(\mathbf{w}; \boldsymbol{\theta}) \triangleq \frac{\mathbf{w} \circ \nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})}{D_x^w(\mathbf{w}; \boldsymbol{\theta})} \in \Delta_K. \quad (13)$$

Since  $D_x^w(\cdot; \boldsymbol{\theta})$  is homogeneous of degree 1, Euler's identity yields  $D_x^w(\mathbf{w}; \boldsymbol{\theta}) = \langle \mathbf{w}, \nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta}) \rangle$ , confirming that (13) indeed defines a probability vector. Such IDS distribution (13) at round  $n$ , i.e.  $h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)$  can be computed explicitly using Lemma 1. It relates to the PAN algorithm 1 sampling distribution  $H_i^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n) \propto p_{n,i} \text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}_x, i}) / C_i(\boldsymbol{\theta}_n)$  through the following identity:

$$\begin{aligned} h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) &= \frac{w_{n,i} \text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}_x, i}) / C_i(\boldsymbol{\theta}_n)}{\sum_{j=1}^K w_j \text{KL}(P_{\boldsymbol{\theta}, j} \| P_{\boldsymbol{\vartheta}_x, j}) / C_j(\boldsymbol{\theta}_n)} = \frac{\frac{C_i(\boldsymbol{\theta})}{C_{\boldsymbol{\theta}}} p_{n,i} \text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}_x, i}) / C_i(\boldsymbol{\theta}_n)}{\sum_{j=1}^K \frac{C_j(\boldsymbol{\theta})}{C_{\boldsymbol{\theta}}} p_j \text{KL}(P_{\boldsymbol{\theta}, j} \| P_{\boldsymbol{\vartheta}_x, j}) / C_j(\boldsymbol{\theta}_n)} \\ &= \frac{C_i(\boldsymbol{\theta}) p_{n,i} \text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}_x, i}) / C_i(\boldsymbol{\theta}_n)}{\sum_{j=1}^K C_j(\boldsymbol{\theta}) p_j \text{KL}(P_{\boldsymbol{\theta}, j} \| P_{\boldsymbol{\vartheta}_x, j}) / C_j(\boldsymbol{\theta}_n)} = \frac{C_i(\boldsymbol{\theta}) H_i^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n)}{\sum_{j=1}^K C_j(\boldsymbol{\theta}) H_j^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n)}. \end{aligned} \quad (14)$$

Thus, while  $\mathbf{p}$  tracks sampling *frequencies*, the  $\mathbf{w}$  captures the evolution of how the *budget* is distributed across arms, and the IDS rule has a particularly simple form in these coordinates.

**Expected drift in cost coordinates** Let  $q_{n,i} \triangleq \mathbb{P}(I_{n+1} = i \mid \mathcal{H}_n)$  denote the actual sampling distribution at round  $n$  (including forced explorations). Define the expected stepsize

$$\tilde{\alpha}_{n+1} \triangleq \mathbb{E}[\alpha_{n+1} \mid \mathcal{H}_n] = \sum_{i=1}^K q_{n,i} \frac{C_i(\boldsymbol{\theta})}{B_n + C_i(\boldsymbol{\theta})},$$

and introduce the cost-weighted sampling distribution  $\tilde{\mathbf{h}}_n \in \Delta_K$  such that  $\mathbb{E}[\alpha_{n+1} \mathbf{e}_{I_{n+1}} \mid \mathcal{H}_n] = \tilde{\alpha}_{n+1} \tilde{\mathbf{h}}_n$ . Explicitly, the components of  $\tilde{\mathbf{h}}_n$  are given by

$$\tilde{h}_{n,i} \triangleq q_{n,i} \frac{C_i(\boldsymbol{\theta})}{B_n + C_i(\boldsymbol{\theta})} / \left[ \sum_{j=1}^K q_{n,j} \frac{C_j(\boldsymbol{\theta})}{B_n + C_j(\boldsymbol{\theta})} \right].$$

Taking conditional expectation in the update rule (11) yields the expected drift of  $\mathbf{w}_n$ :

$$\mathbb{E}[\mathbf{w}_{n+1} - \mathbf{w}_n \mid \mathcal{H}_n] = \tilde{\alpha}_{n+1} (\tilde{\mathbf{h}}_n - \mathbf{w}_n). \quad (15)$$

The following lemma characterizes the asymptotic relationship between the actual sampling distribution  $\tilde{\mathbf{h}}_n$  and the ideal IDS distribution  $\mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)$ .

**Lemma 2.** *With estimator sequence  $\boldsymbol{\theta}_n \rightarrow \boldsymbol{\theta}$  almost surely,*

1. **Detection consistency:**  $\lim_{n \rightarrow \infty} |D_{x_n}^w(\mathbf{w}_n, \boldsymbol{\theta}_n) - \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_n, \boldsymbol{\theta}_n)| = 0$  almost surely.

2. **Selection consistency:** *On non-forced rounds, the sampling distribution  $\tilde{\mathbf{h}}_n$  satisfies*

$$|\tilde{h}_{n,i} - h_i^{x_n}(\mathbf{w}_n, \boldsymbol{\theta}_n)| \leq \frac{c_{\max}}{n c_{\min}} h_i^{x_n}(\mathbf{w}_n, \boldsymbol{\theta}_n) \quad \text{for all } i \in [K], \quad (16)$$

*and consequently  $\|\tilde{\mathbf{h}}_n - \mathbf{h}^{x_n}(\mathbf{w}_n, \boldsymbol{\theta}_n)\|_{\infty} \rightarrow 0$  at rate  $O(1/n)$ .*

Lemma 2 shows that, in cost coordinates, the stochastic sampling process asymptotically follows a deterministic drift direction given by the IDS distribution  $\mathbf{h}^{x_n}(\mathbf{w}_n, \boldsymbol{\theta}_n)$ . This gradient-like form of IDS naturally provides a convenient optimization perspective for our analysis.

## 4.2 Continuous-Time Dynamics

We study the discrete-time iterates via their limiting continuous-time dynamics as  $n \rightarrow \infty$ , i.e., as the step size vanishes. Throughout this subsection, we fix the true instance  $\boldsymbol{\theta}$  and suppress it in the notation: write  $D_x^w(\mathbf{w}) \equiv D_x^w(\mathbf{w}; \boldsymbol{\theta})$  and  $\mathcal{X} \equiv \mathcal{X}(\boldsymbol{\theta})$ .

Recall that the discrete IDS rule in (13) uses  $\nabla_{\mathbf{w}} D_x^w$  and divides by  $D_x^w$ . Two issues therefore arise on simplex boundary: (i)  $D_x^w$  may fail to be differentiable, and (ii) the denominator can vanish. To isolate points where the discrete rule is well defined, define the *regular region*<sup>4</sup>

$$\mathcal{R} \triangleq \left\{ \mathbf{w} \in \Delta_K : \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}) > 0 \text{ and } D_x^w \text{ is } C^1 \text{ at } \mathbf{w}, \forall x \in \arg \min_{x' \in \mathcal{X}} D_{x'}^w(\mathbf{w}) \right\}.$$

For each  $\mathbf{w} \in \mathcal{R}$ , the IDS sampling distribution is well defined. By Lemma 1, for every  $x \in \mathcal{X}$  and  $\mathbf{p} \in \text{int}(\Delta_K)$ , the function  $D_x^w(\cdot; \boldsymbol{\theta})$  is strictly positive and  $C^1$ ; hence  $\text{int}(\Delta_K) \subseteq \mathcal{R}$  and  $\overline{\mathcal{R}} = \Delta_K$ . Therefore IDS can be extended from  $\mathcal{R}$  to the full simplex via an appropriate limit construction.

Defining a continuous-time analogue of IDS entails an additional subtlety. As  $n \rightarrow \infty$ , the detected index  $x_n$  can switch on a fast time scale, while the sample allocation  $\mathbf{w}_n$  evolves on a slower time scale. Consequently, in the continuous-time limit, the sampling decision at time  $t$  should respond not to a single scenario, but to the entire *active set*  $\arg \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t))$ . To capture the limiting effect of rapid switching among active scenarios, we introduce an *IDS correspondence*: a set-valued extension of the discrete IDS rule that collects all limiting sampling responses induced by such fast switching. These limiting responses are parameterized by an “average” detection-frequency vector  $\boldsymbol{\nu} \in \text{det}(\mathbf{w})$ , where  $\text{det}(\mathbf{w})$  is the detection correspondence defined as

$$\text{det}(\mathbf{w}) \triangleq \arg \min_{\boldsymbol{\nu} \in \Delta_{|\mathcal{X}|}} F(\mathbf{w}, \boldsymbol{\nu}).$$

At the slower time scale, the discrete IDS updates track a *smoothed* response to an average frequency of confusing scenarios, represented by some  $\boldsymbol{\nu} \in \text{det}(\mathbf{w})$ . Define

$$F(\mathbf{w}, \boldsymbol{\nu}) \triangleq \sum_{x \in \mathcal{X}} \nu_x D_x^w(\mathbf{w}), \quad \forall \mathbf{w} \in \Delta_K, \quad \forall \boldsymbol{\nu} \in \Delta_{|\mathcal{X}|}.$$

For  $\mathbf{w} \in \mathcal{R}$  and  $\boldsymbol{\nu} \in \text{det}(\mathbf{w})$ , define the averaged IDS response

$$\mathbf{h}(\mathbf{w}, \boldsymbol{\nu}) \triangleq \sum_{x \in \mathcal{X}} \nu_x \mathbf{h}^x(\mathbf{w}) = \frac{\mathbf{w} \circ \nabla_{\mathbf{w}} F(\mathbf{w}, \boldsymbol{\nu})}{F(\mathbf{w}, \boldsymbol{\nu})} \in \Delta_K, \quad (17)$$

and the IDS selection correspondence

$$\text{sel}(\mathbf{w}) \triangleq \text{conv} \left\{ \lim_{n \rightarrow \infty} \mathbf{h}(\mathbf{w}^n, \boldsymbol{\nu}^n) : \mathbf{w}^n \rightarrow \mathbf{w}, \mathbf{w}^n \in \mathcal{R}, \boldsymbol{\nu}^n \in \text{det}(\mathbf{w}^n) \right\}.$$

The continuous-time evolution of  $\mathbf{w}$  is then described by the differential inclusion (DI)

$$\dot{\mathbf{w}}(t) \in G(\mathbf{w}(t)) \triangleq \text{sel}(\mathbf{w}(t)) - \mathbf{w}(t), \quad \mathbf{w}(0) = \mathbf{w}_0 \in \Delta_K. \quad (18)$$

It is worth noting the similarity between the DI and its discrete version in (15).

<sup>4</sup>For boundary  $\mathbf{w}$ , we interpret continuous differentiability as  $C^1$  on a neighborhood of  $\mathbf{w}$  within the relative interior of the smallest face containing  $\mathbf{w}$ .

**Proposition 1** (Properties of the DI). *The correspondences  $\det$  and  $\sel$  are nonempty, convex-valued, compact-valued, and upper hemicontinuous on  $\Delta_K$ . For every  $\mathbf{w}_0 \in \Delta_K$ , the differential inclusion (18) admits an absolutely continuous solution  $\mathbf{w}(\cdot)$  with  $\mathbf{w}(0) = \mathbf{w}_0$ , and  $\Delta_K$  is forward invariant. Moreover, for any solution  $\mathbf{w}(\cdot)$  there exists a measurable selection  $t \mapsto \mathbf{s}(t) \in \sel(\mathbf{w}(t))$  such that  $\dot{\mathbf{w}}(t) = \mathbf{s}(t) - \mathbf{w}(t)$  for a.e.  $t \geq 0$ .*

## 5 Convergence Analysis

The main result of this paper is the *value convergence* of Algorithm 1. We show that both the cost allocation and the sampling frequencies induce the optimal value.

**Theorem 3.** *Under Assumptions 1–7, let  $\{\mathbf{w}_n\}_{n \geq 0}$  be the cost allocation sequence generated by Algorithm 1 and iteration (11). Then,  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}_n) \rightarrow \Gamma^*$  almost surely. As a consequence, for  $\{\mathbf{p}_n\}_{n \geq 0}$  from Algorithm 1, we also have  $\min_{x \in \mathcal{X}} D_x(\mathbf{p}_n; \boldsymbol{\theta}_n) \rightarrow \Gamma^*$  almost surely.*

In view of the sufficient condition in Theorem 1, the value convergence of the sampling allocations generated by Algorithm 1 (Theorem 3) certifies the asymptotic optimality claimed in Theorem 2. Our analysis proceeds in four steps.

First, in Appendix E.1, we construct a globally well-defined information value that serves as a Lyapunov function for the limiting differential inclusion and show that it is nondecreasing along every solution trajectory. By LaSalle’s invariance principle, the associated  $\omega$ -limit set has empty interior. Second, in Appendix E.2, we transfer this continuous-time structure to the stochastic iterates via stochastic-approximation tools—the asymptotic pseudo-trajectory (APT) framework and the characterization of internally chain transitive (ICT) sets—which yields convergence of the information value along the discrete-time trajectory. Third, in Appendix E.3, we rule out convergence to the degenerate value 0 under interior initialization, ensuring the dynamics remains in a positive-information regime. Finally, in Appendix E.4, we use the Kullback–Leibler divergence as an energy function on the active set and establish a uniform negative-drift bound. This forces the limiting information value to coincide with the optimal value  $F^*$ , completing the proof.

## 6 Numerical Experiment

We implement Algorithm 1 and compare against four baselines: (i) *Track-and-Stop*, where the sampling proportions are obtained by running the deterministic version (i.e., remove the estimation step and assume that  $\hat{\boldsymbol{\theta}}_n$  is the ground truth) of Algorithm 1 for 100 iterations on the current estimate  $\hat{\boldsymbol{\theta}}_n$ , coupled with a classical track-and-stop rule; (ii) *LinGapE* (Xu et al. 2018); (iii) *OD-LinBAI* (Yang and Tan 2022); and (iv) *Uniform* sampling. All methods use a ridge-regression estimation oracle. Additional experiments are deferred to Appendix G (heterogeneous costs), and we refer to Qin and You (2025) for unit-cost variants in the fixed-confidence setting.

We consider fixed-confidence linear best-arm identification with  $d = 5$  and  $K = 6$ . Arms  $a_i = e_i$  for  $i \in \{1, \dots, 5\}$ , and  $a_6 = \cos(\omega)e_1 + \sin(\omega)e_2$  with  $\omega = 0.01$ , so  $a_6$  is nearly collinear with  $a_1$ .

The true parameter is  $\boldsymbol{\theta} = (2, 0, 0, 0, 0)$  with unit noise variance and unit costs, yielding  $\mu_1 = 2$  and  $\mu_6 = 2 \cos(\omega) \approx 1.9999$ . This is the classical hard instance of Soare et al. (2014): distinguishing  $a_1$  from  $a_6$  requires shrinking uncertainty along the nearly  $-e_2$  contrast direction, so accurately estimating  $\theta_2$  is essential even though arm 2 is far from optimal in mean.

For each replicate, we record the smallest budget  $n$  such that the posterior error  $\Pi_n(\text{Alt}(\boldsymbol{\theta})) \leq 10^{-3}$ , censoring at  $B = 10^6$  if the target is not reached. Figure 1 reports the distribution of budgets needed to reach the target over 1000 replicates. Algorithm 1 is the best-performing method, with a median budget of  $2.77 \times 10^4$ , while Track-and-Stop is slightly slower with median  $3.32 \times 10^4$ . In contrast, LinGapE, OD-LinBAI, and Uniform have markedly larger medians ( $8.85 \times 10^4$ ,  $1.08 \times 10^5$ , and  $1.44 \times 10^5$ , respectively).

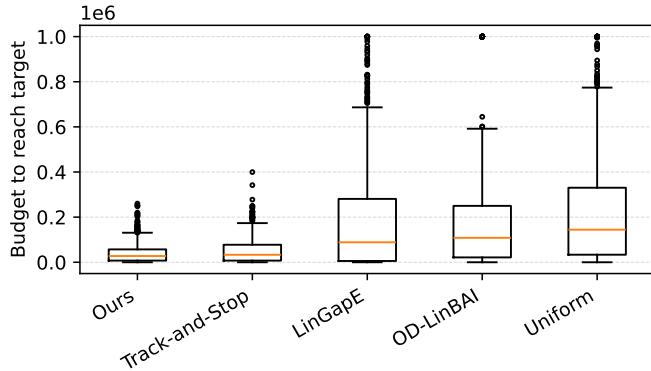


Figure 1: Budget to reach target posterior error  $\Pi_n(\text{Alt}(\boldsymbol{\theta})) \leq 10^{-3}$  over 1000 replicates.

The behavior matches the oracle structure: the optimal allocation places almost all mass on arm 2 ( $p_2^* \approx 0.995$ ). Algorithm 1 quickly discovers and tracks this, with average final allocation  $\approx 0.994$  on arm 2. Track-and-Stop (with our approximate oracle) assigns less to arm 2 on average ( $\approx 0.756$ ), while LinGapE, OD-LinBAI, and Uniform overweight the contender arms (1 and 6), slowing contraction along the critical contrast and producing heavy-tailed budgets.

## 7 Conclusion

We studied pure exploration with heterogeneous per-measurement costs and characterized the optimal large-deviation rate at which the posterior probability of error can decay *per unit spent budget*. Our main contribution is a cost-aware extension of the Pitfall-Adapted Nomination (PAN) framework that attains this optimal cost-normalized posterior error exponent almost surely under mild regularity conditions, for general one-parameter exponential-family rewards and structured bandit models. Conceptually, the results clarify that cost heterogeneity fundamentally changes the design criterion from “information per sample” to *information per unit cost*, and that asymptotically optimal behavior is governed by budget allocations rather than sampling frequencies.

Technically, we developed a proof strategy based on continuous-time approximation via differential inclusions, which accommodates fast switching in the detection step, nonsmoothness of the Chernoff

information objective, and the set-valued nature of IDS at ties and on the boundary of the simplex. This perspective may be useful beyond the present setting, offering a general toolkit for analyzing bandit sampling rules that behave like nonsmooth first-order methods.

Finally, our analysis incorporates forced exploration to streamline technical arguments and ensure sufficient excitation. Empirically, however, we observe that forced exploration is not essential for practical performance. Establishing asymptotic optimality without this mechanism is an interesting direction for future work.

## References

Jean-Pierre Aubin and Arrigo Cellina. *Differential Inclusions: Set-Valued Maps and Viability Theory*. Springer Berlin Heidelberg, 1984.

Brian Beavis and Ian M Dobbs. *Optimization and stability theory for economic analysis*. Cambridge university press, 1990.

Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Stochastic subgradient descent escapes active strict saddles on weakly convex functions. *Mathematics of Operations Research*, 49(3):1761–1790, 2024.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188(1):19–51, 2021.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 9. Springer, 2008.

Haim Brezis. *Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, volume 5. Elsevier, 1973.

Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27, 2014.

Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics*, 20(1):119–154, 2020.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

Ignacio Esponda, Demian Pouzo, and Yuichi Yamamoto. Corrigendum to “asymptotic behavior of bayesian learners with misspecified models”[j. econ. theory 195 (2021) 105260]. *Journal of Economic Theory*, 204:105513, 2022.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Joy He-Yueya, Jonathan Lee, Matthew Jörke, and Emma Brunskill. Cost-aware near-optimal policy learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28088–28096, 2025.

