The Price of Knowledge: Optimal Algorithms for Costly Bandits

Felix Schur[*1] **Jesus Lago**[2] **Tanner Fiez**[2]

[1]ETH Zurich, Department of Mathematics
[2]Amazon

Abstract

We study stochastic bandits in which observing a reward is optional but incurs an action-dependent cost. This setting captures applications where feedback acquisition (e.g., human evaluation or randomized testing) is expensive, and the learner must trade off exploration, exploitation, and observation cost. We formulate regret to include both reward loss and the cumulative cost of requested observations. Our first result is structural: for minimizing regret, it is without loss of generality to consider two-phase policies that first request observations during an exploration phase and then commit to a single action without further observations. Building on this reduction, we introduce two cost-sensitive complexity measures that extend maximum information gain: a cost-adjusted information gain $\Gamma_T(c)$ for minimax analysis, and a cost- and gap-adjusted information gain $\Gamma_T^{\mathrm{gap}}(\Delta_c)$ for instance-dependent analysis. Using these quantities, we develop two Gaussian-process-based algorithms, `C3-GP` and `GP-C-LUCB`, and derive regret upper bounds for correlated-action settings with heterogeneous (i.e. action-dependent) observation costs. In the finite independent-action setting, we further prove matching lower and upper bounds (up to constants/logarithmic factors), yielding a tight characterization of both minimax and instance-dependent regret in terms of the proposed cost-aware complexity measures.

1 INTRODUCTION

In many sequential decision problems, the acquisition of feedback is itself costly, and this cost can vary substantially between actions. In online advertising, estimating the causal effect of an intervention may require a randomized controlled trial, but the cost of such a trial depends on the audience being targeted: testing an ad on a premium or narrowly defined demographic can be much more expensive than testing it on a broad, inexpensive audience. A similar phenomenon arises in the evaluation of generative AI systems, where obtaining human feedback on a short factual response is cheap, while evaluating a long mathematical derivation, legal argument, or piece of code may require substantial expert effort. In both examples, the value of an action is tied not only to its expected reward, but also to the price of learning about it. This creates a richer exploration problem: the learner must decide not only which actions are promising, but also whether the information gained from observing a particular action is worth its action-dependent cost.

The stochastic multi-armed bandit problem formalizes sequential decision-making under uncertainty: over T rounds, a learner repeatedly selects an action and seeks to maximize cumulative reward. In the classical stochastic setting, minimax regret scales as $\Theta(\sqrt{|\mathcal{X}|T})$, while instance-dependent regret grows only logarithmically in T under suitable gap assumptions when reward feedback is immediate and free [Lai and Robbins, 1985, Garivier and Moulines, 2008, Bubeck et al., 2012, Gerchinovitz and Lattimore, 2016]. In many applications, however, observing the reward of a chosen action incurs a non-negligible cost. The Bandits with Costly Reward Observations (BwCRO) model captures this by allowing the learner, after choosing an action, to either pay a cost to observe its reward or forgo feedback and save cost at the expense of learning [Schulze and Evans, 2018, Krueger et al., 2020, Tucker et al., 2023].

Existing work on BwCRO has largely focused on finite bandits with uniform observation costs and independent arms (we use "arm" and "action" interchangeably throughout). This misses two features that are often central in practice. First, observation costs can vary substantially between actions. Second, rewards can be correlated across similar ac-

[*]felix.schur@stat.math.ethz.ch. This work was done while Felix Schur was at Amazon.

tions, so that observing one action may reduce uncertainty about many others. These two aspects interact in an essential way: an expensive action need not be queried directly if its value can be inferred from cheaper correlated actions nearby. To address this, we study stochastic bandits with optional, action-dependent reward observations under both finite independent-arm models and correlated-action models.

Contributions. We make four main contributions. First, we prove a structural result showing that, for regret minimization in BwCRO, it is without loss of generality to restrict attention to two-phase policies: an initial exploration phase in which the learner may request reward observations, followed by a commitment phase in which it stops querying and repeatedly plays a single action. Second, we introduce two cost-aware complexity measures, $\Gamma_T(c)$ and $\Gamma_T^{\mathrm{gap}}(\Delta_c)$, which extend maximum information gain to capture how statistical difficulty interacts with observation cost. Third, we develop two Gaussian-process-based algorithms, `C3-GP` and `GP-C-LUCB`, and derive regret upper bounds for correlated-action settings with heterogeneous costs. Finally, in the finite independent-arm setting, we establish matching lower and upper bounds up to constants and logarithmic factors, showing that these complexity measures characterize both minimax and instance-dependent regret in that regime.

2 PROBLEM SETTING

For $n \in \mathbb{N}$ we write $[n] \coloneqq \{1, \dots, n\}$. Throughout, we employ the "soft-O" notation $\tilde{O}(\cdot)$ to suppress polylogarithmic factors (that is, to ignore any $\log$-terms).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{X}$ be an action set, and fix a time horizon $T \in \mathbb{N}$. We assume there exists an (unknown) reward function $f : \mathcal{X} \to \mathbb{R}$ and that, whenever the agent selects actions $x_1, \dots, x_T \in \mathcal{X}$, it observes noisy rewards

$$y_t \coloneqq f(x_t) + \varepsilon_t, \quad t \in [T],$$

where $\{\varepsilon_t\}_{t=1}^T$ are independent, mean-zero, σ-sub-Gaussian random variables.

Standard stochastic bandits. Define the instantaneous regret of choosing $x \in \mathcal{X}$ by $r(x) \coloneqq \max_{z \in \mathcal{X}} f(z) - f(x)$, and denote by $x^* \in \arg\max_{x \in \mathcal{X}} f(x)$ any maximizer of f (we assume the supremum is attained). Let $\mathcal{Y} \coloneqq \mathbb{R}$ denote the reward space, and define the history set $\mathcal{H} \coloneqq \bigcup_{n=0}^{T} (\mathcal{X} \times \mathcal{Y})^n$. A policy is a measurable mapping $\bar{\mathcal{A}} : \mathcal{H} \to \mathcal{X}$, which we interpret recursively by $x_t = \bar{\mathcal{A}}((x_1, y_1, \dots, x_{t-1}, y_{t-1}))$. The cumulative regret of any policy $\bar{\mathcal{A}}$ after $t \in [T]$ rounds is $\bar{R}_t(\bar{\mathcal{A}}) \coloneqq \sum_{s=1}^{t} r(x_s)$, where $x_s \in \mathcal{X}$ is the action chosen at time $s \in [T]$.

Bandits with Costly Reward Observations (BwCRO). In many applications, observing the reward y_t incurs a *known* cost

$$c : \mathcal{X} \to [0, \infty),$$

which we also call the labeling or observation cost or just "cost". At each round $t \in [T]$, after choosing $x_t \in \mathcal{X}$, the agent must decide whether to pay $c(x_t)$ to observe y_t ($d_t = 1$) or to forgo the observation ($d_t = 0$). We then define an extended (policy-dependent) history $H_t \coloneqq \big((x_s, y_s) \mid 1 \le s \le t,\ d_s = 1\big) \in \mathcal{H}, H_0 \coloneqq \emptyset$, and a BwCRO policy $\mathcal{A} : \mathcal{H} \to \mathcal{X} \times \{0,1\}, H_{t-1} \mapsto (x_t, d_t)$. In other words, $(x_t, d_t) = \mathcal{A}(H_{t-1})$, and if $d_t = 1$ then $H_t = (H_{t-1}, (x_t, y_t))$. We measure performance via the cost-aware cumulative regret

$$R_t(\mathcal{A}) \coloneqq \sum_{s=1}^{t} \Big[r(x_s) + c(x_s) \mathbb{1}\{d_s = 1\} \Big],$$

which reduces to the standard definition when all $c(x) \equiv 0$. For brevity, we often drop $\mathcal{A}$ whenever the policy is clear from context. Our aim is to design a policy $\mathcal{A}$ whose total regret $R_T(\mathcal{A})$ grows sublinearly in T. A key simplification arises:

Lemma 1 (Reduction to Two-Phase Policies)**.** *Let $\tilde{\mathcal{A}}$ be the class of all (possibly randomized) BwCRO policies that, at each round, may depend on the full past including past decisions $d_1, \dots, d_{t-1}$, past actions (labeled or not), and all observed rewards. Then for every policy $\mathcal{A} \in \tilde{\mathcal{A}}$ there exists a (possibly randomized) policy $\mathcal{A}'$ (with label decisions d'_t and actions x'_t) such that:*