Xiaoyan Hu, Lauren Pick, Ho-fung Leung, and Farzan Farnia. Promptwise: Online learning for cost-aware prompt assignment in generative models. *arXiv preprint arXiv:2505.18901*, 2025.

Alexander D Ioffe. Variational analysis of regular mappings. *Springer Monographs in Mathematics. Springer, Cham*, 2017.

Marc Jourdan. *Solving pure exploration problems with the Top Two approach*. PhD thesis, Université de Lille, 2024.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. *Advances in Neural Information Processing Systems*, 35:26791–26803, 2022.

Kellen Kanarios, Qining Zhang, and Lei Ying. Cost aware best arm identification. *Reinforcement Learning Journal*, 2024.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

Pierre Ménard. Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*, 2019.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

Riccardo Poiani, Rémy Degenne, Emilie Kaufmann, Alberto Maria Metelli, and Marcello Restelli. Optimal multi-fidelity best-arm identification. *Advances in Neural Information Processing Systems*, 37:121882–121927, 2024.

Chao Qin and Daniel Russo. Optimizing adaptive experiments: A unified approach to regret minimization and best-arm identification. *arXiv preprint arXiv:2402.10592*, 2024.

Chao Qin and Wei You. Dual-directed algorithm design for efficient pure exploration. *Operations Research*, 2025.

Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 30, 2017.

Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

Xuedong Shang, Rianne Heide, Pierre Menard, Emilie Kaufmann, and Michal Valko. Fixed-confidence guarantees for bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR, 2020.

Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-arm identification in linear bandits. *Advances in neural information processing systems*, 27, 2014.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via Frank-Wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

Di Wu, Chengshuai Shi, Ruida Zhou, and Cong Shen. Cost-aware optimal pairwise pure exploration. *arXiv preprint arXiv:2503.07877*, 2025.

Liyuan Xu, Junya Honda, and Masashi Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851. PMLR, 2018.

Junwen Yang and Vincent Tan. Minimax optimal fixed-budget best arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 35:12253–12266, 2022.

Wei You, Chao Qin, Zihao Wang, and Shuoguang Yang. Information-directed selection for top-two algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2850–2851. PMLR, 2023.

# Appendix

## A Proof of Theorem 1

Our proof follows the basic template of Russo (2020, Theorem 1), adapted to the cost-aware and structural bandit setting. The execution is relatively routine, nevertheless, we include full arguments for completeness.

### A.1 Preliminaries

Recall the average KL divergence under a probability vector  $\mathbf{p} \in \Delta_K$ :

$$\Gamma(\mathbf{p}; \boldsymbol{\vartheta}) \triangleq \sum_{i=1}^K p_i \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i}).$$

Let the log-likelihood ratio up to time  $n$  be

$$\Lambda_n(\boldsymbol{\theta} \| \boldsymbol{\vartheta}) \triangleq \sum_{\ell=1}^n \log \frac{p_{\boldsymbol{\theta},I_\ell}(Y_\ell)}{p_{\boldsymbol{\vartheta},I_\ell}(Y_\ell)}.$$

**Lemma 3** (Uniform log-likelihood approximation). *Under Assumption 1,*

$$\lim_{n \rightarrow \infty} \sup_{\boldsymbol{\vartheta} \in \Theta} \left| \frac{1}{n} \Lambda_n(\boldsymbol{\theta} \| \boldsymbol{\vartheta}) - \Gamma(\bar{\boldsymbol{\psi}}_n; \boldsymbol{\vartheta}) \right| = 0 \quad \text{a.s.}$$

*Proof.* Fix the true instance  $\boldsymbol{\theta}$ . For each fixed  $\boldsymbol{\vartheta} \in \Theta$ , define

$$X_\ell(\boldsymbol{\vartheta}) \triangleq \log \frac{p_{\boldsymbol{\theta},I_\ell}(Y_\ell)}{p_{\boldsymbol{\vartheta},I_\ell}(Y_\ell)}.$$

Then  $X_\ell(\boldsymbol{\vartheta})$  is  $\mathcal{H}_\ell$ -measurable and

$$\mathbb{E}_{\boldsymbol{\theta}}[X_\ell(\boldsymbol{\vartheta}) | \mathcal{H}_{\ell-1}] = \sum_{i=1}^K \psi_{\ell,i} \mathbb{E}_{\boldsymbol{\theta}} \left[ \log \frac{p_{\boldsymbol{\theta},i}(Y)}{p_{\boldsymbol{\vartheta},i}(Y)} \right] = \sum_{i=1}^K \psi_{\ell,i} \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i}).$$

Therefore,

$$\Lambda_n(\boldsymbol{\theta} \| \boldsymbol{\vartheta}) = \sum_{\ell=1}^n X_\ell(\boldsymbol{\vartheta}) = \sum_{\ell=1}^n \mathbb{E}_{\boldsymbol{\theta}}[X_\ell(\boldsymbol{\vartheta}) | \mathcal{H}_{\ell-1}] + M_n(\boldsymbol{\vartheta}) = n\Gamma(\bar{\boldsymbol{\psi}}_n; \boldsymbol{\vartheta}) + M_n(\boldsymbol{\vartheta}),$$

where  $M_n(\boldsymbol{\vartheta})$  is the martingale

$$M_n(\boldsymbol{\vartheta}) \triangleq \sum_{\ell=1}^n \delta_\ell(\boldsymbol{\vartheta}), \quad \delta_\ell(\boldsymbol{\vartheta}) \triangleq X_\ell(\boldsymbol{\vartheta}) - \mathbb{E}_{\boldsymbol{\theta}}[X_\ell(\boldsymbol{\vartheta}) | \mathcal{H}_{\ell-1}].$$

Thus it remains to show that  $\sup_{\boldsymbol{\vartheta} \in \Theta} |M_n(\boldsymbol{\vartheta})| = o(n)$  a.s.

Under Assumption 1, we have a finite uniform bound

$$\kappa_{\text{KL}} \triangleq \sup_{i \in [K]} \sup_{\boldsymbol{\vartheta} \in \Theta} \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i}) \leq \sup_{i \in [K]} \sup_{\boldsymbol{\vartheta} \in \Theta} \left( \mathbb{E}_{\boldsymbol{\theta}} \left[ \left( \log \frac{p_{\boldsymbol{\theta},i}(Y)}{p_{\boldsymbol{\vartheta},i}(Y)} \right)^2 \right] \right)^{1/2} < \infty. \quad (19)$$

In particular, for each fixed  $\boldsymbol{\vartheta}$ ,  $\{M_n(\boldsymbol{\vartheta})\}_{n \geq 1}$  is square-integrable and

$$\mathbb{E}_{\boldsymbol{\theta}} [\delta_\ell(\boldsymbol{\vartheta})^2 \mid \mathcal{H}_{\ell-1}] \leq 4\mathbb{E}_{\boldsymbol{\theta}} [X_\ell(\boldsymbol{\vartheta})^2 \mid \mathcal{H}_{\ell-1}] \leq 4\kappa_{\text{KL}}^2,$$

where we used  $(a - \mathbb{E}[a \mid \mathcal{H}])^2 \leq 4a^2$  and the definition of  $\kappa_X$ . Hence,

$$\sum_{\ell=1}^{\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}} [\delta_\ell(\boldsymbol{\vartheta})^2]}{\ell^2} \leq 4\kappa_{\text{KL}}^2 \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} < \infty.$$

By the strong law of large numbers for square-integrable martingales, this implies  $M_n(\boldsymbol{\vartheta})/n \rightarrow 0$  a.s. for each fixed  $\boldsymbol{\vartheta} \in \Theta$ . Next fix  $\delta > 0$ . By Assumption 1, there exists a finite  $\delta$ -net  $\mathcal{N}_\delta \subset \Theta$ . Since  $\mathcal{N}_\delta$  is finite and  $M_n(\boldsymbol{\nu})/n \rightarrow 0$  a.s. for each  $\boldsymbol{\nu} \in \mathcal{N}_\delta$ , we have

$$\max_{\boldsymbol{\nu} \in \mathcal{N}_\delta} \frac{|M_n(\boldsymbol{\nu})|}{n} \rightarrow 0 \quad \text{a.s.} \quad (20)$$

We now control  $|M_n(\boldsymbol{\vartheta}) - M_n(\boldsymbol{\vartheta}')|$  for nearby parameters. Assumption 1 implies the following Lipschitz property: there exists a finite constant  $L < \infty$  such that for all  $i \in [K]$ , all  $\boldsymbol{\vartheta}, \boldsymbol{\vartheta}' \in \Theta$ , and  $\nu$ -a.e.  $y$ ,

$$\left| \log \frac{p_{\boldsymbol{\vartheta}',i}(y)}{p_{\boldsymbol{\vartheta},i}(y)} \right| \leq L \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\| (1 + |T(y)|). \quad (21)$$

Using (21) with  $i = I_\ell$  and  $y = Y_\ell$  yields, for all  $\boldsymbol{\vartheta}, \boldsymbol{\vartheta}' \in \Theta$ ,

$$|X_\ell(\boldsymbol{\vartheta}) - X_\ell(\boldsymbol{\vartheta}')| = \left| \log \frac{p_{\boldsymbol{\vartheta}',I_\ell}(Y_\ell)}{p_{\boldsymbol{\vartheta},I_\ell}(Y_\ell)} \right| \leq L \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\| (1 + |T(Y_\ell)|). \quad (22)$$

By Jensen's inequality,

$$\begin{aligned} |\mathbb{E}_{\boldsymbol{\theta}} [X_\ell(\boldsymbol{\vartheta}) - X_\ell(\boldsymbol{\vartheta}') \mid \mathcal{H}_{\ell-1}]| &\leq \mathbb{E}_{\boldsymbol{\theta}} [|X_\ell(\boldsymbol{\vartheta}) - X_\ell(\boldsymbol{\vartheta}')| \mid \mathcal{H}_{\ell-1}] \\ &\leq L \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\| (1 + \mathbb{E}_{\boldsymbol{\theta}} [|T(Y_\ell)| \mid \mathcal{H}_{\ell-1}]). \end{aligned} \quad (23)$$

Combining (22)–(23) and the definition of  $\delta_\ell(\cdot)$  gives

$$|\delta_\ell(\boldsymbol{\vartheta}) - \delta_\ell(\boldsymbol{\vartheta}')| \leq |X_\ell(\boldsymbol{\vartheta}) - X_\ell(\boldsymbol{\vartheta}')| + |\mathbb{E}_{\boldsymbol{\theta}} [X_\ell(\boldsymbol{\vartheta}) - X_\ell(\boldsymbol{\vartheta}') \mid \mathcal{H}_{\ell-1}]|$$

and hence

$$|\delta_\ell(\boldsymbol{\vartheta}) - \delta_\ell(\boldsymbol{\vartheta}')| \leq L \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\| (2 + |T(Y_\ell)| + \mathbb{E}_{\boldsymbol{\theta}} [|T(Y_\ell)| \mid \mathcal{H}_{\ell-1}]). \quad (24)$$

Summing (24) over  $\ell \leq n$  yields

$$\frac{|M_n(\boldsymbol{\vartheta}) - M_n(\boldsymbol{\vartheta}')|}{n} \leq L \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\| \left( 2 + \frac{1}{n} \sum_{\ell=1}^n |T(Y_\ell)| + \frac{1}{n} \sum_{\ell=1}^n \mathbb{E}_{\boldsymbol{\theta}} [|T(Y_\ell)| \mid \mathcal{H}_{\ell-1}] \right). \quad (25)$$

Define the constants

$$\kappa_T^{(2)} \triangleq \sup_{i \in [K]} \mathbb{E}_{\boldsymbol{\theta}} [T(Y)^2] < \infty, \quad \kappa_T^{(1)} \triangleq \sup_{i \in [K]} \mathbb{E}_{\boldsymbol{\theta}} [|T(Y)|] \leq \sqrt{\kappa_T^{(2)}} < \infty.$$

Then for every  $\ell$ ,

$$\mathbb{E}_{\boldsymbol{\theta}}[|T(Y_\ell)| \mid \mathcal{H}_{\ell-1}] = \sum_{i=1}^K \psi_{\ell,i} \mathbb{E}_{\boldsymbol{\theta}}[|T(Y)| \mid I_\ell = i] \leq \kappa_T^{(1)},$$

and therefore

$$\frac{1}{n} \sum_{\ell=1}^n \mathbb{E}_{\boldsymbol{\theta}}[|T(Y_\ell)| \mid \mathcal{H}_{\ell-1}] \leq \kappa_T^{(1)}. \quad (26)$$

Moreover, letting

$$U_\ell \triangleq |T(Y_\ell)| - \mathbb{E}_{\boldsymbol{\theta}}[|T(Y_\ell)| \mid \mathcal{H}_{\ell-1}], \quad S_n \triangleq \frac{1}{n} \sum_{\ell=1}^n U_\ell,$$

we have that  $\{U_\ell\}_{\ell \geq 1}$  is a martingale difference sequence and

$$\mathbb{E}_{\boldsymbol{\theta}}[U_\ell^2 \mid \mathcal{H}_{\ell-1}] \leq \mathbb{E}_{\boldsymbol{\theta}}[|T(Y_\ell)|^2 \mid \mathcal{H}_{\ell-1}] = \sum_{i=1}^K \psi_{\ell,i} \mathbb{E}_{\boldsymbol{\theta}}[T(Y)^2 \mid I_\ell = i] \leq \kappa_T^{(2)}.$$

Hence  $\sum_{\ell \geq 1} \mathbb{E}_{\boldsymbol{\theta}}[U_\ell^2]/\ell^2 < \infty$ , and by the martingale strong law,

$$S_n \rightarrow 0 \quad \text{a.s.} \quad (27)$$

Finally,

$$\frac{1}{n} \sum_{\ell=1}^n |T(Y_\ell)| = \frac{1}{n} \sum_{\ell=1}^n \mathbb{E}_{\boldsymbol{\theta}}[|T(Y_\ell)| \mid \mathcal{H}_{\ell-1}] + S_n \leq \kappa_T^{(1)} + S_n. \quad (28)$$

Plugging (26) and (28) into (25) gives

$$\frac{|M_n(\boldsymbol{\vartheta}) - M_n(\boldsymbol{\vartheta}')|}{n} \leq L \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\| (2 + 2\kappa_T^{(1)} + S_n) \quad \text{a.s.} \quad (29)$$

Fix  $\delta > 0$  and let  $\mathcal{N}_\delta$  be a finite  $\delta$ -net of  $\Theta$ . For each  $\boldsymbol{\vartheta} \in \Theta$ , choose  $\pi_\delta(\boldsymbol{\vartheta}) \in \mathcal{N}_\delta$  such that  $\|\boldsymbol{\vartheta} - \pi_\delta(\boldsymbol{\vartheta})\| \leq \delta$ . Then

$$\frac{|M_n(\boldsymbol{\vartheta})|}{n} \leq \frac{|M_n(\pi_\delta(\boldsymbol{\vartheta}))|}{n} + \frac{|M_n(\boldsymbol{\vartheta}) - M_n(\pi_\delta(\boldsymbol{\vartheta}))|}{n}.$$

Taking  $\sup_{\boldsymbol{\vartheta} \in \Theta}$  and using (29) yields

$$\begin{aligned} \sup_{\boldsymbol{\vartheta} \in \Theta} \frac{|M_n(\boldsymbol{\vartheta})|}{n} &\leq \max_{\boldsymbol{\nu} \in \mathcal{N}_\delta} \frac{|M_n(\boldsymbol{\nu})|}{n} + \sup_{\substack{\boldsymbol{\vartheta} \in \Theta: \\ \|\boldsymbol{\vartheta} - \pi_\delta(\boldsymbol{\vartheta})\| \leq \delta}} L \|\boldsymbol{\vartheta} - \pi_\delta(\boldsymbol{\vartheta})\| (2 + 2\kappa_T^{(1)} + |S_n|) \\ &\leq \max_{\boldsymbol{\nu} \in \mathcal{N}_\delta} \frac{|M_n(\boldsymbol{\nu})|}{n} + L\delta(2 + 2\kappa_T^{(1)}) + L\delta|S_n|. \end{aligned} \quad (30)$$

Now let  $n \rightarrow \infty$  in (30). Using (20) and  $S_n \rightarrow 0$  a.s. yields

$$\limsup_{n \rightarrow \infty} \sup_{\boldsymbol{\vartheta} \in \Theta} \frac{|M_n(\boldsymbol{\vartheta})|}{n} \leq L\delta(2 + 2\kappa_T^{(1)}) \quad \text{a.s.}$$

Since  $\delta > 0$  is arbitrary, letting  $\delta \rightarrow 0$  gives  $\sup_{\boldsymbol{\vartheta} \in \Theta} |M_n(\boldsymbol{\vartheta})|/n \rightarrow 0$  a.s. Therefore,

$$\sup_{\boldsymbol{\vartheta} \in \Theta} \left| \frac{1}{n} \Lambda_n(\boldsymbol{\theta} \|\boldsymbol{\vartheta}) - \Gamma(\bar{\psi}_n; \boldsymbol{\vartheta}) \right| = \sup_{\boldsymbol{\vartheta} \in \Theta} \frac{|M_n(\boldsymbol{\vartheta})|}{n} \rightarrow 0 \quad \text{a.s.},$$

which proves the lemma.  $\square$

**Lemma 4** (Uniform posterior density ratio). *Assume Assumptions 1–4. Then,*

$$\sup_{\vartheta \in \Theta} \left| \frac{1}{n} \log \frac{\pi_n(\vartheta)}{\pi_n(\boldsymbol{\theta})} + \Gamma(\bar{\psi}_n; \vartheta) \right| \rightarrow 0 \quad a.s.$$

*Proof.* By Bayes' rule,

$$\log \frac{\pi_n(\vartheta)}{\pi_n(\boldsymbol{\theta})} = \log \frac{\pi_0(\vartheta)}{\pi_0(\boldsymbol{\theta})} + \sum_{\ell=1}^n \log \frac{p_{\vartheta, I_\ell}(Y_\ell)}{p_{\boldsymbol{\theta}, I_\ell}(Y_\ell)} = \log \frac{\pi_0(\vartheta)}{\pi_0(\boldsymbol{\theta})} - \Lambda_n(\boldsymbol{\theta} \parallel \vartheta).$$

Divide by  $n$ , we obtain

$$\frac{1}{n} \log \frac{\pi_n(\vartheta)}{\pi_n(\boldsymbol{\theta})} + \Gamma(\bar{\psi}_n; \vartheta) = \frac{1}{n} \log \frac{\pi_0(\vartheta)}{\pi_0(\boldsymbol{\theta})} - \left( \frac{1}{n} \Lambda_n(\boldsymbol{\theta} \parallel \vartheta) - \Gamma(\bar{\psi}_n; \vartheta) \right).$$

The first term goes to 0 uniformly in  $\vartheta$  by Assumption 4 and the second term goes to 0 uniformly by Lemma 3.  $\square$

Next we need a mild regularity of  $\vartheta \mapsto \Gamma(\bar{\psi}_n; \vartheta)$ .

**Lemma 5** (Uniform continuity of  $\Gamma(\boldsymbol{q}; \vartheta)$  in  $\vartheta$ ). *Assume Assumptions 1 and 1. Then for every  $\epsilon > 0$  there exists  $\delta > 0$  such that for all  $\boldsymbol{q} \in \Delta_K$  and all  $\vartheta, \vartheta' \in \Theta$ ,*

$$\|\vartheta - \vartheta'\| \leq \delta \quad \Rightarrow \quad |\Gamma(\boldsymbol{q}; \vartheta) - \Gamma(\boldsymbol{q}; \vartheta')| \leq \epsilon.$$

*Proof.* For each arm  $i$ , continuity of  $\eta_i$  on compact  $\Theta$  implies uniform continuity. Recall that

$$\text{KL}(P_{\boldsymbol{\theta}, i} \parallel P_{\vartheta, i}) = (\eta_i(\boldsymbol{\theta}) - \eta_i(\vartheta)) A'(\eta_i(\boldsymbol{\theta})) - A(\eta_i(\boldsymbol{\theta})) + A(\eta_i(\vartheta)).$$

On compact  $\eta_i(\Theta)$ ,  $A$  is Lipschitz as  $A'$  is bounded by Assumption 1. Thus  $\vartheta \mapsto \text{KL}(P_{\boldsymbol{\theta}, i} \parallel P_{\vartheta, i})$  is uniformly continuous, uniformly in  $i$ . A convex combination over  $i$  with weights  $q_i$  preserves uniform continuity uniformly over  $\boldsymbol{q} \in \Delta_K$ .  $\square$

## A.2 Posterior Large Deviation

**Definition 1** (Logarithmic equivalence). *For positive sequences  $\{a_n\}$  and  $\{b_n\}$ , write  $a_n \doteq b_n$  if  $\lim_{n \rightarrow \infty} \frac{1}{n} \log(a_n/b_n) = 0$ . If the sequences are random, the relation holds almost surely.*

**Proposition 2** (Posterior large deviations). *Assume Assumptions 1–4. Let  $\tilde{\Theta} \subset \Theta$  be open. Then,*

$$\Pi_n(\tilde{\Theta}) \doteq \exp \left( -n \inf_{\vartheta \in \tilde{\Theta}} \Gamma(\bar{\psi}_n; \vartheta) \right). \quad (31)$$

*Proof.* Fix  $\tilde{\Theta}$  open. Write

$$\Pi_n(\tilde{\Theta}) = \frac{\int_{\tilde{\Theta}} \pi_n(\vartheta) d\vartheta}{\int_{\Theta} \pi_n(\vartheta) d\vartheta} = \frac{\int_{\tilde{\Theta}} \frac{\pi_n(\vartheta)}{\pi_n(\boldsymbol{\theta})} d\vartheta}{\int_{\Theta} \frac{\pi_n(\vartheta)}{\pi_n(\boldsymbol{\theta})} d\vartheta}.$$

By Lemma 4, for some deterministic  $\epsilon_n \downarrow 0$ ,

$$\exp \{-n(\Gamma(\bar{\psi}_n; \vartheta) + \epsilon_n)\} \leq \frac{\pi_n(\vartheta)}{\pi_n(\boldsymbol{\theta})} \leq \exp \{-n(\Gamma(\bar{\psi}_n; \vartheta) - \epsilon_n)\} \quad \forall \vartheta \in \Theta.$$

Integrating over  $\tilde{\Theta}$  and  $\Theta$  yields that the numerator and denominator are each log-equivalent to the corresponding Laplace integrals

$$\int_{\tilde{\Theta}} \exp\{-n\Gamma(\bar{\psi}_n; \boldsymbol{\vartheta})\} d\boldsymbol{\vartheta} \quad \text{and} \quad \int_{\Theta} \exp\{-n\Gamma(\bar{\psi}_n; \boldsymbol{\vartheta})\} d\boldsymbol{\vartheta}.$$

Thus it suffices to show the Laplace principle

$$\int_{\tilde{\Theta}} \exp\{-nW_n(\boldsymbol{\vartheta})\} d\boldsymbol{\vartheta} \doteq \exp\{-n \inf_{\boldsymbol{\vartheta} \in \tilde{\Theta}} W_n(\boldsymbol{\vartheta})\} \quad \text{with} \quad W_n(\boldsymbol{\vartheta}) = \Gamma(\bar{\psi}_n; \boldsymbol{\vartheta}),$$

and similarly with  $\tilde{\Theta} = \Theta$ .

Let  $\hat{\boldsymbol{\vartheta}}_n \in \text{cl}(\tilde{\Theta})$  attain the minimum  $W_n(\hat{\boldsymbol{\vartheta}}_n) = \inf_{\boldsymbol{\vartheta} \in \tilde{\Theta}} W_n(\boldsymbol{\vartheta})$ , which exists because  $W_n$  is continuous (Lemma 5) and  $\text{cl}(\tilde{\Theta})$  is compact. Define  $\gamma_n = \int_{\tilde{\Theta}} \exp(-nW_n(\boldsymbol{\vartheta})) d\boldsymbol{\vartheta}$ . Then

$$\gamma_n \leq \text{Vol}(\tilde{\Theta}) \exp\{-nW_n(\hat{\boldsymbol{\vartheta}}_n)\},$$

so  $\limsup_{n \rightarrow \infty} \frac{1}{n} \log \gamma_n + W_n(\hat{\boldsymbol{\vartheta}}_n) \leq 0$ .

For the reverse bound, fix  $\epsilon > 0$ . By uniform continuity of  $W_n$  (Lemma 5), there exists  $\delta > 0$  such that  $\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}'\| \leq \delta$  implies  $|W_n(\boldsymbol{\vartheta}) - W_n(\boldsymbol{\vartheta}')| \leq \epsilon$  for all  $n$ . Because  $\tilde{\Theta}$  is open, for each  $n$  we can choose a point  $\boldsymbol{\vartheta}_n^\circ \in \tilde{\Theta}$  arbitrarily close to  $\hat{\boldsymbol{\vartheta}}_n$ . Then the ball  $B(\boldsymbol{\vartheta}_n^\circ, \delta/2)$  has positive intersection with  $\tilde{\Theta}$  and volume bounded below by some  $c_\delta > 0$  uniformly in  $n$  (compactness plus finiteness of a cover argument). On that intersection,  $W_n(\boldsymbol{\vartheta}) \leq W_n(\hat{\boldsymbol{\vartheta}}_n) + \epsilon$ . Therefore

$$\gamma_n \geq \int_{\tilde{\Theta} \cap B(\boldsymbol{\vartheta}_n^\circ, \delta/2)} \exp\{-nW_n(\boldsymbol{\vartheta})\} d\boldsymbol{\vartheta} \geq c_\delta \exp\{-n(W_n(\hat{\boldsymbol{\vartheta}}_n) + \epsilon)\}.$$

Taking logs and dividing by  $n$  gives

$$\frac{1}{n} \log \gamma_n + W_n(\hat{\boldsymbol{\vartheta}}_n) \geq \frac{1}{n} \log c_\delta - \epsilon \rightarrow -\epsilon.$$

Since  $\epsilon$  is arbitrary,  $\liminf_{n \rightarrow \infty} \frac{1}{n} \log \gamma_n + W_n(\hat{\boldsymbol{\vartheta}}_n) \geq 0$ . This establishes the Laplace principle for  $\tilde{\Theta}$ .

For  $\Theta$  itself, note that  $\inf_{\boldsymbol{\vartheta} \in \Theta} W_n(\boldsymbol{\vartheta}) = 0$  because  $W_n(\boldsymbol{\vartheta}) = 0$  and  $W_n \geq 0$ . Thus the denominator integral is log-equivalent to  $\exp(0) = 1$  and cancels. This yields (31).  $\square$

### A.3 From Sample-Normalized to Cost-Normalized Exponents

**Lemma 6.** *For each arm  $i$ , we have  $\lim_{n \rightarrow \infty} (p_{n,i} - \bar{\psi}_{n,i}) = 0$  a.s. Consequently,  $\|\mathbf{p}_n - \bar{\psi}_n\|_1 \rightarrow 0$  a.s.*

*Proof.* Let  $X_\ell = \mathbf{1}\{I_\ell = i\}$  and  $Z_\ell = \mathbb{E}[X_\ell \mid \mathcal{H}_{\ell-1}] = \psi_{\ell,i}$ . Then  $M_n = \sum_{\ell=1}^n (X_\ell - Z_\ell)$  is a martingale with bounded increments, hence  $M_n/n \rightarrow 0$  a.s.  $\square$

**Proposition 3** (Cost-normalized posterior exponent). *Assume Assumptions 1–4. Then for any open  $\tilde{\Theta} \subset \Theta$ ,*

$$-\frac{1}{B_n} \log \Pi_n(\tilde{\Theta}) \doteq \inf_{\boldsymbol{\vartheta} \in \tilde{\Theta}} \frac{\Gamma(\mathbf{p}_n; \boldsymbol{\vartheta})}{\bar{C}(\mathbf{p}_n)} \quad \text{a.s.}$$

*Proof.* From Proposition 2,

$$-\frac{1}{n} \log \Pi_n(\tilde{\Theta}) \doteq \inf_{\vartheta \in \tilde{\Theta}} \Gamma(\bar{\psi}_n; \vartheta). \quad (32)$$

For any  $\vartheta \in \Theta$ ,

$$\Gamma(\mathbf{p}_n; \vartheta) - \Gamma(\bar{\psi}_n; \vartheta) = \sum_{i=1}^K (p_{n,i} - \bar{\psi}_{n,i}) \text{KL}(P_{\theta,i} \| P_{\vartheta,i}),$$

hence by (19),

$$\sup_{\vartheta \in \Theta} |\Gamma(\mathbf{p}_n; \vartheta) - \Gamma(\bar{\psi}_n; \vartheta)| \leq \kappa_{\text{KL}} \|\mathbf{p}_n - \bar{\psi}_n\|_1 \longrightarrow 0 \quad \text{a.s.},$$

where the convergence is Lemma 6. Therefore,

$$\left| \inf_{\vartheta \in \tilde{\Theta}} \Gamma(\mathbf{p}_n; \vartheta) - \inf_{\vartheta \in \tilde{\Theta}} \Gamma(\bar{\psi}_n; \vartheta) \right| \leq \sup_{\vartheta \in \Theta} |\Gamma(\mathbf{p}_n; \vartheta) - \Gamma(\bar{\psi}_n; \vartheta)| \longrightarrow 0 \quad \text{a.s.} \quad (33)$$

Combining (32) and (33) yields

$$\Pi_n(\tilde{\Theta}) \doteq \exp \left\{ -n \inf_{\vartheta \in \tilde{\Theta}} \Gamma(\mathbf{p}_n; \vartheta) \right\} \quad \text{a.s.},$$

Multiply by  $n/B_n = 1/\bar{C}(\mathbf{p}_n)$  to obtain

$$-\frac{1}{B_n} \log \Pi_n(\tilde{\Theta}) \doteq \inf_{\vartheta \in \tilde{\Theta}} \frac{\Gamma(\mathbf{p}_n; \vartheta)}{\bar{C}(\mathbf{p}_n)},$$

using that multiplying by deterministic sequences bounded away from 0 and  $\infty$  preserves log-equivalence. The second statement follows if the right-hand side converges.  $\square$

#### A.4 Completing the Proof

We are now ready to prove Theorem 1.

*Proof.* We prove the upper bound first. Fix  $\epsilon > 0$  and let  $\tilde{\Theta}_\epsilon$  be the open superset of  $\text{Alt}(\boldsymbol{\theta})$  from Assumption 3. Then  $\Pi_n(\text{Alt}(\boldsymbol{\theta})) \leq \Pi_n(\tilde{\Theta}_\epsilon)$ , hence

$$-\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) \geq -\frac{1}{B_n} \log \Pi_n(\tilde{\Theta}_\epsilon).$$

This direction is not useful for an upper bound; instead we use  $\Pi_n(\text{Alt}(\boldsymbol{\theta})) \geq \Pi_n(\tilde{\Theta})$  for any open  $\tilde{\Theta} \subset \text{Alt}(\boldsymbol{\theta})$ . Let  $\tilde{\Theta}$  be any open subset of  $\text{Alt}(\boldsymbol{\theta})$  (e.g. an interior approximation). Then Proposition 3 yields

$$-\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) \leq -\frac{1}{B_n} \log \Pi_n(\tilde{\Theta}) \doteq \inf_{\vartheta \in \tilde{\Theta}} \frac{\Gamma(\mathbf{p}_n; \vartheta)}{\bar{C}(\mathbf{p}_n)}.$$

Taking  $\limsup$  in  $n$  and using  $\tilde{\Theta} \subset \text{Alt}(\boldsymbol{\theta})$  gives

$$\limsup_{n \rightarrow \infty} -\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) \leq \limsup_{n \rightarrow \infty} \inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \frac{\Gamma(\mathbf{p}_n; \vartheta)}{\bar{C}(\mathbf{p}_n)}.$$

Finally, since for each  $n$ ,  $\mathbf{p}_n \in \Delta_K$ ,

$$\inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \frac{\Gamma(\mathbf{p}_n; \vartheta)}{\bar{C}(\mathbf{p}_n)} \leq \sup_{\mathbf{p} \in \Delta_K} \inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \frac{\Gamma(\mathbf{p}; \vartheta)}{\bar{C}(\mathbf{p})} = \Gamma^*.$$

Thus the  $\limsup$  is bounded by  $\Gamma^*$ .

For sufficiency, recall that

$$\inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \frac{\Gamma(\mathbf{p}_n; \vartheta)}{\bar{C}(\mathbf{p}_n)} = \inf_{\vartheta \in \text{Alt}(\boldsymbol{\theta})} \Gamma^c(\mathbf{p}_n; \vartheta),$$

Assume (3) holds, then so the right-hand side above converges to  $\Gamma^*$  a.s. Now apply Proposition 3 to the (open) sets  $\tilde{\Theta}_\epsilon$  that approximate  $\text{Alt}(\boldsymbol{\theta})$ . Because  $\text{Alt}(\boldsymbol{\theta}) \subset \tilde{\Theta}_\epsilon$ ,

$$-\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) \geq -\frac{1}{B_n} \log \Pi_n(\tilde{\Theta}_\epsilon) \doteq \inf_{\vartheta \in \tilde{\Theta}_\epsilon} \frac{\Gamma(\bar{\psi}_n; \vartheta)}{\bar{C}(\mathbf{p}_n)}.$$

Taking  $\liminf$  in  $n$  and then letting  $\epsilon \downarrow 0$ , we obtain

$$\liminf_{n \rightarrow \infty} -\frac{1}{B_n} \log \Pi_n(\text{Alt}(\boldsymbol{\theta})) \geq \lim_{\epsilon \downarrow 0} \liminf_{n \rightarrow \infty} \inf_{\vartheta \in \tilde{\Theta}_\epsilon} \frac{\Gamma(\bar{\psi}_n; \vartheta)}{\bar{C}(\mathbf{p}_n)} = \Gamma^*,$$

where the last equality uses that  $\tilde{\Theta}_\epsilon$  shrinks to  $\text{Alt}(\boldsymbol{\theta})$  and the objective is continuous in  $\vartheta$  by Lemma 5 and in  $\mathbf{p}$ ; thus the infimum over  $\tilde{\Theta}_\epsilon$  converges to the infimum over  $\text{Alt}(\boldsymbol{\theta})$ .

Combine this lower bound with the upper bound conclude that the limit exists and equals  $\Gamma^*$ .  $\square$

## B Proof of Lemma 1

Fix  $\boldsymbol{\theta}$  and  $x \in \mathcal{X}(\boldsymbol{\theta})$ . Recall

$$\tilde{\Gamma}^c(\mathbf{w}, \vartheta) = \sum_{i=1}^K w_i \frac{\text{KL}(P_{\boldsymbol{\theta},i} \| P_{\vartheta,i})}{C_i(\boldsymbol{\theta})}, \quad D_x^w(\mathbf{w}; \boldsymbol{\theta}) = \inf_{\vartheta \in \text{Alt}_x(\boldsymbol{\theta})} \tilde{\Gamma}^c(\mathbf{w}, \vartheta).$$

Since  $\Theta$  is compact (Assumption 1), the feasible set  $\text{cl}(\text{Alt}_x(\boldsymbol{\theta})) \subset \Theta$  is compact. For each  $i$ , the map  $\vartheta \mapsto \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\vartheta,i})$  is continuous on  $\Theta$  (one-dimensional canonical exponential family with continuous  $\eta_i$  and  $A \in C^2$  on a neighborhood of  $\eta_i(\Theta)$ ). Hence  $\vartheta \mapsto \tilde{\Gamma}^c(\mathbf{w}, \vartheta)$  is continuous on  $\text{cl}(\text{Alt}_x(\boldsymbol{\theta}))$  for each fixed  $\mathbf{w}$ , and therefore

$$D_x^w(\mathbf{w}; \boldsymbol{\theta}) = \min_{\vartheta \in \text{cl}(\text{Alt}_x(\boldsymbol{\theta}))} \tilde{\Gamma}^c(\mathbf{w}, \vartheta).$$

**Shape on  $\mathbb{R}_{\geq 0}^K$ .** Fix any  $\vartheta$ . The map  $\mathbf{w} \mapsto \tilde{\Gamma}^c(\mathbf{w}, \vartheta)$  is linear with nonnegative coefficients (since  $\text{KL} \geq 0$  and  $C_i(\boldsymbol{\theta}) \geq c_{\min} > 0$  by Assumption 2). Thus it is concave, continuous, coordinatewise nondecreasing, and degree-one homogeneous on  $\mathbb{R}_{\geq 0}^K$ . Taking the minimum over  $\vartheta \in \text{cl}(\text{Alt}_x(\boldsymbol{\theta}))$  preserves concavity, monotonicity, homogeneity, and nonnegativity, so  $D_x^w(\cdot; \boldsymbol{\theta})$  has the claimed properties.

**Unique minimizer and  $C^1$  smoothness on  $\mathbb{R}_{>0}^K$ .** Fix  $\mathbf{w} \in \mathbb{R}_{>0}^K$  so that  $w_i > 0$  for all  $i$ . Existence of a minimizer follows from compactness of  $\text{cl}(\text{Alt}_x(\boldsymbol{\theta}))$ .

For uniqueness, note that for a one-dimensional canonical exponential family the KL divergence is a Bregman divergence of  $A$  and is strictly convex in the natural parameter; under our structured models (e.g., linear bandits where  $\eta_i$  is affine in  $\boldsymbol{\vartheta}$ ), each map  $\boldsymbol{\vartheta} \mapsto \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i})$  is convex and not affine on any nontrivial segment. Since  $w_i/C_i(\boldsymbol{\theta}) > 0$  for all  $i$  and  $\text{cl}(\text{Alt}_x(\boldsymbol{\theta}))$  is convex (Assumption 5), the weighted sum  $\boldsymbol{\vartheta} \mapsto \tilde{\Gamma}^c(\mathbf{w}, \boldsymbol{\vartheta})$  is strictly convex on  $\text{cl}(\text{Alt}_x(\boldsymbol{\theta}))$ , hence the minimizer is unique; denote it by  $\boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta})$ .

By Danskin's theorem,  $D_x^w(\cdot; \boldsymbol{\theta})$  is differentiable at  $\mathbf{w}$  and

$$[\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})]_i = \frac{\partial}{\partial w_i} \tilde{\Gamma}^c(\mathbf{w}, \boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta})) = \frac{\text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta}),i})}{C_i(\boldsymbol{\theta})}, \quad i \in [K].$$

Moreover, the argmin map  $\mathbf{w} \mapsto \boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta})$  is continuous on  $\mathbb{R}_{>0}^K$  by Berge's maximum theorem (constant compact feasible set and unique minimizer), hence the displayed gradient is continuous in  $\mathbf{w}$ . Therefore  $D_x^w(\cdot; \boldsymbol{\theta}) \in C^1(\mathbb{R}_{>0}^K)$ .