- *(i) (Two-phase) There exists a (random) stopping time $\tau \in \{0, 1, \dots, T\}$ (adapted to A's internal randomness and labeled history (x'_s, y'_s)) with $d'_t = 1$ for all $t \le \tau$ and $d'_t = 0$ for all $t > \tau$, and $x'_t = x'_{\tau+1}$ for all $t > \tau$.*
- *(ii) (Label-only) $\mathcal{A}'$ can be written as a measurable map $\mathcal{A}' : \mathcal{H} \to \mathcal{X} \times \{0,1\}$ that depends only on the labeled history $H_{t-1} = ((x_s, y_s) \mid s < t,\ d'_s = 1)$.*
- *(iii) (No worse regret uniformly) For every bandit instance ν, $\mathbb{E}_\nu[R_T(\mathcal{A}')] \le \mathbb{E}_\nu[R_T(\mathcal{A})]$.*

Proof sketch. We construct $\mathcal{A}'$ by running a virtual copy of $\mathcal{A}$ with the same internal randomness. All virtual rounds in which $\mathcal{A}$ requests a label are executed first, in their original order; stationarity of the rewards and the fact that unlabeled rounds reveal no information preserve the distribution of the labeled history. After these labeled rounds, $\mathcal{A}'$ samples one action uniformly from the multiset of virtual unlabeled actions that $\mathcal{A}$ would have played and commits to it. By linearity of expectation, this randomized commitment has the same expected unlabeled regret as the original unlabeled sequence, while the labeled costs and labeled regret have the same distribution. The full coupling argument is given in the appendix.

Lemma 1 implies that it is sufficient for regret minimization in the BwCRO to focus on two-phase policies $\mathcal{A} : \mathcal{H} \to \mathcal{X} \times \{0, 1\}$:

- *Label-requesting phase* (exploration): the agent may pay costs to collect observations and improve its estimate of f.
- *No-label phase* (exploitation): once it stops requesting observations, it repeats a single action indefinitely.

Unlike classical bandits—where one balances exploration vs. exploitation—BwCRO involves a three-way trade-off among exploration, exploitation, and labeling costs.

3 PRELIMINARIES

We use Gaussian process (GP) regression as a unified inference model. This covers several standard bandit settings as special cases, including finite independent-arm bandits (delta kernel) and linear bandits (linear kernel). Our analysis only relies on the posterior mean/variance and a high-probability confidence event.

GP posterior notation. Fix a positive semidefinite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a noise proxy $\sigma^2 > 0$. At round t, let H_t denote the labeled history (as defined in Section 2), and let n_t be the number of labeled observations contained in H_t. If $H_t = (z_i, \tilde{y}_i)_{i=1}^{n_t}$, define $K_t := (k(z_i, z_j))_{i,j=1}^{n_t}$, and $k_t(x) := (k(z_1, x), \dots, k(z_{n_t}, x))^\top$. The GP posterior mean and variance after H_t are $\mu_t(x) := k_t(x)^\top (K_t + \sigma^2 I)^{-1} \tilde{y}_{1:n_t}$, and $\sigma_t^2(x) := k(x, x) - k_t(x)^\top (K_t + \sigma^2 I)^{-1} k_t(x)$. When $n_t = 0$, we use the convention $\mu_t(x) = 0$ and $\sigma_t^2(x) = k(x, x)$. For a confidence parameter $\beta_T > 0$, define the event $\mathcal{E} := \{\forall t \in [T],\ \forall x \in \mathcal{X} :\ |f(x) - \mu_{t-1}(x)| \le \sqrt{\beta_T}\sigma_{t-1}(x)\}$. Throughout, we assume that $\|f\|_k \le B < \infty$. Our regret bounds are stated on $\mathcal{E}$. Standard choices of β_T guarantee that $\mathcal{E}$ holds with high probability under common GP/RKHS assumptions. For $S \subseteq \mathcal{X}$, define the maximum information gain restricted to S by

$$\gamma_T(S) := \sup_{x_{1:T} \in S^T} \frac{1}{2} \log\det\Big(I_T + \sigma^{-2} K_T(x_{1:T})\Big),$$

where $K_T(x_{1:T})$ is the $T \times T$ kernel matrix with entries $k(x_i, x_j)$. We write $\gamma_T := \gamma_T(\mathcal{X})$. The maximum information gain is a commonly used complexity measure in the GP-bandit literature [Srinivas et al., 2009, Chowdhury and Gopalan, 2017].