**Gradient lower bound and strict positivity on  $\text{int}(\Delta_K)$ .** Assume  $D_x^w(\cdot; \boldsymbol{\theta}) \not\equiv 0$  on  $\Delta_K$ . Consider the continuous function on the compact set  $\text{cl}(\text{Alt}_x(\boldsymbol{\theta}))$ ,

$$\boldsymbol{\vartheta} \mapsto \max_{i \in [K]} \frac{\text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i})}{c_{\max}}.$$

If this maximum equals 0 at some  $\boldsymbol{\vartheta}$ , then  $\text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i}) = 0$  for all  $i$ , hence  $P_{\boldsymbol{\theta},i} = P_{\boldsymbol{\vartheta},i}$  for all  $i$ . By the identifiability condition in Assumption 3, this implies  $\mathcal{I}(\boldsymbol{\vartheta}) = \mathcal{I}(\boldsymbol{\theta})$ , contradicting  $\boldsymbol{\vartheta} \in \text{cl}(\text{Alt}_x(\boldsymbol{\theta})) \subset \text{Alt}(\boldsymbol{\theta})$ . Therefore the maximum is strictly positive on  $\text{cl}(\text{Alt}_x(\boldsymbol{\theta}))$ , and by compactness,

$$d_x(\boldsymbol{\theta}) \triangleq \min_{\boldsymbol{\vartheta} \in \text{cl}(\text{Alt}_x(\boldsymbol{\theta}))} \max_{i \in [K]} \frac{\text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta},i})}{c_{\max}} > 0.$$

Now fix any  $\mathbf{w} \in \text{int}(\Delta_K)$ . Using the gradient formula and  $C_i(\boldsymbol{\theta}) \leq c_{\max}$ ,

$$\sum_{i=1}^K [\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})]_i = \sum_{i=1}^K \frac{\text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta}),i})}{C_i(\boldsymbol{\theta})} \geq \frac{1}{c_{\max}} \max_{i \in [K]} \text{KL}(P_{\boldsymbol{\theta},i} \| P_{\boldsymbol{\vartheta}_x(\mathbf{w}; \boldsymbol{\theta}),i}) \geq d_x(\boldsymbol{\theta}).$$

Finally, since  $D_x^w(\cdot; \boldsymbol{\theta})$  is degree-one homogeneous and differentiable at  $\mathbf{w} \in \text{int}(\Delta_K)$ , Euler's identity yields

$$D_x^w(\mathbf{w}; \boldsymbol{\theta}) = \langle \mathbf{w}, \nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta}) \rangle \geq \left( \min_{i \in [K]} w_i \right) \sum_{i=1}^K [\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})]_i \geq \left( \min_{i \in [K]} w_i \right) d_x(\boldsymbol{\theta}) > 0,$$

so  $D_x^w(\cdot; \boldsymbol{\theta})$  is strictly positive on  $\text{int}(\Delta_K)$ .

## C Proof of Lemma 2

Detection consistency follows directly from (12), the definition (10) of  $D_x^w$  and the continuity of  $C_i(\cdot)$  (assumption 2). For selection consistency on non-forced rounds, note that  $q_{n,i} = H_i^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n)$  and use the relation (14),

$$\begin{aligned}\tilde{h}_{n,i} &= \frac{q_{n,i} \frac{C_i(\boldsymbol{\theta})}{B_n + C_i(\boldsymbol{\theta})}}{\sum_{j=1}^K q_{n,j} \frac{C_j(\boldsymbol{\theta})}{B_n + C_j(\boldsymbol{\theta})}} = \frac{H_i^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n) \cdot \frac{C_i(\boldsymbol{\theta})}{B_n + C_j(\boldsymbol{\theta})}}{\sum_{j=1}^K H_j^{x_n}(\mathbf{p}_n; \boldsymbol{\theta}_n) \cdot \frac{C_j(\boldsymbol{\theta})}{B_n + C_j(\boldsymbol{\theta})}} = \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot \frac{1}{B_n + C_j(\boldsymbol{\theta})}}{\sum_{j=1}^K h_j^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot \frac{1}{B_n + C_j(\boldsymbol{\theta})}} \\ &= \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot \frac{B_n}{B_n + C_i(\boldsymbol{\theta})}}{\sum_{j=1}^K h_j^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot \frac{B_n}{B_n + C_j(\boldsymbol{\theta})}} = \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot (1 + x_i)^{-1}}{\sum_{j=1}^K h_j^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot (1 + x_j)^{-1}}\end{aligned}$$

where  $x_i \triangleq C_i(\boldsymbol{\theta})/B_n \in (\frac{c_{\min}}{nc_{\max}}, \frac{c_{\max}}{nc_{\min}})$  then by  $1 - x \leq (1 + x)^{-1} \leq 1$ ,

$$\begin{aligned}\tilde{h}_{n,i} &\geq \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot (1 - x_i)}{\sum h_j^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)} \geq (1 - x_i) h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \geq \left(1 - \frac{c_{\max}}{nc_{\min}}\right) h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n), \\ \tilde{h}_{n,i} &\leq \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)}{\sum h_j^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \cdot (1 + x_j)^{-1}} \leq \left(1 + \frac{c_{\max}}{nc_{\min}}\right) h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n).\end{aligned}$$

Thus

$$\left| \tilde{h}_{n,i} - h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \right| \leq \frac{c_{\max}}{nc_{\min}} h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) \leq \frac{c_{\max}}{nc_{\min}} \rightarrow 0.$$

## D Proof of Proposition 1

**Properties of  $\det$ .** Fix  $\mathbf{w} \in \Delta_K$ . Since  $F(\mathbf{w}, \cdot)$  is linear on the compact convex set  $\Delta_{|\mathcal{X}|}$ , the minimizer set  $\det(\mathbf{w}) = \arg \min_{\boldsymbol{\nu} \in \Delta_{|\mathcal{X}|}} F(\mathbf{w}, \boldsymbol{\nu})$  is nonempty and is a face of  $\Delta_{|\mathcal{X}|}$ , hence convex and compact. Because  $\mathcal{X}$  is finite and each  $D_x^w(\mathbf{w})$  is continuous in  $\mathbf{w}$ , the function  $F(\mathbf{w}, \boldsymbol{\nu}) = \sum_{x \in \mathcal{X}} \nu_x D_x^w(\mathbf{w})$  is jointly continuous on  $\Delta_K \times \Delta_{|\mathcal{X}|}$ . By Berge's maximum theorem,  $\det$  is upper hemicontinuous and compact-valued; in particular, its graph is closed.

**The limit-set  $S(\mathbf{w})$ .** For each  $\mathbf{w} \in \Delta_K$ , define

$$S(\mathbf{w}) \triangleq \left\{ \lim_{n \rightarrow \infty} \mathbf{h}(\mathbf{w}^n, \boldsymbol{\nu}^n) : \mathbf{w}^n \rightarrow \mathbf{w}, \mathbf{w}^n \in \mathcal{R}, \boldsymbol{\nu}^n \in \det(\mathbf{w}^n) \right\} \subseteq \Delta_K.$$

*Nonemptiness.* Pick any sequence  $\mathbf{w}^n \in \mathcal{R}$  with  $\mathbf{w}^n \rightarrow \mathbf{w}$ . For each  $n$ , choose  $\boldsymbol{\nu}^n \in \det(\mathbf{w}^n)$ . Then  $\mathbf{h}(\mathbf{w}^n, \boldsymbol{\nu}^n) \in \Delta_K$ , and by compactness of  $\Delta_K$  the sequence admits a convergent subsequence. Hence  $S(\mathbf{w}) \neq \emptyset$ .

*Compactness.* Since  $S(\mathbf{w}) \subseteq \Delta_K$  and  $\Delta_K$  is compact, it suffices to show that  $S(\mathbf{w})$  is closed. Take any sequence  $\mathbf{s}^m \in S(\mathbf{w})$  with  $\mathbf{s}^m \rightarrow \mathbf{s}$ . For each  $m$ , by definition of  $S(\mathbf{w})$  there exist sequences  $\{\mathbf{w}^{m,n}\}_{n \geq 1} \subset \mathcal{R}$  and  $\{\boldsymbol{\nu}^{m,n}\}_{n \geq 1}$  such that

$$\mathbf{w}^{m,n} \rightarrow \mathbf{w}, \quad \boldsymbol{\nu}^{m,n} \in \det(\mathbf{w}^{m,n}), \quad \mathbf{h}(\mathbf{w}^{m,n}, \boldsymbol{\nu}^{m,n}) \rightarrow \mathbf{s}^m \quad \text{as } n \rightarrow \infty.$$

Choose an  $\varepsilon$ -selection: pick  $n(m)$  large enough so that

$$\|\mathbf{w}^{m,n(m)} - \mathbf{w}\| \leq \frac{1}{m}, \quad \|\mathbf{h}(\mathbf{w}^{m,n(m)}, \boldsymbol{\nu}^{m,n(m)}) - \mathbf{s}^m\| \leq \frac{1}{m}.$$

Set  $\tilde{\mathbf{w}}^m := \mathbf{w}^{m,n(m)} \in \mathcal{R}$ ,  $\tilde{\boldsymbol{\nu}}^m := \boldsymbol{\nu}^{m,n(m)} \in \det(\tilde{\mathbf{w}}^m)$ , and  $\tilde{\mathbf{s}}^m := \mathbf{h}(\tilde{\mathbf{w}}^m, \tilde{\boldsymbol{\nu}}^m) \in \Delta_K$ . Then  $\tilde{\mathbf{w}}^m \rightarrow \mathbf{w}$  and  $\|\tilde{\mathbf{s}}^m - \mathbf{s}^m\| \leq 1/m$ , hence  $\tilde{\mathbf{s}}^m \rightarrow \mathbf{s}$ .

By compactness of  $\Delta_{|\mathcal{X}|}$ , along a subsequence (not relabeled)  $\tilde{\boldsymbol{\nu}}^m \rightarrow \boldsymbol{\nu}$ . Since  $\tilde{\mathbf{w}}^m \rightarrow \mathbf{w}$  and  $\tilde{\boldsymbol{\nu}}^m \in \det(\tilde{\mathbf{w}}^m)$  with  $\tilde{\boldsymbol{\nu}}^m \rightarrow \boldsymbol{\nu}$ , the closed-graph property of  $\det$  implies  $\boldsymbol{\nu} \in \det(\mathbf{w})$ . Moreover, by construction,

$$\tilde{\mathbf{s}}^m = \mathbf{h}(\tilde{\mathbf{w}}^m, \tilde{\boldsymbol{\nu}}^m) \rightarrow \mathbf{s} \quad \text{with} \quad \tilde{\mathbf{w}}^m \in \mathcal{R}, \quad \tilde{\mathbf{w}}^m \rightarrow \mathbf{w}, \quad \tilde{\boldsymbol{\nu}}^m \in \det(\tilde{\mathbf{w}}^m),$$

so  $\mathbf{s} \in S(\mathbf{w})$ . Thus  $S(\mathbf{w})$  is closed, hence compact.

**Upper hemicontinuity of  $S(\cdot)$ .** We show  $\text{graph}(S)$  is closed. Let  $\mathbf{w}^m \rightarrow \mathbf{w}$  and take  $\mathbf{s}^m \in S(\mathbf{w}^m)$  with  $\mathbf{s}^m \rightarrow \mathbf{s}$ . For each  $m$ , pick sequences  $\{\mathbf{w}^{m,n}\}_{n \geq 1} \subset \mathcal{R}$  and  $\{\boldsymbol{\nu}^{m,n}\}_{n \geq 1}$  such that

$$\mathbf{w}^{m,n} \rightarrow \mathbf{w}^m, \quad \boldsymbol{\nu}^{m,n} \in \det(\mathbf{w}^{m,n}), \quad \mathbf{h}(\mathbf{w}^{m,n}, \boldsymbol{\nu}^{m,n}) \rightarrow \mathbf{s}^m \quad (n \rightarrow \infty).$$

As above, choose  $n(m)$  so that

$$\|\mathbf{w}^{m,n(m)} - \mathbf{w}^m\| \leq \frac{1}{m}, \quad \|\mathbf{h}(\mathbf{w}^{m,n(m)}, \boldsymbol{\nu}^{m,n(m)}) - \mathbf{s}^m\| \leq \frac{1}{m}.$$

Define  $\tilde{\mathbf{w}}^m := \mathbf{w}^{m,n(m)} \in \mathcal{R}$ ,  $\tilde{\boldsymbol{\nu}}^m := \boldsymbol{\nu}^{m,n(m)} \in \det(\tilde{\mathbf{w}}^m)$ , and  $\tilde{\mathbf{s}}^m := \mathbf{h}(\tilde{\mathbf{w}}^m, \tilde{\boldsymbol{\nu}}^m) \in \Delta_K$ . Then

$$\|\tilde{\mathbf{w}}^m - \mathbf{w}^m\| \leq \|\tilde{\mathbf{w}}^m - \mathbf{w}^m\| + \|\mathbf{w}^m - \mathbf{w}\| \rightarrow 0, \quad \|\tilde{\mathbf{s}}^m - \mathbf{s}^m\| \leq \|\tilde{\mathbf{s}}^m - \mathbf{s}^m\| + \|\mathbf{s}^m - \mathbf{s}\| \rightarrow 0,$$

so  $\tilde{\mathbf{w}}^m \rightarrow \mathbf{w}$  and  $\tilde{\mathbf{s}}^m \rightarrow \mathbf{s}$ . By compactness of  $\Delta_{|\mathcal{X}|}$ , along a subsequence  $\tilde{\boldsymbol{\nu}}^m \rightarrow \boldsymbol{\nu}$ . Using again the closed graph of  $\det$ , we get  $\boldsymbol{\nu} \in \det(\mathbf{w})$ . Therefore  $\mathbf{s} = \lim_m \mathbf{h}(\tilde{\mathbf{w}}^m, \tilde{\boldsymbol{\nu}}^m) \in S(\mathbf{w})$ , proving the closed-graph property. Since  $S$  is compact-valued (Step 2), it follows that  $S$  is upper hemicontinuous.

**The correspondence  $\mathbf{sel}$  inherits regularity.** By definition,  $\mathbf{sel}(\mathbf{w}) = \text{conv}(S(\mathbf{w}))$ . Since  $S(\mathbf{w})$  is nonempty and compact,  $\mathbf{sel}(\mathbf{w})$  is nonempty, convex, and compact.

It remains to show u.h.c. of  $\mathbf{sel}$ . We use a closed-graph argument and Carathéodory. Let  $\mathbf{w}^m \rightarrow \mathbf{w}$  and  $\mathbf{s}^m \in \mathbf{sel}(\mathbf{w}^m)$  with  $\mathbf{s}^m \rightarrow \mathbf{s}$ . By Carathéodory in  $\mathbb{R}^K$ , for each  $m$  we may write

$$\mathbf{s}^m = \sum_{j=1}^K \lambda^{m,j} \mathbf{x}^{m,j}, \quad \lambda^{m,j} \geq 0, \quad \sum_{j=1}^K \lambda^{m,j} = 1, \quad \mathbf{x}^{m,j} \in S(\mathbf{w}^m).$$

By compactness of  $\Delta_K$  and  $\Delta_K$  (for the coefficients), after passing to a subsequence we may assume

$$\lambda^{m,j} \rightarrow \lambda^j, \quad \mathbf{x}^{m,j} \rightarrow \mathbf{x}^j \quad \text{for each } j = 1, \dots, K.$$

Since  $S$  is u.h.c. with compact values, its graph is closed; thus  $\mathbf{x}^{m,j} \in S(\mathbf{w}^m)$  and  $(\mathbf{w}^m, \mathbf{x}^{m,j}) \rightarrow (\mathbf{w}, \mathbf{x}^j)$  imply  $\mathbf{x}^j \in S(\mathbf{w})$ . Taking limits in the convex combination yields

$$\mathbf{s} = \lim_{m \rightarrow \infty} \mathbf{s}^m = \sum_{j=1}^K \lambda^j \mathbf{x}^j \in \text{conv}(S(\mathbf{w})) = \mathbf{sel}(\mathbf{w}).$$

Hence  $\text{graph}(\mathbf{sel})$  is closed. Because  $\mathbf{sel}$  is compact-valued, it is upper hemicontinuous.

**Existence of DI solutions and forward invariance.** Because  $\text{sel}$  is nonempty, convex, compact-valued, and u.h.c., the same holds for  $G(\mathbf{w}) = \text{sel}(\mathbf{w}) - \mathbf{w}$ . Also, for any  $\mathbf{w} \in \Delta_K$  and  $\mathbf{g} \in G(\mathbf{w})$ , we can write  $\mathbf{g} = \mathbf{s} - \mathbf{w}$  with  $\mathbf{s} \in \text{sel}(\mathbf{w}) \subseteq \Delta_K$ , hence  $\|\mathbf{g}\|_1 \leq 2$ ; thus  $G$  is bounded. Standard existence theorems for differential inclusions with u.h.c. nonempty convex compact right-hand sides (e.g. Aubin and Cellina 1984) yield an absolutely continuous solution to (18) for any initial condition  $\mathbf{w}_0 \in \Delta_K$ .

Moreover,  $\Delta_K$  is forward invariant: if  $\dot{\mathbf{w}}(t) = \mathbf{s}(t) - \mathbf{w}(t)$  with  $\mathbf{s}(t) \in \text{sel}(\mathbf{w}(t)) \subseteq \Delta_K$  a.e., then  $\sum_{i=1}^K \dot{w}_i(t) = 1 - 1 = 0$  a.e., and if  $p_i(t) = 0$  then  $\dot{w}_i(t) = s_i(t) \geq 0$  a.e. Hence  $\mathbf{w}(t) \in \Delta_K$  for all  $t \geq 0$ .

Finally, for any absolutely continuous solution  $\mathbf{w}(\cdot)$ , define  $\mathbf{s}(t) := \dot{\mathbf{w}}(t) + \mathbf{w}(t)$  for a.e.  $t$ . Then  $\mathbf{s}(\cdot)$  is measurable and satisfies  $\mathbf{s}(t) \in \text{sel}(\mathbf{w}(t))$  for a.e.  $t$ .

## E Proof of Theorem 3

### E.1 Continuous-Time Dynamics

**Lemma 7.** *Under Assumption 7, the function  $\mathbf{w} \mapsto \min_{x \in \mathcal{X}} D_x^w(\mathbf{w})$  is  $M$ -Lipschitz on  $\Delta_K$  with respect to  $\|\cdot\|_1$ , i.e.,*

$$\left| \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_1) - \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_2) \right| \leq M \|\mathbf{w}_1 - \mathbf{w}_2\|_1, \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \Delta_K.$$

*Proof.* By Assumption 7, each  $D_x$  is  $M$ -Lipschitz on  $\Delta_K$  with respect to  $\|\cdot\|_1$ . Fix  $\mathbf{w}_1, \mathbf{w}_2 \in \Delta_K$  and choose  $x_2 \in \arg \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_2)$ . Then

$$\min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_1) - \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_2) \leq D_{x_2}^w(\mathbf{w}_1) - D_{x_2}^w(\mathbf{w}_2) \leq M \|\mathbf{w}_1 - \mathbf{w}_2\|_1.$$

Swapping  $\mathbf{w}_1$  and  $\mathbf{w}_2$  yields the reverse inequality, proving the claim.  $\square$

**Lemma 8** (Chain rule for  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w})$ ). *The information function  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w})$  admits the Clarke chain rule: for any absolutely continuous curve  $\mathbf{p}(t) : [0, \infty) \rightarrow \Delta_K$ ,  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t))$  is differentiable for almost every  $t \in [0, \infty)$  and*

$$\frac{d}{dt} \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t)) = \mathbf{f}(\mathbf{w}(t))^\top \frac{d}{dt} \mathbf{w}(t) \quad \text{for every } \mathbf{f}(\mathbf{w}(t)) \in \partial^* \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t)), \quad (34)$$

for almost every  $t$ .

*Proof.* Since  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w})$  is concave (Lemma 1) and locally Lipschitz on  $\Delta_K$  (Lemma 7), the chain rule for concave function along absolutely continuous curves (see, e.g., Brezis (1973)) yields the claim.  $\square$

**Remark 1** (More general chain-rule conditions). *Lemma 8 is a special case of general chain rules for subdifferentially regular functions, Davis et al. (2020, Lemma 5.2) and Bolte and Pauwels (2021, Section 3.2) gives a broader characterization, known as path differentiable functions.*

The main result for this section is the convergence of the continuous-time dynamics to stationary points via Lyapunov analysis. Define the information value along the DI solution  $\mathbf{w}(\cdot)$  as

$$V(t) = \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t)).$$

We show that such a information value *does not decrease*, and serve as a (weak) Lyapunov function.

**Theorem 4** (Global Lyapunov function). *Let  $\mathbf{w}(\cdot)$  be any absolutely continuous solution of (18), then  $V(\cdot)$  is absolutely continuous and*

$$\frac{d}{dt} V(t) \geq 0 \quad \text{for a.e. } t \geq 0.$$

*Proof.* By the definition of the selection correspondence and Carathéodory's theorem, there exist sequences

$$(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j})_{n \rightarrow \infty} \subseteq \mathcal{R} \times \Delta_{|\mathcal{X}|}, \quad j \in [K],$$

and weights  $(\lambda^j)_{j \in [K]} \in \Delta_K$  such that  $\mathbf{w}^{n,j} \xrightarrow{n \rightarrow \infty} \mathbf{w}(t)$  for each  $j$  and

$$\sum_{j \in [K]} \lambda^j \mathbf{h}(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) - \mathbf{w}(t) \xrightarrow{n \rightarrow \infty} \frac{d}{dt} \mathbf{w}(t). \quad (35)$$

The limit (upon possibly passing to a subsequence)

$$\mathbf{f}^*(\mathbf{w}(t)) \triangleq \lim_{n \rightarrow \infty} \sum_{j \in [K]} \lambda^j \nabla_{\mathbf{w}} F(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) \in \partial^{\circ} \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t)) \quad (36)$$

since the Clarke subdifferential is closed under such limits. Since  $\mathbf{w}(\cdot)$  is absolutely continuous and  $\mathbf{w} \mapsto \min_{x \in \mathcal{X}} D_x^w(\mathbf{w})$  is Lipschitz on  $\Delta_K$  (Lemma 7), the map  $V(t) = \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t))$  is absolutely continuous and hence differentiable for a.e.  $t \geq 0$  and lemma 8 gives

$$\frac{d}{dt} V(t) = \mathbf{f}^*(\mathbf{w}(t))^{\top} \frac{d}{dt} \mathbf{w}(t), \quad \text{for a.e. } t.$$

We claim that this time derivative is nonnegative, which by (35) and (36) is sufficient to show

$$\lim_{n \rightarrow \infty} \left( \sum_{j \in [K]} \lambda^j \nabla_{\mathbf{w}} F(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) \right)^{\top} \left( \sum_{j \in [K]} \lambda^j \mathbf{h}(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) \right) \geq \mathbf{f}^*(\mathbf{w}(t)) \circ \mathbf{w}(t).$$

Indeed, for any fixed  $n$ , introduce the shorthand

$$w_i^j \triangleq w_i^{n,j}, \quad f_i^j \triangleq \lambda^j [\nabla_{\mathbf{w}} F(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j})]_i, \quad F^j \triangleq \sum_{i \in [K]} w_i^j f_i^j = \lambda^j F(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}).$$

Then the weighted IDS probability satisfies  $\lambda^j h_i(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) = \frac{w_i^j f_i^j}{F^j}$ , and we can compute

$$\begin{aligned}
& \left( \sum_{j \in [K]} \lambda^j \nabla_{\mathbf{w}} F(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) \right)^\top \left( \sum_{j \in [K]} \lambda^j \mathbf{h}(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) \right) \\
&= \sum_{i \in [K]} \left( \sum_{j \in [K]} f_i^j \right) \left( \sum_{j \in [K]} \frac{w_i^j f_i^j}{F^j} \right) \\
&= \sum_{i \in [K]} \sum_{j \in [K]} \frac{w_i^j f_i^j}{F^j} \left( \sum_{k \in [K]} f_i^k \right) \\
&= \sum_{j \in [K]} \frac{1}{F^j} \sum_{i \in [K]} w_i^j f_i^j \left( \sum_{k \in [K]} f_i^k \right), \tag{37}
\end{aligned}$$

For each fixed  $j \in [K]$ , using  $\sum_{k \in [K]} f_i^k \geq f_i^j$  and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\sum_{i \in [K]} w_i^j f_i^j \left( \sum_{k \in [K]} f_i^k \right) &\geq \sum_{i \in [K]} w_i^j (f_i^j)^2 = \left( \sum_{i \in [K]} w_i^j (f_i^j)^2 \right) \left( \sum_{i \in [K]} w_i^j \right) \\
&\geq \left( \sum_{i \in [K]} w_i^j f_i^j \right)^2 = (F^j)^2. \tag{38}
\end{aligned}$$

Summing over  $j$  and using (37), we obtain

$$(37) \geq \sum_{j \in [K]} \frac{1}{F^j} (F^j)^2 = \sum_{j \in [K]} F^j = \sum_{j \in [K]} \lambda^j F(\mathbf{w}^{n,j}, \boldsymbol{\nu}^{n,j}) \xrightarrow{n \rightarrow \infty} \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t)).$$

An application of the Euler identity for the Clarke gradient gives

$$\mathbf{f}^*(\mathbf{w}(t))^\top \mathbf{w}(t) = \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t)),$$

which establishes the desired claim and

$$\frac{d}{dt} V(t) = \mathbf{f}^*(\mathbf{w}(t))^\top \frac{d}{dt} \mathbf{w}(t) \geq 0 \quad \text{for a.e. } t.$$

□

**Corollary 1** (LaSalle limit points and stationary values). *Let  $\mathbf{w}^* \in \Delta_K$  be any  $\omega$ -limit point of the DI (18). Then there exists  $\boldsymbol{\nu}^* \in \det(\mathbf{w}^*)$  and  $\mathbf{f}(\mathbf{w}^*, \boldsymbol{\nu}^*) \in \partial^* \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}(t))$  such that*

$$[\mathbf{f}(\mathbf{w}^*, \boldsymbol{\nu}^*)]_i = F(\mathbf{w}^*, \boldsymbol{\nu}^*) \quad \text{for all } i \in \text{Supp}(\mathbf{w}^*). \tag{39}$$

Writing  $J \triangleq \text{Supp}(\mathbf{w}^*)$  and the face  $\Delta_J \triangleq \{\mathbf{w} \in \Delta_K : p_i = 0 \ (i \notin J)\}$ , we have the facewise maximality

$$\min_{x \in \mathcal{X}} D_x^w(\mathbf{w}^*) = \max_{\mathbf{w} \in \Delta_J} \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}).$$

Consequently, the set of all stationary values is

$$\left\{ \max_{\mathbf{w} \in \Delta_J} \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}) : \emptyset \neq J \subseteq [K] \right\},$$

which is finite, hence has empty interior.

*Proof.* By Theorem 4,  $t \mapsto \min_x D_x^w(\mathbf{w}(t))$  is nondecreasing. LaSalle's invariance principle for DIs implies that every  $\omega$ -limit point lies in the largest weakly invariant subset of the set where the time derivative of  $t \mapsto \min_x D_x^w(\mathbf{w}(t))$  vanishes.

Take limit to the derivative identity (38) where the Cauchy–Schwarz inequality is used,

$$\sum_{i \in [K]} w_i^* [\mathbf{f}(\mathbf{w}^*, \boldsymbol{\nu}^*)]_i^2 - \left( \sum_{i \in [K]} w_i^* [\mathbf{f}(\mathbf{w}^*, \boldsymbol{\nu}^*)]_i \right)^2 = \sum_{i,j} w_i^* w_j^* \left( [\mathbf{f}(\mathbf{w}^*, \boldsymbol{\nu}^*)]_i - [\mathbf{f}(\mathbf{w}^*, \boldsymbol{\nu}^*)]_j \right)^2.$$

Vanishing of this quantity forces the bracketed components to be equal on  $J = \text{Supp}(\mathbf{w}^*)$ . By Euler's identity  $\sum_i w_i^* [\mathbf{f}(\mathbf{w}^*, \boldsymbol{\nu}^*)]_i = F(\mathbf{w}^*, \boldsymbol{\nu}^*)$ , that common value is  $F(\mathbf{w}^*, \boldsymbol{\nu}^*)$ , yielding the equality condition (39).

For any  $\mathbf{w} \in \Delta_J$ , concavity of  $F(\cdot, \boldsymbol{\nu}^*)$  gives

$$F(\mathbf{w}, \boldsymbol{\nu}^*) \leq F(\mathbf{w}^*, \boldsymbol{\nu}^*) + \mathbf{f}(\mathbf{w}^*, \boldsymbol{\mu}^*)^\top (\mathbf{w} - \mathbf{w}^*) = F(\mathbf{w}^*, \boldsymbol{\nu}^*),$$

since the gradient components are equal on  $J$  and  $\sum_{i \in J} (w_i - w_i^*) = 0$ . Because  $\min_x D_x^w(\mathbf{w}) \leq F(\mathbf{w}, \boldsymbol{\nu}^*)$  for all  $\mathbf{w}$ , with equality at  $\mathbf{w}^*$  (as  $\boldsymbol{\nu}^* \in \det(\mathbf{w}^*)$ ), we obtain

$$\min_x D_x^w(\mathbf{w}) \leq \min_x D_x^w(\mathbf{w}^*), \quad \forall \mathbf{w} \in \Delta_J,$$

proving the facewise maximization in (ii).

The description and finiteness of stationary values then follow by enumerating nonempty faces  $J \subseteq [K]$ .  $\square$

**Remark 2** (Weak Sard property). *Let  $g(\mathbf{w}) := \min_{x \in \mathcal{X}} D_x(\mathbf{w})$ . A standard first-order (Clarke) stationarity condition for the constrained maximization  $\max_{\mathbf{w} \in \Delta_K} g(\mathbf{w})$  is the normal-cone inclusion*

$$\mathbf{0} \in -\partial^\circ g(\mathbf{w}^*) + N_{\Delta_K}(\mathbf{w}^*), \quad (40)$$

where for any closed convex  $C \subseteq \mathbb{R}^K$  and  $\mathbf{w} \in C$  the normal cone is

$$N_C(\mathbf{w}) := \{\mathbf{u} \in \mathbb{R}^K : \langle \mathbf{u}, \mathbf{q} - \mathbf{w} \rangle \leq 0 \quad \forall \mathbf{q} \in C\}.$$

For  $C = \Delta_K$  and  $J = \text{Supp}(\mathbf{w}^*)$ , one has

$$N_{\Delta_K}(\mathbf{w}^*) = \left\{ \mathbf{u} : \exists \lambda \in \mathbb{R} \text{ s.t. } u_i = \lambda \ (i \in J), \ u_i \leq \lambda \ (i \notin J) \right\}.$$

Thus (40) requires a Clarke subgradient to be constant on the support and no larger off the support; (39) in Corollary 1 recovers the equality part.

The corresponding critical values are the objective values attained at such points. Corollary 1 shows that this set is finite (hence has empty interior), which is exactly the “weak Sard property” postulated in Davis et al. (2020, Assumption B); see also Ioffe (2017, Section 8.4) and Bolte and Pauwels (2021, Theorem 5).

## E.2 Stochastic Approximation

We relate the discrete updates (11) to a compact, convex, upper hemicontinuous (u.h.c.) differential inclusion (DI) via a standard interpolation and an  $\varepsilon$ -perturbation scheme, as in Benaïm et al. (2005).

**Definition 2** (Continuous-time interpolation). *Set  $t_0 = 0$  and  $t_n = \sum_{m=0}^{n-1} \alpha_m$ . For  $n \in \mathbb{N}$ , we define the linear interpolation of  $\mathbf{w}_n$  by*

$$\hat{\mathbf{w}}(t_n + s) \triangleq \mathbf{w}_n + s \frac{\mathbf{w}_{n+1} - \mathbf{w}_n}{t_{n+1} - t_n}. \quad (41)$$

Then  $\hat{\mathbf{w}}(t_n) = \mathbf{w}_n$  for all  $n$ .

Recall  $G : \Delta_K \times \mathbb{R}^d \rightrightarrows T_{\Delta_K}$  denote the right-hand side correspondence of the continuous dynamics with the state  $\boldsymbol{\theta}$ :

$$G(\mathbf{w}; \boldsymbol{\theta}) \triangleq \mathbf{sel}(\mathbf{w}; \boldsymbol{\theta}) - \mathbf{w},$$

where  $\mathbf{sel}(\mathbf{w}; \boldsymbol{\theta})$  is defined by replacing  $D_x^w(\mathbf{w})$  with  $D_x^w(\mathbf{w}; \boldsymbol{\theta})$ .

**Definition 3** ( $\varepsilon$ -Perturbed DI). *For  $\varepsilon \geq 0$  define*

$$G^\varepsilon(\mathbf{w}; \boldsymbol{\theta}) \triangleq \left\{ \mathbf{y} \in \mathbb{R}^K : \exists \mathbf{w}' \in \Delta_K \text{ s.t. } \|\mathbf{w} - \mathbf{w}'\|_1 + d(\mathbf{y}, G(\mathbf{w}'; \boldsymbol{\theta})) \leq \varepsilon \right\}.$$

Define the solution set  $\mathcal{T}_{\boldsymbol{\theta}}^{\varepsilon, T} : \Delta_K \rightrightarrows \mathcal{AC}([0, T]; \mathbb{R}^K)$  that maps an initial condition in  $\Delta_K$  to the (nonempty) set of absolutely continuous solutions of the  $\varepsilon$ -perturbed DI on  $[0, T]$ :

$$\frac{d}{dt} \mathbf{w}(t) \in G^\varepsilon(\mathbf{w}(t); \boldsymbol{\theta}) + U(t), \quad \mathbf{w}(0) = \mathbf{w}_0,$$

where locally integrable (stochastic) process  $U(t)$  satisfying

$$\sup_{0 \leq t \leq T} \left\| \int_0^t U(t) dt \right\|_1 \leq \epsilon, \text{ with probability one.}$$

To show that our allocation update indeed realizes a perturbed differential inclusion in the sense of Definition 3, we first use the following consistency property of the instance estimator.

**Lemma 9** (Sufficient exploration ensures consistency of  $\boldsymbol{\theta}_n$ ). *Under the forced-exploration schedule in (7), the instance estimate satisfies  $\boldsymbol{\theta}_n \rightarrow \boldsymbol{\theta}$  almost surely.*

Assumption 3 implies that there exists a neighborhood  $U$  of the true  $\boldsymbol{\theta}$  such that for all  $\boldsymbol{\vartheta} \in U$ , we have  $\mathcal{X}(\boldsymbol{\vartheta}) = \mathcal{X}(\boldsymbol{\theta})$  and  $\text{Alt}_x(\boldsymbol{\vartheta}) = \text{Alt}_x(\boldsymbol{\theta})$  for all  $x$  in  $\mathcal{X}(\boldsymbol{\theta})$ . In view of Lemma 1, the minimizer is unique for  $\mathbf{w} \in \mathcal{R}$ . Then, Degenne and Koolen (2019, Theorem 4) implies that, for a fixed true parameter  $\boldsymbol{\theta} \in \mathbb{R}^d$  and for every  $\mathbf{w} \in \mathcal{R}$  and every  $x \in \mathcal{X}$ , the maps

$$\boldsymbol{\vartheta} \mapsto D_x^w(\mathbf{w}, \boldsymbol{\vartheta}) \quad \text{and} \quad \boldsymbol{\vartheta} \mapsto \nabla_{\mathbf{w}} D_x^w(\mathbf{w}, \boldsymbol{\vartheta})$$

are continuous at  $\boldsymbol{\vartheta} = \boldsymbol{\theta}$ .

**Lemma 10** (U.h.c. at the true state parameter). *Fix  $\mathbf{w} \in \mathcal{R}$  (so  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w}; \boldsymbol{\theta}) > 0$ ). Then  $\boldsymbol{\theta}' \mapsto G(\mathbf{w}, \boldsymbol{\theta}')$  is upper hemicontinuous at  $\boldsymbol{\theta}$ : for every  $\varepsilon > 0$  there exists  $\delta > 0$  such that*

$$\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_\infty \leq \delta \implies G(\mathbf{w}, \boldsymbol{\theta}') \subseteq G(\mathbf{w}; \boldsymbol{\theta}) + \varepsilon B,$$

where  $B$  is the unit ball of  $\ell_1(\mathbb{R}^K)$ .