Cost-adjusted information gain. We introduce cost-sensitive analogues of information gain to quantify the difficulty of the BwCRO problem. The key idea is that the difficulty measure should reflect how observation costs align with the statistical structure of the action space. If costs are high in regions that are easy to learn (for example, where arms are strongly correlated) and low in regions that are hard to learn (where correlations are weak), then the overall problem should be considered relatively easy, since costly exploration is needed only in low-complexity parts of the space. In contrast, if costs are high precisely in regions that are difficult to learn, then exploration becomes both statistically and economically expensive, and the problem should be deemed harder. We now introduce adjusted information-gain measures that formalize this intuition.

Definition 1 (Cost-adjusted information gain)**.** *Assume $c_{\max} := \sup_{x \in \mathcal{X}} c(x) < \infty$. The cost-adjusted information gain is*

$$\Gamma_T(c) := \int_0^{c_{\max}} \gamma_T(\{x \in \mathcal{X} : c(x) \ge v\})\, dv.$$

The cost-adjusted information gain integrates the cost with respect to the measure induced by the information gain. Intuitively, this quantity reweights information gain by observation cost through superlevel sets of c. Regions that are both information-rich and expensive contribute more strongly to $\Gamma_T(c)$. A trivial bound is $\Gamma_T(c) \le c_{\max}\gamma_T$. For finite independent arms the cost-adjusted information gain is controlled by the maximum information gain multiplied by the sum of the costs.

Corollary 2. *Assume $\mathcal{X} = [K]$ and the kernel is the independent (delta) kernel. Then*

$$\Gamma_T(c) \le \frac{1}{2} \log\left(1 + \frac{T}{\sigma^2}\right) \sum_{i=1}^{K} c(i).$$

Computable upper bounds. The algorithms below do not require the exact values of γ_T or $\Gamma_T(c)$. It suffices to use deterministic upper bounds $\bar{\gamma}_T \ge \gamma_T$ and $\bar{\Gamma}_T(c) \ge \Gamma_T(c)$. When $c_{\min} > 0$, we take these bounds to be consistent in the sense that $\bar{\Gamma}_T(c) \ge c_{\min}\bar{\gamma}_T$, which can always be enforced by increasing $\bar{\Gamma}_T(c)$. For finite or discretized action sets, these bounds can be computed from the kernel matrix and costs: sorting the distinct cost values reduces $\bar{\Gamma}_T(c)$ to a finite sum of restricted log-determinant information-gain bounds over cost superlevel sets. For linear kernels, one can use restricted linear-bandit information-gain bounds depending on the feature radius and span dimension of each superlevel set. For RBF kernels, standard information-gain bounds based on covering dimension or intrinsic dimension can be applied on each superlevel set and then integrated over the cost levels [Srinivas et al., 2009, Chowdhury and Gopalan, 2017].

For instance-dependent bounds, we additionally encode the optimality gaps and labeling costs into a single difficulty profile.

Definition 2 (Cost-adjusted gaps)**.** *Assume a unique optimizer x^*, and define the optimality gap $\Delta(x) := f(x^*) -$*

$f(x)$, *and* $\Delta_{\min} \coloneqq \min_{x \neq x^*} \Delta(x) > 0$. *Define* $\Delta_c : \mathcal{X} \to \mathbb{R}_+$ *by*

$$\Delta_c(x) \coloneqq \begin{cases} \frac{\Delta(x)+c(x)}{\Delta(x)^2}, & x \neq x^*, \\ \frac{c(x^*)}{\Delta_{\min}^2}, & x = x^*. \end{cases}$$

The quantity $\Delta_c(x)$ is large when an arm is hard to distinguish (small gap) and/or expensive to observe (large cost). The term at x^* captures the cost of confirming the optimal arm. If all costs are equal to zero we recover the standard gaps.