*Proof.* Recall that  $F(\mathbf{w}, \boldsymbol{\nu}; \boldsymbol{\vartheta}) = \sum_x \nu_x D_x^w(\mathbf{w}; \boldsymbol{\vartheta})$  is continuous in  $(\boldsymbol{\nu}; \boldsymbol{\vartheta})$  and linear in  $\boldsymbol{\nu}$  on the compact set  $\Delta_{|\mathcal{X}|}$ ; hence by Berge's maximum theorem

$$\det(\mathbf{w}; \boldsymbol{\vartheta}) \triangleq \arg \min_{\boldsymbol{\nu} \in \Delta_{|\mathcal{X}|}} F(\mathbf{w}, \boldsymbol{\nu}; \boldsymbol{\vartheta})$$

is nonempty, compact-valued, and u.h.c. at  $\boldsymbol{\theta}$ . Since  $\min_x D_x^w(\mathbf{w}; \boldsymbol{\theta}) > 0$  and  $D_x^w(\mathbf{w}, \cdot)$  is continuous at  $\boldsymbol{\theta}$ , there exists a neighborhood  $U$  of  $\boldsymbol{\theta}$  with  $\sum_x \nu_x D_x^w(\mathbf{w}; \boldsymbol{\vartheta}) \geq \min_{x \in \mathcal{X}} D_x^w(\mathbf{w}; \boldsymbol{\theta})/2 > 0$  for all  $\boldsymbol{\vartheta} \in U$  and all  $\boldsymbol{\nu} \in \det(\mathbf{w}; \boldsymbol{\vartheta})$ . Thus

$$(\boldsymbol{\nu}; \boldsymbol{\vartheta}) \mapsto \mathbf{h}(\mathbf{w}, \boldsymbol{\nu}; \boldsymbol{\vartheta}) = \left( \frac{p_i \sum_x \nu_x [\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\vartheta})]_i}{\sum_x \nu_x D_x^w(\mathbf{w}; \boldsymbol{\vartheta})} \right)_{i \in [K]}$$

is continuous on  $\Delta_{|\mathcal{X}|} \times U$ . The image correspondence

$$\boldsymbol{\vartheta} \mapsto \{\mathbf{h}(\mathbf{w}, \boldsymbol{\nu}; \boldsymbol{\vartheta}) - \mathbf{w} : \boldsymbol{\nu} \in \det(\mathbf{w}; \boldsymbol{\vartheta})\} = G(\mathbf{w}; \boldsymbol{\vartheta})$$

is therefore the continuous image of a compact-valued u.h.c. correspondence, hence u.h.c. at  $\boldsymbol{\theta}$ . The stated  $\varepsilon$ -inclusion follows.  $\square$

**Lemma 11** (Martingale on non-forced rounds). *Define*

$$\mathsf{NF}_n \triangleq \mathbb{1} \left\{ \sqrt{\lfloor n/K \rfloor} \notin \mathbb{Z} \right\} \quad \text{and} \quad \mathsf{F}_n \triangleq \mathbb{1} \left\{ \sqrt{\lfloor n/K \rfloor} \in \mathbb{Z} \right\},$$

i.e.,  $\mathsf{NF}_n = 1$  on non-forced rounds and  $\mathsf{NF}_n = 0$  on forced rounds. Define

$$\boldsymbol{\xi}_m \triangleq \sum_{n=0}^{m-1} \mathsf{NF}_n \alpha_{n+1} (\mathbf{e}_{I_{n+1}} - \mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)), \quad m \geq 0.$$

Then  $(\boldsymbol{\xi}_m)_{m \geq 0}$  is an  $\mathbb{R}^K$ -valued martingale with respect to  $(\mathcal{H}_m)_{m \geq 0}$ , and  $\boldsymbol{\xi}_m$  converges almost surely to a finite random limit  $\boldsymbol{\xi}_\infty$ .

*Proof.* On non-forced rounds  $\mathbb{E}[\alpha_{n+1} \mathbf{e}_{I_{n+1}} \mid \mathcal{H}_n] = \tilde{\alpha}_{n+1} \tilde{\mathbf{h}}_n$ , while  $\mathsf{NF}_n = 0$  on forced rounds. Hence

$$\mathbb{E} \left[ \mathsf{NF}_n \alpha_{n+1} (\mathbf{e}_{I_{n+1}} - \tilde{\mathbf{h}}_n) \mid \mathcal{H}_n \right] = \mathbf{0},$$

so  $(\boldsymbol{\xi}_m)$  is a martingale. Since  $\|\mathbf{e}_{I_n} - \tilde{\mathbf{h}}_n\|_1 \leq 2$ , we have

$$\sum_{n=0}^{\infty} \mathbb{E} \left[ \left\| \mathsf{NF}_n \alpha_{n+1} (\mathbf{e}_{I_n} - \tilde{\mathbf{h}}_n) \right\|_1^2 \mid \mathcal{H}_n \right] \leq 4 \sum_{n=0}^{\infty} \tilde{\alpha}_{n+1}^2 < \infty.$$

By the martingale convergence theorem for square-summable increments,  $\boldsymbol{\xi}_m \rightarrow \boldsymbol{\xi}_\infty$  a.s.  $\square$

**Proposition 4** (Interpolation process to perturbed solution). *For any  $\epsilon > 0$  and  $T \geq 0$ , there exists almost surely finite  $S(\epsilon, T)$  such that the translated interpolation*

$$\hat{\mathbf{w}}(s+t)_{t \in [0, T]} \in \mathcal{T}_{\boldsymbol{\theta}}^{\epsilon, T}(\hat{\mathbf{w}}(s)), \quad \forall s \geq S(\epsilon, T)$$

with probability one.

*Proof.* Fix  $t \in [s, s+T]$  satisfying  $t \in (t_n, t_{n+1})$ . Define

$$U(t) = \mathbf{e}_{I_{n+1}} - \mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n),$$

By (41) and  $t_{n+1} - t_n = \alpha_{n+1}$ , we have:

$$\frac{d}{dt} \hat{\mathbf{w}}(t) = \frac{\mathbf{w}_{n+1} - \mathbf{w}_n}{\alpha_{n+1}} = \mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) - \mathbf{w}_n + U(t).$$

For  $t = t_n + r$  with  $r \in [0, \alpha_{n+1}]$ ,

$$\|\hat{\mathbf{w}}(t_n + r) - \mathbf{w}_n\|_1 = \frac{r}{\alpha_{n+1}} \|\mathbf{w}_{n+1} - \mathbf{w}_n\|_1 \leq \|\mathbf{w}_{n+1} - \mathbf{w}_n\|_1 = \alpha_{n+1} \|\mathbf{e}_{I_{n+1}} - \mathbf{w}_n\|_1 \leq 2\alpha_{n+1}.$$

Note that  $\mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) - \mathbf{w}_n \in G(\mathbf{w}_n; \boldsymbol{\theta}_n)$ , thus by Lemma 9 and Lemma 10 (where we use  $\mathbf{w}_n \in \text{int}(\Delta_K) \subseteq \mathcal{R}$  for finite  $n$ ), there exists  $N(\epsilon)$  such that when  $n \geq N(\epsilon)$ , we have the definition of  $G^\epsilon$  is satisfied with  $\mathbf{w}' = \mathbf{w}_n$ ,

$$\|\hat{\mathbf{w}}(t) - \mathbf{w}'\|_1 + d(\mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) - \mathbf{w}_n, G(\mathbf{w}_n; \boldsymbol{\theta})) \leq 2\alpha_{n+1} + d(G(\mathbf{w}_n; \boldsymbol{\theta}_n), G(\mathbf{w}_n; \boldsymbol{\theta})) \leq \epsilon.$$

For time interval  $[s, s+T]$ , we partition it into two disjoint sets  $A_1(s), A_2(s)$ , for any  $t \in A_1(s)$  the algorithm take IDS, for any  $t \in A_2(s)$  the algorithm take forced exploration. Since  $\sum_{m=1}^{\infty} \sum_{k=0}^{K-1} \alpha_{Km^2+k} < \infty$ , we have  $\lim_{s \rightarrow \infty} |A_1(s)| \rightarrow 0$ . Thus there exists  $S_0(\epsilon, T) < \infty$  such that when  $s \geq S_0(\epsilon, T)$

$$\left\| \int_{A_1(s)} U(t) dt \right\|_1 \leq 2|A_1(s)| \leq \epsilon/2.$$

To transform interpolation time  $t$  to round  $n$ , we use

$$\frac{c_{\min}}{c_{\max}} \log(n+1) = \frac{c_{\min}}{c_{\max}} \int_1^{n+1} \left( \frac{1}{t} dt \right) \leq t_n \leq \frac{c_{\max}}{c_{\min}} \left( 1 + \int_1^n \frac{1}{t} dt \right) = \frac{c_{\max}}{c_{\min}} (\log(n) + 1),$$

then  $t_{n(s)} \leq s \leq s+T \leq t_{n(s+T)}$  where  $n(s) = \lfloor \exp(s-1)c_{\min}/c_{\max} \rfloor$  and  $n(s+T) = \lfloor \exp(s+T)c_{\max}/c_{\min} \rfloor$ . For non-forced rounds, by (16)  $\|\tilde{\mathbf{h}}_n - \mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)\|_1 \rightarrow 0$  as  $n \rightarrow \infty$ . By Lemma 11 that  $\xi_m = \sum_{n=0}^{m-1} \mathbf{NF}_n \alpha_{n+1} (\mathbf{e}_{I_{n+1}} - \tilde{\mathbf{h}}_n)$  converges almost surely, there exists  $S_1(\epsilon, T) < \infty$  such that for all  $s \geq S_1(\epsilon, T)$ , with probability one,

$$\begin{aligned} \left\| \int_{A_2(s)} U(t) dt \right\|_1 &\leq \|\xi_{n(s+T)} - \xi_{n(s)}\|_1 + \left\| \sum_{n=n(s)}^{n(s+T)} \mathbf{NF}_n \alpha_{n+1} (\tilde{\mathbf{h}}_n - \mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)) \right\|_1 \\ &\leq \epsilon/2, \end{aligned}$$

where the first inequality follows from the triangle inequality and the definition of  $\xi_n$ , while the second is due to the martingale convergence and the limit of  $\|\tilde{\mathbf{h}}_n - \mathbf{h}^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)\|_1$ .

Therefore there exists  $T(\epsilon) \triangleq \max\{T_0(\epsilon), T_1(\epsilon)\} < \infty$  such that when  $s \geq T(\epsilon)$

$$\sup_{0 \leq t \leq T} \left\| \int_s^{s+t} U(t) dt \right\|_1 \leq \left\| \int_s^{s+T} U(t) dt \right\|_1 \leq \left\| \int_{A_1(s)} U(t) dt \right\|_1 + \left\| \int_{A_2(s)} U(t) dt \right\|_1 \leq \epsilon.$$

□

**Definition 4** (Asymptotic pseudo-trajectory (APT) for u.h.c DI). *Consider a u.h.c. differential inclusion and let  $S_{x(0)}$  be the set of solutions with initial value  $x(0)$ ; write  $S \triangleq \bigcup_{x(0)} S_{x(0)}$  for the set of all solutions. A bounded continuous curve  $z : [0, \infty) \rightarrow \mathbb{R}^m$  is an asymptotic pseudo-trajectory (APT) of the DI if for every  $T > 0$ ,*

$$\lim_{t \rightarrow \infty} \inf_{\sigma \in S} \sup_{s \in [0, T]} \|z(t + s) - \sigma(s)\| = 0.$$

*Equivalently: for all  $\varepsilon, T > 0$  there exists  $t_0 < \infty$  such that for all  $t \geq t_0$  there is a DI solution segment  $\sigma : [0, T] \rightarrow \mathbb{R}^m$  with  $\sup_{s \in [0, T]} \|z(t + s) - \sigma(s)\| \leq \varepsilon$ .*

Note that in this definition the comparison solution  $\sigma$  need not start exactly from  $z(t)$ ; we only require that its initial condition be within  $\varepsilon$  of  $z(t)$ . This is precisely the definition adopted in Esponda et al. (2022), and see also Appendix C of Bianchi et al. (2024).

**Definition 5**  $((\varepsilon, T)$ -chains and internally chain transitive (ICT) sets). *Let  $(X, d)$  be a compact metric space and  $\{\Phi_t\}_{t \geq 0}$  a continuous semiflow on  $X$ . For  $\varepsilon > 0$  and  $T > 0$ , an  $(\varepsilon, T)$ -chain from  $x$  to  $y$  is a finite sequence*

$$x = z_0, z_1, \dots, z_m = y, \quad t_1, \dots, t_m \geq T,$$

*such that  $d(\Phi_{t_i}(z_{i-1}), z_i) < \varepsilon$  for all  $i$ . A nonempty compact set  $L \subset X$  is ICT if: (i)  $L$  is invariant ( $\Phi_t(L) = L$  for all  $t \geq 0$ ); and (ii) for all  $x, y \in L$  and all  $\varepsilon, T > 0$ , there exists an  $(\varepsilon, T)$ -chain entirely contained in  $L$  from  $x$  to  $y$ .*

**Proposition 5** (Interpolation  $\Rightarrow$  APT  $\Rightarrow$  ICT  $\omega$ -limit set). *The interpolation process  $\hat{\mathbf{w}}$  is an a.s. asymptotic pseudo-trajectory of the differential inclusion (18); and its  $\omega$ -limit set  $\Omega \triangleq \omega(\hat{\mathbf{w}}) \subset \Delta_K$  is internally chain transitive.*

*Proof.* The first part follows from our Porposition 4 and Benaïm et al. (2005, Theorem 4.2). The second conclusion follows from Benaïm et al. (2005, Theorem 4.3). □

**Proposition 6** (Constancy of  $\min_x D_x$  on ICT sets). *The map  $\mathbf{w} \mapsto \min_{x \in \mathcal{X}} D_x^w(\mathbf{w})$  is constant on  $\Omega$ . Consequently (along the PAN iterates),  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}_n)$  converges almost surely.*

*Proof.* By Theorem 4,  $t \mapsto \min_x D_x^w(\mathbf{w}(t))$  is nondecreasing along any solution and strictly increasing off the stationary set of Corollary 1; the set of stationary values has empty interior. Hence Proposition 3.27 of Benaïm et al. (2005) applies: on the ICT set  $\Omega = \omega(\hat{\mathbf{w}})$  the function  $\min_x D_x$  is constant. □

**Remark 3.** The framework of Benaïm et al. (2005) using the internally chain transitive set is essentially the same as the “non-escape argument” in Davis et al. (2020, Section 3.3). In Benaïm’s framework, an ICT set is a compact invariant set with no proper attracting subset, see Benaïm et al. (2005, Proposition 3.20). If a Lyapunov function is nondecreasing along the DI and its stationary values form a thin set (weak sard), then every small sublevel that is forward invariant would have to attract the whole ICT set; this forces the Lyapunov to be constant on that set. The proof in Davis et al. (2020, Section 3.3) does the similar thing: they show the APT cannot keep re-entering them—hence “non-escape”—so the Lyapunov must converge and be constant on the limit set.

### E.3 Algorithm 1 Attains Positive Information

Recall from Lemma 1 that for each  $x \in \mathcal{X}$  there exists  $d_x(\boldsymbol{\theta}) > 0$  such that  $\sum_{i=1}^K [\nabla_{\mathbf{w}} D_x^w(\mathbf{w}; \boldsymbol{\theta})]_i \geq d_x(\boldsymbol{\theta})$  on  $\text{int}(\Delta_K)$ . We set

$$d_* \triangleq \min_{x \in \mathcal{X}} d_x(\boldsymbol{\theta}) > 0. \quad (42)$$

**Proposition 7** (Interior start rules out zero information). *There exists path independent constant  $D_*$  such that*

$$\lim_{n \rightarrow \infty} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n) \geq D_* \triangleq \frac{c_{\min}^2 d_*}{8c_{\max}^2(K+2)} \quad \text{with probability one}$$

where the limit exists by Proposition 6 and Lemma 2.

*Proof.* Assume for contradiction that on a set  $A$  with  $\mathbb{P}(A) > 0$ ,  $\lim_{n \rightarrow \infty} \min_x D_x^w(\mathbf{w}_n; \boldsymbol{\theta}_n) < D_*$ .

**Non-forced rounds give a uniform multiplicative lift** Conditioning on  $\mathcal{H}_n$  and for non-forced rounds  $n$  large enough such that  $\tilde{h}_{n,i} \geq h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)/2 = w_{n,i}[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i/2D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)$ ,

$$\begin{aligned} \mathbb{E} \left[ \sum_{i=1}^K \frac{w_{n+1,i}}{w_{n,i}} \mid \mathcal{H}_n \right] &= K + \tilde{\alpha}_{n+1} \left( \sum_{i=1}^K \frac{\tilde{h}_{n,i}}{w_{n,i}} - K \right) \\ &\geq K + \tilde{\alpha}_{n+1} \left( \frac{\sum_i [\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{2D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} - K \right). \end{aligned}$$

By (42), continuity in  $\boldsymbol{\theta}$ , and  $\boldsymbol{\theta}_n \rightarrow \boldsymbol{\theta}$  a.s., there exists  $n_0$  such that, on  $A$  and for all non-forced  $n \geq n_0$ ,

$$\sum_i [\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i \geq d_*/2, \quad D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n) \leq 2D_* = \frac{c_{\min}^2 d_*}{4c_{\max}^2(K+2)}.$$

Hence, for non-forced  $n \geq n_0$ ,

$$\mathbb{E} \left[ \sum_{i=1}^K \frac{w_{n+1,i}}{w_{n,i}} \mid \mathcal{H}_n \right] \geq K + \frac{2c_{\max}^2}{c_{\min}^2} \tilde{\alpha}_{n+1} \geq K + 2 \frac{c_{\max}}{c_{\min}(n+1)}. \quad (43)$$

Since all the ratios satisfy  $w_{n+1,i}/w_{n,i} \geq 1 - \frac{c_{\max}}{c_{\min}(n+1)}$ , by the Bauer maximum principle (concavity of  $\sum_i \log(\cdot)$ ) with  $G_n \triangleq \prod_{i=1}^K w_{n,i}$  we have

$$\frac{G_{n+1}}{G_n} = \prod_{i=1}^K \frac{w_{n+1,i}}{w_{n,i}} \geq \left( 1 - \frac{c_{\max}}{c_{\min}(n+1)} \right)^{K-1} \left( \sum_{i=1}^K \frac{w_{n+1,i}}{w_{n,i}} - (K-1) \left( 1 - \frac{c_{\max}}{c_{\min}(n+1)} \right) \right).$$

Since this inequality holds for every realization conditional on  $\mathcal{H}_n$ , we may take conditional expectations on both sides and use (43) for the sum to obtain, for all large non-forced  $n$ ,

$$\mathbb{E} \left[ \frac{G_{n+1}}{G_n} \mid \mathcal{H}_n \right] \geq \left( 1 - \frac{c_{\max}}{c_{\min}(n+1)} \right)^{K-1} \left( 1 + \frac{(K+1)c_{\max}}{(n+1)c_{\min}} \right) \geq 1 + \frac{c_{\max}}{c_{\min}(n+1)}, \quad (44)$$

where in the last inequality we used  $(1-x)^{K-1} \geq 1 - (K-1)x$  and  $(1-(K-1)x)(1+(K+1)x) = 1+2x-(K^2-1)x^2 \geq 1+x$  for  $x = \frac{c_{\max}}{c_{\min}(n+1)}$  and  $n$  large enough.