Definition 3 (Cost- and gap-adjusted information gain)**.** *Let* $\Delta_c^{\max} \coloneqq \max_{x \in \mathcal{X}} \Delta_c(x)$. *Define*

$$\Gamma_T^{\text{gap}}(\Delta_c) \coloneqq \int_0^{\Delta_c^{\max}} \gamma_T(\{x \in \mathcal{X} : \Delta_c(x) \geq u\})\, du.$$

This is the instance-dependent counterpart of $\Gamma_T(c)$: it emphasizes regions of the action space that are statistically hard and expensive to explore. In the finite independent-arm setting, it reduces (up to logarithmic factors) to a weighted sum of the quantities $\Delta_c(x)$, which yields the gap-dependent rates proved later.

Corollary 3. *Assume that* $\mathcal{X} = [K]$ *and the kernel is the independent (delta) kernel (so* $k(i,j) = \mathbf{1}\{i = j\}$ *and* $\kappa = 1$*). We then have*

$$\Gamma_T^{\text{gap}}(\Delta_c) \leq \frac{1}{2} \log\Big(1 + \frac{T}{\sigma^2}\Big) \sum_{x \in \mathcal{X}} \Delta_c(x).$$

Note that when $c = 0$, the cost-adjusted gaps reduce to the standard gaps, and we recover the usual gap-dependent quantities for MAB settings.

4 THEORETICAL GUARANTEES

We now introduce `C3-GP` and `GP-C-LUCB` and provide upper and lower bounds for their regret based on the cost-adjusted information gain measures introduced in Section 3. Proofs are given in the Appendix.

Algorithm 1 `C3-GP`

Require: T, $c : \mathcal{X} \to \mathbb{R}_+$, β_T, upper bounds $\bar{\gamma}_T \geq \gamma_T$, $\bar{\Gamma}_T(c) \geq \Gamma_T(c)$, κ^2

Set $m_{\max} \coloneqq \lceil T^{4/7} \rceil$, $m \coloneqq 0$
Burn-in phase
for $t = 1, 2, \ldots, m_{\max}$ **do**
  Compute $\mu_{t-1}(\cdot)$ and $\sigma_{t-1}(\cdot)$
  Let $\sigma_{\max} \coloneqq \max_{x \in \mathcal{X}} \sigma_{t-1}(x)$
  if $\sigma_{\max} \bar{\gamma}_T \leq \bar{\Gamma}_T(c)$ **then**
    break
  end if
  Choose $x_t \in \arg\max_{x \in \mathcal{X}} \sigma_{t-1}(x)$
  Request an observation at x_t and set $m \coloneqq t$
end for
Let $\sigma_{\max} \coloneqq \max_{x \in \mathcal{X}} \sigma_m(x)$
Set $\lambda \coloneqq 2\left(\beta_T(\sigma^2 + \kappa^2)\big(\sigma_{\max}\bar{\gamma}_T + \bar{\Gamma}_T(c)\big)/T\right)^{1/3}$
for $t = m+1, m+2, \ldots, T$ **do**
  Compute $\mu_{t-1}(\cdot)$ and $\sigma_{t-1}(\cdot)$
  Define $u_{t-1}(\cdot) \coloneqq \mu_{t-1}(\cdot) + \sqrt{\beta_T}\sigma_{t-1}(\cdot)$
  Define $\ell_{t-1}(\cdot) \coloneqq \mu_{t-1}(\cdot) - \sqrt{\beta_T}\sigma_{t-1}(\cdot)$
  Set $x_t^\ell \in \arg\max_{x \in \mathcal{X}} \ell_{t-1}(x)$
  Set $x_t^u \in \arg\max_{x \in \mathcal{X}} u_{t-1}(x)$
  Let $M_{t-1} \coloneqq \max_{y \in \mathcal{X}} u_{t-1}(y)$
  $\hat{\Delta}_{t-1} \coloneqq u_{t-1} - \ell_{t-1}(x_t^\ell)$
  $D_{t-1} \coloneqq M_{t-1} - \ell_{t-1} + c$
  $\tilde{D}_{t-1} \coloneqq 2\sqrt{\beta_T}(\sigma_{t-1} + \sigma_{t-1}(x_t^u)) + c$
  $\tilde{s}_{t-1} \coloneqq \frac{\sqrt{\tilde{D}_{t-1}}}{\sqrt{D_{t-1}}}\hat{\Delta}_{t-1}$
  Set $\bar{x}_t \in \arg\max_{x \neq x_t^\ell} \tilde{s}_{t-1}(x)$
  Define $s(x) \coloneqq \frac{\sqrt{\tilde{D}_{t-1}(x)}}{\sqrt{D_{t-1}(x)}}\left(u_{t-1}(x) - \ell_{t-1}(x)\right)$
  Choose $x_t \in \arg\max_{x \in \{\bar{x}_t, x_t^\ell\}} s(x)$
  if $\max_{x \neq x_t^\ell} \tilde{s}_{t-1}(x) \geq \lambda$ **then**
    Request an observation at x_t
  else
    Stop requesting observations, play x_t^ℓ forever
  end if
end for