**Forced rounds only decrease mildly** At a forced round, since  $w_{n+1,i} \geq (1-c_{\max}/(n+1)c_{\min})w_{n,i}$ , we have the deterministic bound

$$\mathbb{E} \left[ \frac{G_{n+1}}{G_n} \mid \mathcal{H}_n \right] = \mathbb{E} \left[ \prod_{i=1}^K \frac{w_{n+1,i}}{w_{n,i}} \mid \mathcal{H}_n \right] \geq \left( 1 - \frac{c_{\max}}{c_{\min}(n+1)} \right)^K \geq 1 - \frac{c_{\max}K}{c_{\min}(n+1)}.$$

We obtain that on event  $A$  and for all  $n \geq N$ ,

$$\mathbb{E}[G_{n+1} \mid \mathcal{H}_n] = G_n \mathbb{E} \left[ \frac{G_{n+1}}{G_n} \mid \mathcal{H}_n \right] \geq a_{n+1} G_n, \quad (45)$$

where

$$a_{n+1} := \begin{cases} 1 + \frac{c_{\max}}{c_{\min}(n+1)}, & \sqrt{\left\lfloor \frac{n}{K} \right\rfloor} \notin \mathbb{Z}, \\ 1 - \frac{c_{\max}K}{c_{\min}(n+1)}, & \sqrt{\left\lfloor \frac{n}{K} \right\rfloor} \in \mathbb{Z}. \end{cases}$$

Iterating (45) from  $N$  to  $N'$  using the tower property gives, on  $A$ ,

$$\mathbb{E}[G'_N \mid \mathcal{H}_N] \geq G_N \prod_{n=N}^{N'-1} a_{n+1} \geq G_N \exp \left( \sum_{n=N}^{N'-1} \frac{c_{\max}}{2c_{\min}(n+1)} \mathsf{NF}_n - \frac{2c_{\max}K}{c_{\min}(n+1)} \mathsf{F}_n \right)$$

where we use for all sufficiently large  $n$ ,  $\log(1+x) \geq x/2$  with  $x = \frac{c_{\max}}{c_{\min}(n+1)}$  and  $\log(1-y) \geq -2y$  with  $y = \frac{c_{\max}K}{c_{\min}(n+1)}$ . Therefore on  $A$ ,  $\mathbb{E}[G'_N \mid \mathcal{H}_N] \rightarrow +\infty$  as  $N' \rightarrow \infty$ . This contradicts with  $G_{N'} = \prod_{i=1}^K w_{N',i} \leq K^{-K}$ .  $\square$

#### E.4 Completing the Proof: Algorithm 1 Attains Optimal Value

A pair  $(\mathbf{w}^*, \boldsymbol{\mu}^*)$  is a *saddle point* of  $F$  if

$$F(\mathbf{w}, \boldsymbol{\mu}^*) \leq F(\mathbf{w}^*, \boldsymbol{\mu}^*) \leq F(\mathbf{w}^*, \boldsymbol{\mu}) \quad \forall \mathbf{w} \in \Delta_K, \boldsymbol{\mu} \in \Delta_{|\mathcal{X}|}.$$

The common value  $F^* \triangleq F(\mathbf{w}^*, \boldsymbol{\mu}^*)$  is the minimax value. Let

$$\mathcal{E} \triangleq \{(\mathbf{w}, \boldsymbol{\mu}) \in \Delta_K \times \Delta_{|\mathcal{X}|} : (\mathbf{w}, \boldsymbol{\mu}) \text{ is a saddle point of } F\}.$$

For each  $i \in [K]$ , define the attainable  $i$ -th coordinate set

$$\mathcal{E}_{w_i} \triangleq \{w_i \in [0, 1] : \exists (\mathbf{w}, \boldsymbol{\mu}) \in \mathcal{E} \text{ with } (\mathbf{w})_i = w_i\},$$

and let  $\mathcal{K}_0 \triangleq \{i : \mathcal{E}_{w_i} = \{0\}\}$  and  $\mathcal{K}_1 \triangleq [K] \setminus \mathcal{K}_0$ .

**Lemma 12** (A strictly positive representative). *The saddle-point set  $\mathcal{E}$  is nonempty, compact, and convex. Consequently, there exists  $(\mathbf{w}^*, \boldsymbol{\mu}^*) \in \mathcal{E}$  such that  $(\mathbf{w}^*)_i = 0$  for all  $i \in \mathcal{K}_0$  and  $(\mathbf{w}^*)_i > 0$  for all  $i \in \mathcal{K}_1$ .*

*Proof.* Nonemptiness follows from Sion's minimax theorem (using Lemma 1). Compactness follows from a classical saddle-point theorem; see, e.g., Bertsekas et al. (2003, Prop. 2.6.9).

For convexity, take  $(\mathbf{w}^1, \boldsymbol{\mu}^1), (\mathbf{w}^2, \boldsymbol{\mu}^2) \in \mathcal{E}$  and  $\lambda \in [0, 1]$ , and set  $\mathbf{w}^\lambda = \lambda \mathbf{w}^1 + (1 - \lambda) \mathbf{w}^2$  and  $\boldsymbol{\mu}^\lambda = \lambda \boldsymbol{\mu}^1 + (1 - \lambda) \boldsymbol{\mu}^2$ . For any  $\mathbf{w}' \in \Delta_K$ , linearity of  $F(\mathbf{w}', \cdot)$  in  $\boldsymbol{\mu}$  gives

$$F(\mathbf{w}', \boldsymbol{\mu}^\lambda) = \lambda F(\mathbf{w}', \boldsymbol{\mu}^1) + (1 - \lambda) F(\mathbf{w}', \boldsymbol{\mu}^2) \leq \lambda F(\mathbf{w}^1, \boldsymbol{\mu}^1) + (1 - \lambda) F(\mathbf{w}^2, \boldsymbol{\mu}^2),$$

where the inequality uses that  $(\mathbf{w}^k, \boldsymbol{\mu}^k)$  is a saddle point ( $k = 1, 2$ ). Next, linearity in  $\boldsymbol{\mu}$  and concavity of  $F(\cdot, \boldsymbol{\mu}^\lambda)$  yield

$$\lambda F(\mathbf{w}^1, \boldsymbol{\mu}^1) + (1 - \lambda) F(\mathbf{w}^2, \boldsymbol{\mu}^2) = \lambda F(\mathbf{w}^1, \boldsymbol{\mu}^\lambda) + (1 - \lambda) F(\mathbf{w}^2, \boldsymbol{\mu}^\lambda) \leq F(\mathbf{w}^\lambda, \boldsymbol{\mu}^\lambda).$$

Combining the two displays gives  $F(\mathbf{w}', \boldsymbol{\mu}^\lambda) \leq F(\mathbf{w}^\lambda, \boldsymbol{\mu}^\lambda)$  for all  $\mathbf{w}'$ . Similarly, for any  $\boldsymbol{\mu}' \in \Delta_{|\mathcal{X}|}$ , convexity of  $F(\mathbf{w}^\lambda, \cdot)$  implies  $F(\mathbf{w}^\lambda, \boldsymbol{\mu}^\lambda) \leq F(\mathbf{w}^\lambda, \boldsymbol{\mu}')$ . Hence  $(\mathbf{w}^\lambda, \boldsymbol{\mu}^\lambda) \in \mathcal{E}$ .

For the final claim, for each  $i \in \mathcal{K}_1$  choose  $(\mathbf{w}^{(i)}, \boldsymbol{\mu}^{(i)}) \in \mathcal{E}$  with  $(\mathbf{w}^{(i)})_i > 0$ , which exists by definition of  $\mathcal{K}_1$ . Define

$$(\mathbf{w}^*, \boldsymbol{\mu}^*) \triangleq \frac{1}{|\mathcal{K}_1|} \sum_{i \in \mathcal{K}_1} (\mathbf{w}^{(i)}, \boldsymbol{\mu}^{(i)}).$$

By convexity,  $(\mathbf{w}^*, \boldsymbol{\mu}^*) \in \mathcal{E}$ . If  $j \in \mathcal{K}_0$ , then  $(\mathbf{w})_j = 0$  for all  $(\mathbf{w}, \boldsymbol{\mu}) \in \mathcal{E}$ , hence  $(\mathbf{w}^*)_j = 0$ . If  $j \in \mathcal{K}_1$ , then

$$(\mathbf{w}^*)_j = \frac{1}{|\mathcal{K}_1|} \sum_{i \in \mathcal{K}_1} (\mathbf{w}^{(i)})_j \geq \frac{1}{|\mathcal{K}_1|} (\mathbf{w}^{(j)})_j > 0,$$

so  $(\mathbf{w}^*)_j > 0$ . □

**Theorem 3.** *Under Assumptions 1–7, let  $\{\mathbf{w}_n\}_{n \geq 0}$  be the cost allocation sequence generated by Algorithm 1 and iteration (11). Then,  $\min_{x \in \mathcal{X}} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}_n) \rightarrow \Gamma^*$  almost surely. As a consequence, for  $\{\mathbf{p}_n\}_{n \geq 0}$  from Algorithm 1, we also have  $\min_{x \in \mathcal{X}} D_x(\mathbf{p}_n; \boldsymbol{\theta}_n) \rightarrow \Gamma^*$  almost surely.*

*Proof.* Define the KL potential on the active face

$$V_n \triangleq D_{\text{KL}}(\mathbf{w}^* \parallel \mathbf{w}_n) = \sum_{i \in \mathcal{K}_1} w_i^* \log \frac{w_i^*}{w_{n,i}} \geq 0.$$

On non-forced rounds, using the PAN update,

$$\begin{aligned}
& \mathbb{E}[V_{n+1} - V_n \mid \mathcal{H}_n] \\
&= \mathbb{E}\left[-\sum_{i \in \mathcal{K}_1} w_i^* \log \frac{w_{n+1,i}}{w_{n,i}} \mid \mathcal{H}_n\right] \\
&= -\sum_{i \in \mathcal{K}_1} w_i^* \log \frac{w_{n,i} + \tilde{\alpha}_{n+1}(\tilde{h}_{n,i} - w_{n,i})}{w_{n,i}} \\
&= -\sum_{i \in \mathcal{K}_1} w_i^* \log \frac{w_{n,i} + \tilde{\alpha}_{n+1}(h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n) - w_{n,i}) + \tilde{\alpha}_{n+1}(\tilde{h}_{n,i} - h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n))}{w_{n,i}} \\
&= -\sum_{i \in \mathcal{K}_1} w_i^* \log \left(1 + \tilde{\alpha}_{n+1}\left(\frac{[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} - 1\right) + \tilde{\alpha}_{n+1}\left(\frac{\tilde{h}_{n,i}}{w_{n,i}} - \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)}{w_{n,i}}\right)\right). \quad (46)
\end{aligned}$$

From assumption 7 and Proposition 7,

$$\|\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)\|_{\infty} \leq M, \quad D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n) \geq D_*/2 \quad \text{when } n \geq n_0, \text{ a finite random time.} \quad (47)$$

Then by (16) in Lemma 2, we have

$$\left|\frac{\tilde{h}_{n,i}}{w_{n,i}} - \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)}{w_{n,i}}\right| \leq \frac{c_{\max}}{nc_{\min}} \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)}{w_{n,i}} = \frac{c_{\max}}{nc_{\min}} \frac{[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (48)$$

Thus using  $-\log(1+u) \leq -u + u^2$  (for  $u \geq -\frac{1}{2}$ ) to (46) we have

$$\mathbb{E}[V_{n+1} - V_n \mid \mathcal{H}_n] \leq -\tilde{\alpha}_{n+1} \sum_{i \in \mathcal{K}_1} w_i^* \left(\frac{[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} - 1\right) + Q_n,$$

where all other terms are collected by

$$\begin{aligned}
Q_n &\triangleq \tilde{\alpha}_{n+1} \left| \frac{\tilde{h}_{n,i}}{w_{n,i}} - \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)}{w_{n,i}} \right| \\
&\quad + \tilde{\alpha}_{n+1}^2 \left[ \left| \frac{[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} - 1 \right| + \left| \frac{\tilde{h}_{n,i}}{w_{n,i}} - \frac{h_i^{x_n}(\mathbf{w}_n; \boldsymbol{\theta}_n)}{w_{n,i}} \right| \right]^2
\end{aligned}$$

Using (47) and (48) again, we conclude that when  $n \geq n_0$

$$Q_n \leq \frac{2c_{\max}^2 M}{n^2 c_{\min}^2 D_*} + \frac{c_{\max}^2}{n^2 c_{\min}^2} \left( \max \left\{ \left(\frac{2M}{D_*}\right)^2, 1 \right\} + \frac{2c_{\max} M}{nc_{\min} D_*} \right) \quad \text{and} \quad \sum_{n=1}^{\infty} Q_n < \infty.$$

By Lemma 9 that  $\boldsymbol{\theta}_n \rightarrow \boldsymbol{\theta}$  almost surely, continuity of  $C_i(\cdot)$  and lemma 2, we have for any  $\epsilon > 0$  there exists  $n_1(\epsilon)$  such that for all  $n \geq n_1(\epsilon)$ ,

$$\sup_{n \geq n_1(\epsilon)} \left| D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n) - \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) \right| \leq \epsilon/2, \quad \sup_{n \geq n_1(\epsilon)} |D_{x_n}^w(\mathbf{w}^*; \boldsymbol{\theta}_n) - D_{x_n}^w(\mathbf{w}^*; \boldsymbol{\theta})| \leq \epsilon/2. \quad (49)$$

Consequently, for all such  $n$ ,

$$-\sum_{i \in \mathcal{K}_1} w_i^* \left( \frac{[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} - 1 \right) = -\sum_{i \in [K]} w_i^* \left( \frac{[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} - 1 \right) \quad (50)$$

$$= \sum_{i \in [K]} (w_{n,i} - w_i^*) \frac{[\nabla_{\mathbf{w}} D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)]_i}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} \quad (51)$$

$$\leq \frac{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n) - D_{x_n}^w(\mathbf{w}^*; \boldsymbol{\theta}_n)}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} \quad (52)$$

$$\leq \frac{\min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) - D_{x_n}^w(\mathbf{w}^*; \boldsymbol{\theta}) + \epsilon}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} \quad (53)$$

$$\leq \frac{\min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) - F^* + \epsilon}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)}. \quad (54)$$

Here, (50) follows from  $w_i^* = 0$  when  $i \in \mathcal{K}_0 = [K] \setminus \mathcal{K}_1$ ; (51) uses positive homogeneity of first degree together with Euler's identity; (52) follows from the concavity of  $D_{x_n}^w(\cdot; \boldsymbol{\theta}_n)$ ; (53) follows from sufficient exploration approximation (49); and (54) follows from the definition of the equilibrium value  $F^* = F(\mathbf{w}^*, \boldsymbol{\mu}^*) \leq F(\mathbf{w}^*, \mathbf{e}_{x_n}) = D_{x_n}^w(\mathbf{w}^*; \boldsymbol{\theta})$ .

Thus for non-forced exploration steps, when  $n \geq n_0(\epsilon)$ ,

$$\mathbb{E}[V_{n+1} - V_n \mid \mathcal{H}_n] \leq -\tilde{\alpha}_n \frac{F^* - \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) - \epsilon}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} + Q_n.$$

For forced exploration steps,

$$\begin{aligned} \mathbb{E}[V_{n+1} - V_n \mid \mathcal{H}_n] &= \mathbb{E} \left[ -\sum_{i \in \mathcal{K}_1} w_i^* \log \frac{w_{n+1,i}}{w_{n,i}} \mid \mathcal{H}_n \right] \\ &\leq -\sum_{i \in \mathcal{K}_1} w_i^* \log (1 - \tilde{\alpha}_n) = -\log (1 - \tilde{\alpha}_n) \\ &\leq \frac{\tilde{\alpha}_n}{1 - \tilde{\alpha}_n} \leq 2\tilde{\alpha}_n, \end{aligned}$$

then when  $n \geq n_0(\epsilon) \in \mathcal{L}$ ,

$$\mathbb{E}[V_{n+1} - V_n \mid \mathcal{H}_n] \leq -\tilde{\alpha}_n \frac{F^* - \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) - \epsilon}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} \mathsf{NF}_n + Q_n \mathsf{NF}_n + 2\tilde{\alpha}_n \mathsf{F}_n.$$

We claim that  $\limsup_{n \rightarrow \infty} \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) = F^*$  almost surely and then prove by contradiction. Suppose there exists  $\delta > 0$  and  $n_2(\delta)$  such that  $F^* - D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}) \geq \delta$  for  $n \geq n_2(\delta)$ . Take  $\epsilon = \delta/2$ , then when  $n \geq N_0 \triangleq \max\{n_0, n_1(\delta/2), n_2(\delta)\}$ ,  $F^* - \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) - \delta/2 \geq \delta/2 \geq 0$ . Define non-negative process

$$U_n \triangleq \tilde{\alpha}_n \frac{F^* - \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) - \delta/2}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} \mathsf{NF}_n, \quad W_n \triangleq Q_n \mathsf{NF}_n + 2\tilde{\alpha}_n \mathsf{F}_n.$$

Then  $\mathbb{E}[V_{n+1} - V_n \mid \mathcal{H}_n] \leq -U_n + W_n$  when  $n \geq N_0$  with  $\sum_{n=N_0} \mathsf{W}_n < \infty$ . By Robbins–Siegmund almost-supermartingale convergence, see, e.g., Borkar (2008, Theorem C.5),  $\sum_{n=N_0} U_n < \infty$  almost surely. This contradicts with

$$\sum_{n=N_0} U_n \geq \sum_{n=N_0} \frac{\delta/2}{D_{x_n}^w(\mathbf{w}_n; \boldsymbol{\theta}_n)} \mathsf{NF}_n \geq \sum_{n=N_0} \frac{\delta/2}{F^* + 1} \mathsf{NF}_n = \infty.$$

Therefore  $\limsup_{n \rightarrow \infty} \min_{x \in \mathcal{X}(\boldsymbol{\theta}_n)} D_x^w(\mathbf{w}_n; \boldsymbol{\theta}) = F^*$  almost surely and the the convergence of the entire value sequence follows from Proposition 6. The second part of the theorem follows from the identity

$$\begin{aligned} D_x(\mathbf{p}_n; \boldsymbol{\theta}_n) &= \frac{1}{\bar{C}_{\boldsymbol{\theta}_n}(\mathbf{p}_n)} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta}_n)} \sum_{i=1}^K p_{n,i} \text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}, i}) \\ &= \frac{\bar{C}_{\boldsymbol{\theta}}(\mathbf{p}_n)}{\bar{C}_{\boldsymbol{\theta}_n}(\mathbf{p}_n)} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x(\boldsymbol{\theta}_n)} \sum_{i=1}^K w_{n,i} \frac{\text{KL}(P_{\boldsymbol{\theta}_n, i} \| P_{\boldsymbol{\vartheta}, i})}{C_i(\boldsymbol{\theta})}, \end{aligned}$$

together with the continuity of  $C_i(\cdot)$ , which implies  $|D_x(\mathbf{p}_n; \boldsymbol{\theta}_n) - D_x^w(\mathbf{w}_n; \boldsymbol{\theta}_n)| \rightarrow 0$ .  $\square$

**Remark 4** (Boundary separation on potentially active arms). *When the saddle-point set is not a singleton, different algorithms may converge to different saddle points. For instance, it may happen that  $\mathcal{E}_{p_i} = [0, 1/2]$  for some coordinate  $i$ , so both boundary ( $p_i = 0$ ) and interior ( $p_i > 0$ ) saddle points exist. The KL-potential argument used in the proof of Theorem 3 implies a simple but useful boundary-separation property for coordinates that are positive in a saddle-point representative.*

Let  $(\mathbf{w}^*, \boldsymbol{\mu}^*)$  be chosen as in Lemma 12, so that  $(\mathbf{w}^*)_i > 0$  for all  $i \in \mathcal{K}_1$ , and recall the KL potential on the active face

$$V_n \triangleq D_{\text{KL}}(\mathbf{w}^* \| \mathbf{w}_n) = \sum_{i \in \mathcal{K}_1} w_i^* \log \frac{w_i^*}{w_{n,i}}.$$

Robbins–Siegmund yields that  $V_n$  converges almost surely in Theorem 3; in particular,  $V_n$  is almost surely bounded, say  $\sup_n V_n \leq \bar{V} < \infty$  on an event of probability one. Fix any  $i \in \mathcal{K}_1$ . Since each term in the sum defining  $V_n$  is nonnegative, we have

$$w_i^* \log \frac{w_i^*}{w_{n,i}} \leq V_n \leq \bar{V} \quad \forall n,$$

hence  $\log \frac{w_i^*}{w_{n,i}} \leq \bar{V}/w_i^*$  and therefore

$$w_{n,i} \geq w_i^* \exp\left(-\frac{\bar{V}}{w_i^*}\right) > 0 \quad \forall n.$$

Consequently,  $\inf_n w_{n,i} > 0$  for all  $i \in \mathcal{K}_1$ , and every accumulation point of  $(\mathbf{w}_n)$  assigns strictly positive mass to all coordinates in  $\mathcal{K}_1$ . Equivalently, each potentially active arm (i.e., one that is positive in some saddle point) receives a linear number of samples under the PAN/IDS dynamics started from the interior.

We view this as a robustness feature of IDS-style sampling dynamics: it prevents the algorithm from selecting sparse boundary saddle points that would “starve” potentially active arms (e.g.,  $p_i \rightarrow 0$ , leading to sublinear sampling).

## F Examples

### F.1 Uniform Allocation Can be Arbitrarily Bad

Consider a linear bandit in dimension  $d = 2$  with unit observation variances and unknown parameter  $\boldsymbol{\theta} = (1, -C)^\top$ . There are  $K \geq 2$  arms:  $a_1 = e_1$ ,  $a_2 = e_1 + \varepsilon e_2$ , and  $a_k = \alpha e_1$  for  $k = 3, \dots, K$ , with

fixed constants  $\alpha \in (0, 1)$  and a small  $0 < \varepsilon < (1 - \alpha)/C$ . The reward means are  $\mu_i = a_i^\top \theta$ :

$$\mu_1 = 1, \quad \mu_2 = 1 - C\varepsilon, \quad \mu_k = \alpha \quad (k \geq 3).$$

Hence arm 1 is optimal and the only *near rival* is arm 2:

$$\delta_2 = \mu_1 - \mu_2 = C\varepsilon, \quad \delta_k = \mu_1 - \mu_k = 1 - \alpha > \delta_2, \quad k = 3, 4, \dots, K.$$

For best arm  $I^* = 1$ , define comparison directions  $d_j = a_1 - a_j$  for  $j \neq 1$ . Then

$$d_2 = a_1 - a_2 = \begin{bmatrix} 0 \\ -\varepsilon \end{bmatrix} = -\varepsilon e_2, \quad d_k = a_1 - a_k = \begin{bmatrix} 1 - \alpha \\ 0 \end{bmatrix} = (1 - \alpha) e_1 \quad (k \geq 3).$$

For an allocation  $\mathbf{p} = (p_1, \dots, p_K) \in \Delta_K$ , the (normalized) design matrix is

$$V_{\mathbf{p}} = \sum_{i=1}^K p_i a_i a_i^\top = p_1 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + p_2 \begin{bmatrix} 1 & \varepsilon \\ \varepsilon & \varepsilon^2 \end{bmatrix} + \left( \sum_{k=3}^K p_k \right) \alpha^2 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Let  $S \triangleq \sum_{k=3}^K p_k$  and  $s \triangleq p_1 + p_2 + \alpha^2 S$ . Then

$$V_{\mathbf{p}} = \begin{bmatrix} s & \varepsilon p_2 \\ \varepsilon p_2 & \varepsilon^2 p_2 \end{bmatrix}, \quad \det(V_{\mathbf{p}}) = \varepsilon^2 p_2 (s - p_2) = \varepsilon^2 p_2 (p_1 + \alpha^2 S).$$

Hence

$$V_{\mathbf{p}}^{-1} = \frac{1}{\det(V_{\mathbf{p}})} \begin{bmatrix} \varepsilon^2 p_2 & -\varepsilon p_2 \\ -\varepsilon p_2 & s \end{bmatrix}.$$

For linear bandits with unit variances, the usual maximin information-rate objective against a specific rival  $j \neq I^*$  is

$$D_j(\mathbf{p}) = \frac{\Delta_j^2}{2d_j^\top V_{\mathbf{p}}^{-1} d_j}.$$

We have

$$\begin{aligned} D_2(\mathbf{p}) &= \frac{(C\varepsilon)^2}{2} \cdot \frac{p_2(s - p_2)}{s} \\ D_k(\mathbf{p}) &= \frac{(1 - \alpha)^2}{2} \cdot \frac{s - p_2}{(1 - \alpha)^2} = \frac{s - p_2}{2} = \frac{p_1 + \alpha^2 S}{2} \quad (k \geq 3). \end{aligned}$$

Consider uniform allocation, i.e.,  $p_i = 1/K$  for all  $i$ . Then  $S = (K - 2)/K$  and

$$s = \frac{2}{K} + \alpha^2 \frac{K - 2}{K} = \frac{2 + \alpha^2(K - 2)}{K}, \quad s - p_2 = \frac{1 + \alpha^2(K - 2)}{K}.$$

Therefore

$$\begin{aligned} D_{\text{unif}}(2) &= \frac{(C\varepsilon)^2}{2} \cdot \frac{\frac{1}{K} \cdot \frac{1 + \alpha^2(K - 2)}{K}}{\frac{2 + \alpha^2(K - 2)}{K}} = \frac{(C\varepsilon)^2}{2} \cdot \frac{1 + \alpha^2(K - 2)}{K[2 + \alpha^2(K - 2)]} \cdot \frac{1}{K} \\ &= \Theta\left(\frac{(C\varepsilon)^2}{K}\right) \quad \text{as } K \rightarrow \infty, \end{aligned}$$

$$D_{\text{unif}}(k) = \frac{s - p_2}{2} = \frac{1 + \alpha^2(K - 2)}{2K} \rightarrow \frac{\alpha^2}{2} \quad \text{as } K \rightarrow \infty.$$

Hence for small  $\varepsilon$ , the *bottleneck* under uniform is arm 2:

$$\min_{j \neq 1} D_{\text{unif}}(j) = D_{\text{unif}}(2) = \Theta\left(\frac{(C\varepsilon)^2}{K}\right).$$

Alternatively, concentrate on arms 1 and 2 only: set  $S = 0$ , so  $s = p_1 + p_2 = 1$  and  $s - p_2 = 1 - p_2$ .

Then

$$D_2(\mathbf{p}) = \frac{(C\varepsilon)^2}{2}p_2(1 - p_2), \quad D_k(\mathbf{p}) = \frac{1 - p_2}{2} \quad (k \geq 3).$$

The min over  $j \neq 1$  is dominated by arm 2 (since  $\varepsilon$  is small). Maximizing  $p_2(1 - p_2)$  gives  $p_2^* = \frac{1}{2}$ , so

$$\max_{\mathbf{p} \in \Delta_K} \min_{j \neq 1} D_j(\mathbf{p}) = \frac{(C\varepsilon)^2}{8}$$

The ratio between the optimal value and the uniform min-payoff is at least

$$\frac{\frac{(C\varepsilon)^2}{8}}{D_{\text{unif}}(2)} = \frac{(C\varepsilon)^2/8}{\frac{(C\varepsilon)^2}{2} \cdot \frac{1 + \alpha^2(K-2)}{K[2 + \alpha^2(K-2)]} \cdot \frac{1}{K}} = \frac{K}{4} \cdot \frac{2 + \alpha^2(K-2)}{1 + \alpha^2(K-2)} \xrightarrow[K \rightarrow \infty]{} \infty.$$

Thus the sequence of instances indexed by  $K$  (with fixed  $\alpha \in (0, 1)$ ,  $C > 0$ , and  $\varepsilon < (1 - \alpha)/C$ ) makes the uniform allocation *arbitrarily worse* than the maximin (IDS-style) allocation.

All distractors lie on the  $e_1$  axis, so they bring no direct information in the distinguishing  $e_2$  direction that separates the optimal arm 1 from the hard rival 2 (whose difference vector is  $d_2 = -\varepsilon e_2$ ). Uniform wastes most samples on these distractors as  $K$  grows, driving the min-payoff down to  $O((C\varepsilon)^2/K)$ , while the optimal design focuses on arms 1 and 2 and stays at  $\Theta((C\varepsilon)^2)$ .

## F.2 $\beta$ -Tuned Top-Two Algorithm Can be Arbitrarily Bad

Consider a linear bandit problem with unit variances and an unknown parameter  $\theta = (-1, -2)^\top \in \mathbb{R}^2$ . There are four arms defined by

$$\mathbf{a}_1 = \mathbf{e}_1, \quad \mathbf{a}_2 = \mathbf{e}_2, \quad \mathbf{a}_3 = l\mathbf{e}_1, \quad \mathbf{a}_4 = \mathbf{e}_1 + l\mathbf{e}_2.$$

For  $l > 1$ , the optimal arm is  $I^* = 1$ , and the suboptimal arms are  $j \in \{2, 3, 4\}$ . The directions associated with these arms are

$$\mathbf{d}_2 = (1, -1)^\top, \quad \mathbf{d}_3 = (1 - l, 0)^\top, \quad \mathbf{d}_4 = (0, -l)^\top,$$

with corresponding gaps  $\delta_2 = 1$ ,  $\delta_3 = l - 1$ , and  $\delta_4 = 2l$ .

For a sampling allocation  $\mathbf{p} \in \Delta_K$ , the design matrix is

$$V_{\mathbf{p}} = \begin{bmatrix} p_1 + l^2 p_3 + p_4 & lp_4 \\ lp_4 & p_2 + l^2 p_4 \end{bmatrix},$$

and its determinant is

$$\det(V_{\mathbf{p}}) = l^4 p_3 p_4 + l^2(p_1 p_4 + p_2 p_3) + p_1 p_2 + p_2 p_4.$$

The inverse of  $V_{\mathbf{p}}$  is given by

$$V_{\mathbf{p}}^{-1} = \frac{1}{\det(V_{\mathbf{p}})} \begin{bmatrix} p_2 + l^2 p_4 & -l p_4 \\ -l p_4 & p_1 + l^2 p_3 + p_4 \end{bmatrix}.$$

The payoff values for each suboptimal arm are:

$$\begin{aligned} D_2(\mathbf{p}) &= \frac{\det(V_{\mathbf{p}})}{2(p_1 + p_2 + l^2 p_3 + (l+1)^2 p_4)}, \\ D_3(\mathbf{p}) &= \frac{\det(V_{\mathbf{p}})}{2(p_2 + l^2 p_4)}, \\ D_4(\mathbf{p}) &= \frac{2 \det(V_{\mathbf{p}})}{p_1 + l^2 p_3 + p_4}. \end{aligned}$$

Note that  $D_2(\mathbf{p}) < \min \{D_3(\mathbf{p}), D_4(\mathbf{p})\}$  for all  $\mathbf{p}$ . Therefore, the maximin value is

$$F^* = \frac{1}{2} \max_{\mathbf{p} \in \Delta_4} \frac{\det(V_{\mathbf{p}})}{p_1 + p_2 + l^2 p_3 + (l+1)^2 p_4} \geq D_2((0, 0, 1/2, 1/2)) = \frac{l^4}{4(2l^2 + 2l + 1)}.$$

For top-two Thompson sampling with tuning parameter  $\beta = 1/2$ , the sampling allocation converges to  $(1/2, 1/2, 0, 0)$ , yielding a maximin value of  $1/8$ . In this case, the  $\beta$ -tuned top-two Thompson sampling can be arbitrarily worse than the IDS selection rule.

## G Additional Numerical Examples

We consider a simple  $d = 3$ ,  $K = 6$  linear instance in which the best arm is  $a_1 = e_1$  and the closest competitor is  $a_2 = e_1 + \frac{1}{2}e_2$  under  $\boldsymbol{\theta} = (1, -0.02, 0)$ , so the gap is only  $\langle a_1 - a_2, \boldsymbol{\theta} \rangle = 0.01$  and identifying the best arm requires accurately estimating the *nuisance* coordinate  $\theta_2$ . Crucially, there are two arms aligned with  $e_2$ , namely  $a_3 = a_4 = e_2$ , but with different noise levels: arm 3 has low variance  $\sigma_3^2 = 0.04$  while arm 4 has variance 1.

In the *equal-cost* version (all  $c_i \equiv 1$ ), the low-variance arm 3 is genuinely preferable, and LinGapE behaves similarly to PAN and Track-and-Stop: the median stopping times are 460 (PAN), 450 (Track-and-Stop), and 595 (LinGapE) pulls (Figure 2, Left Panel).

We then modify *only* one number: we set the cost of arm 3 to  $c_3 = 100$  while leaving the arms,  $\boldsymbol{\theta}$ , and variances unchanged. This creates a “cost trap”: arm 3 is  $25 \times$  less noisy but  $100 \times$  more expensive, so it is *four times worse* per unit budget than its cheap duplicate arm 4. A cost-aware method should therefore switch almost entirely to sampling arm 4 to learn  $\theta_2$  efficiently in budget.

The results are collected in Figure 2, Right Panel. Empirically, PAN and Track-and-Stop remain stable, reaching the target posterior error after about  $1.25 \times 10^4$  and  $1.10 \times 10^4$  budget units in median, respectively. In contrast, a naive (cost-agnostic) LinGapE implementation becomes dramatically cost-inefficient: its median budget to target jumps to  $3.42 \times 10^4$  (about a  $3 \times$  increase relative to PAN/Track-and-Stop and  $\sim 60 \times$  relative to its equal-cost median when “budget” coincides with pulls), with a heavy tail (95th percentile  $1.28 \times 10^5$ ) and occasional failures to reach the target within the budget cap (0.7% of runs).

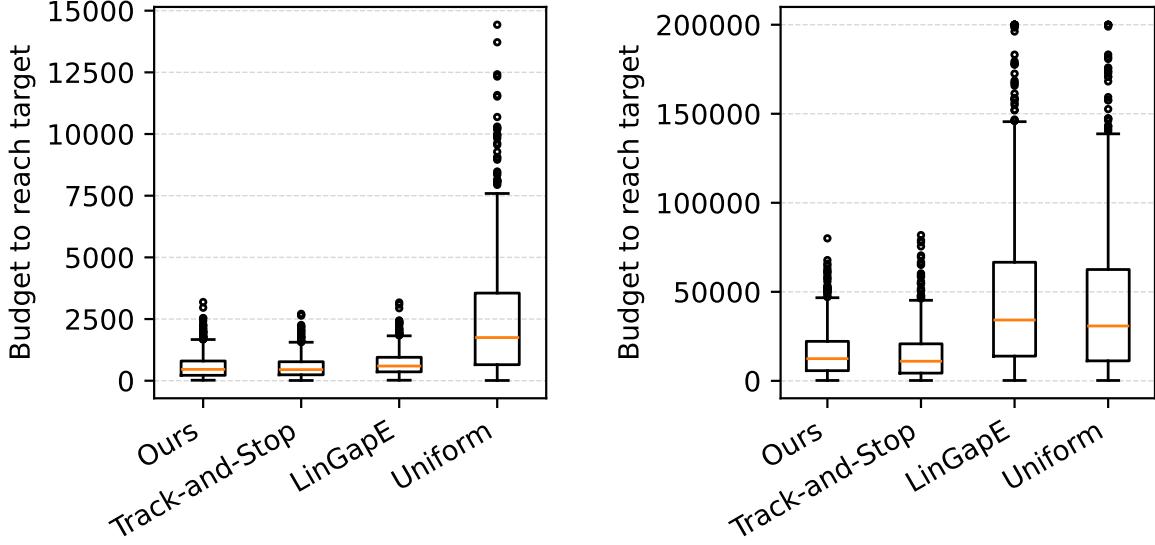


Figure 2: Budget needed to reach posterior threshold. Left: uniform cost. Right: Heterogeneous cost.

This experiment highlights that heterogeneous costs can qualitatively change the optimal allocation, and that plugging a standard LinGapE rule into a cost-budgeted setting without explicitly incorporating costs can lead to a severe performance collapse even when the underlying geometry and noise model are unchanged.

## H Auxiliary facts

In this section, we compile several classical results on set-valued maps and differential inclusions that facilitate our proofs.

**Envelope theorem.** Let  $X$  be a choice set and let  $t \in [0, 1]$  be the relevant parameter. Consider the parameterized objective function  $f : X \times [0, 1] \rightarrow \mathbb{R}$ , and define the value function  $v : [0, 1] \rightarrow \mathbb{R}$  and the optimal choice correspondence (set-valued function)  $X^*$  by

$$v(t) = \sup_{x \in X} f(x, t) \quad \text{and} \quad X^*(t) = \{x \in X : f(x, t) = v(t)\}.$$

The Envelope theorem, Milgrom and Segal (2002, Theorem 1), states that: Assume that for a given  $t \in [0, 1]$  and for some  $x^* \in X^*(t)$  the partial derivative  $f_t(x^*, t)$  exists. If  $v$  is differentiable at  $t$ , then  $v'(t) = f_t(x^*, t)$ .

**Upper hemicontinuity for compact-valued correspondence.** Beavis and Dobbs (1990, Theorem 3.2) provide an equivalent characterization of upper hemicontinuity for compact-valued set-valued map  $F : X \rightrightarrows Y$ .  $F$  is upper hemicontinuous at  $x \in X$  if, and only if, for every sequence  $\{x_n\}$  converging to  $x$  and every sequence  $\{y_n\}$  with  $y_n \in F(x_n)$ , there exists a converging subsequence of  $\{y_n\}$  whose limit belongs to  $F(x)$ .

**Berge's Maximum Theorem.** Below is a rephrase of Beavis and Dobbs (1990, Theorem 3.6). Let  $X \subset \mathbb{R}^m$ ,  $Y \subset \mathbb{R}^k$  and  $\Xi : X \rightrightarrows Y$  be a set-valued map with nonempty, compact values. Let  $f : X \times Y \rightarrow \mathbb{R}$  be a continuous function. Define the set-valued function  $M : X \rightrightarrows Y$ , the maximizers  $M(x) \triangleq \arg \max_{y \in \Xi(x)} f(x, y)$ , and the corresponding value function  $v : X \rightarrow \mathbb{R}$  by  $v(x) = \max_{y \in \Xi(x)} f(x, y)$ . If  $\Xi$  is continuous at  $x$ , then  $v$  is continuous at  $x$  and the set-valued function  $M$  is closed, compact-valued and upper hemicontinuous at  $x$ .

**Existence of solutions to a differential inclusion.** Consider the autonomous differential inclusion

$$\dot{\mathbf{x}}(t) \in F(\mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0,$$

where  $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^m$  satisfies: (i)  $\text{graph}(F) = \{(\mathbf{x}, \mathbf{y}) : \mathbf{y} \in F(\mathbf{x})\}$  is closed; (ii)  $F(\mathbf{x})$  is nonempty, compact, and convex for every  $\mathbf{x} \in \mathbb{R}^m$ ; (iii) there exists  $c > 0$  such that  $\sup_{\mathbf{z} \in F(\mathbf{x})} \|\mathbf{z}\| \leq c(1 + \|\mathbf{x}\|)$  for all  $\mathbf{x} \in \mathbb{R}^m$ . Then, for every initial condition  $\mathbf{x}_0 \in \mathbb{R}^m$ , there exists an absolutely continuous trajectory  $\mathbf{x} : [0, \infty) \rightarrow \mathbb{R}^m$  such that  $\dot{\mathbf{x}}(t) \in F(\mathbf{x}(t))$  for almost every  $t \geq 0$ . See, e.g., Aubin and Cellina (1984, Chapter 2.1).

**Existence of measurable selections.** Aubin and Cellina (1984, Corollary 1, Section 1.14) states that, let  $f : X \times U \rightarrow X$  be continuous, where  $U$  is a compact separable metric space. Assume that there exist an interval  $I$  and an absolutely continuous function  $x : I \rightarrow \mathbb{R}^n$ , such that

$$x'(t) \in f(x(t), U) \quad \text{for almost every } t \in I.$$

Then, there exists a Lebesgue measurable function  $u : I \rightarrow U$  such that

$$x'(t) = f(x(t), u(t)) \quad \text{for almost every } t \in I.$$