4.1 COST-CONSCIOUS CONFIDENCE GP

Based on the two-phase characterization of the BwCRO problem given by Lemma 1 we develop `C3-GP`, a cost-aware querying strategy that (i) selects query points using GP confidence bounds *normalized by observation cost* and (ii) *stops requesting labels* once, under the current confidence intervals, no alternative action can plausibly improve performance enough to justify its labeling cost.

`C3-GP` trades off *potential improvement* against the *price of information*. It maintains GP confidence bounds $u_{t-1}(x) = \mu_{t-1}(x) + \sqrt{\beta_T}\sigma_{t-1}(x)$ and $\ell_{t-1}(x) = \mu_{t-1}(x) - \sqrt{\beta_T}\sigma_{t-1}(x)$. The *incumbent* $x_t^\ell \in \arg\max_x \ell_{t-1}(x)$ is the action with the best guaranteed reward. After an initial burn-in phase that reduces posterior uncertainty by querying

(and labeling) points of maximal variance, the algorithm seeks a *challenger* that could plausibly outperform the incumbent. For each x, define the optimistic advantage over the incumbent as $\hat{\Delta}_{t-1}(x) \coloneqq u_{t-1}(x) - \ell_{t-1}(x_t^\ell)$. To account for costs, C3-GP downweights expensive (or risky) actions through the cost-aware score

$$\tilde{s}_{t-1}(x) \coloneqq \frac{\sqrt{\tilde{D}_{t-1}(x)}}{\sqrt{D_{t-1}(x)}}\,\hat{\Delta}_{t-1}(x),$$

where $D_{t-1}(x) \coloneqq M_{t-1} - \ell_{t-1}(x) + c(x)$, $M_{t-1} \coloneqq \max_y u_{t-1}(y)$, and $\tilde{D}_{t-1}(x) \coloneqq 2\sqrt{\beta_T}(\sigma_{t-1}(x) + \sigma_{t-1}(x_t^u)) + c(x)$.

It is instructive to interpret the cost-aware score $\tilde{s}_{t-1}$ in three steps. Define the optimistic leader $x_t^u \coloneqq \arg\max_{x\in\mathcal{X}} u_{t-1}(x)$. Since $\hat{\Delta}_{t-1}(x) = u_{t-1}(x) - \ell_{t-1}(x_t^\ell)$, the quantity $\hat{\Delta}_{t-1}(x)$ is maximized at x_t^u. Thus, ignoring the multiplicative factor $(\tilde{D}_{t-1}(x)/D_{t-1}(x))^{1/2}$, the action x_t^u is the natural challenger to the incumbent x_t^ℓ. The multiplicative factor $(\tilde{D}_{t-1}(x)/D_{t-1}(x))^{1/2}$ then reweights this optimistic advantage using uncertainty, cost, and potential downside. Writing $D_{t-1}(x) = \big(M_{t-1} - \ell_{t-1}(x)\big) + c(x)$, and $\tilde{D}_{t-1}(x) = a_{t-1}(x) + c(x)$, with $a_{t-1}(x) \coloneqq 2\sqrt{\beta_T}(\sigma_{t-1}(x) + \sigma_{t-1}(x_t^u))$, we see that a smaller lower confidence bound $\ell_{t-1}(x)$ (equivalently, a larger gap $M_{t-1} - \ell_{t-1}(x)$) increases $D_{t-1}(x)$ while leaving $\hat{\Delta}_{t-1}(x)$ unchanged, and therefore reduces $\tilde{s}_{t-1}(x)$. In this sense, actions that are already clearly implausible receive a lower score. For any action that is competitive for the score maximization (in particular, any argmax candidate), cost is always a penalty. Indeed, on this relevant set we have $\tilde{D}_{t-1}(x) \geq D_{t-1}(x)$, equivalently $a_{t-1}(x) \geq M_{t-1} - \ell_{t-1}(x)$. Then

$$\frac{\tilde{D}_{t-1}(x)}{D_{t-1}(x)} = \frac{a_{t-1}(x) + c(x)}{M_{t-1} - \ell_{t-1}(x) + c(x)}$$

is nonincreasing in $c(x)$ (strictly decreasing unless equality holds), so increasing the observation cost decreases the multiplicative factor and hence decreases $\tilde{s}_{t-1}(x)$. Thus, among actions that can win the score maximization, higher cost always makes an action less attractive. Moreover, an action $x \neq x_t^u$ can beat the optimistic leader in the score $\tilde{s}_{t-1}$ only if the multiplicative factor compensates for its smaller optimistic advantage $\hat{\Delta}_{t-1}(x)$. A necessary condition for this is $\tilde{D}_{t-1}(x) \geq D_{t-1}(x)$, which is equivalent to

$$\mu_{t-1}(x_t^u) - \mu_{t-1}(x) \leq \sqrt{\beta_T}\big(\sigma_{t-1}(x_t^u) + \sigma_{t-1}(x)\big).$$

That is, only actions that are statistically competitive with the optimistic leader (within confidence width) can overtake it in the cost-aware score. Among such plausible challengers, the score favors lower-cost actions and actions with smaller potential downside. The effect of cost is adaptive. When $M_{t-1} - \ell_{t-1}(x)$ is large (many actions remain plausible), differences in $c(x)$ have a stronger impact on the score, encouraging cost-efficient exploration. When $M_{t-1} - \ell_{t-1}(x)$ is small (the contender set is narrow), the score becomes less sensitive to cost, allowing the algorithm to focus on resolving the remaining uncertainty near the optimum. Finally, C3-GP stops requesting labels once no challenger has enough cost-aware upside, i.e., $\max_{x\neq x_t^\ell} \tilde{s}_{t-1}(x) < \lambda$, and then plays x_t^ℓ without further observations. In our implementation, β_T is set using the standard finite-action confidence choice unless specified otherwise, $m_{\max} = \lceil T^{4/7} \rceil$ and λ are fixed by the theorem, and $\bar{\gamma}_T, \bar{\Gamma}_T(c)$ are computed from the kernel/cost structure or from deterministic finite-design upper bounds. Thus these quantities are inputs to the guarantee, not tuned per instance.

Algorithm 2 GP-C-LUCB

Require: horizon T, costs $c : \mathcal{X} \to \mathbb{R}_+$, variable v

for $t = 1, 2, \ldots, T$ **do**
  Compute μ_{t-1} and σ_{t-1}
  Define $u_{t-1}(x) \coloneqq \mu_{t-1}(x) + \sqrt{\beta_T}\sigma_{t-1}(x)$
  Define $\ell_{t-1}(x) \coloneqq \mu_{t-1}(x) - \sqrt{\beta_T}\sigma_{t-1}(x)$
  Set $x_t^\ell \in \arg\max_{x\in\mathcal{X}} \ell_{t-1}(x)$
  Set $\bar{x}_t \in \arg\max_{x\in\mathcal{X}\setminus\{x_t^\ell\}} u_{t-1}(x)$
  if $u_{t-1}(\bar{x}_t) - \ell_{t-1}(x_t^\ell) \leq v$ **then**
    Stop requesting observations, play x_t^ℓ forever
  else
    $x_t \in \operatorname{argmax}_{x\in\{x_t^\ell, \bar{x}_t\}} \sigma_{t-1}(x)$
    Request an observation at x_t
  end if
end for

Theorem 4. *Assume $k(x,x) \leq \kappa^2$ for all $x \in \mathcal{X}$ and the confidence event $\mathcal{E}$ holds. Assume also $|f(x)| \leq B$ for all $x \in \mathcal{X}$ and $c_{\min} > 0$. Run Algorithm 1 with deterministic upper bounds $\bar{\gamma}_T \geq \gamma_T$ and $\bar{\Gamma}_T(c) \geq \Gamma_T(c)$ satisfying $\bar{\Gamma}_T(c) \geq c_{\min}\bar{\gamma}_T$. Then on $\mathcal{E}$,*

$$R_T \in \mathcal{O}\left(T^{2/3}\Big(\beta_T(\sigma^2 + \kappa^2)\bar{\Gamma}_T(c)\Big)^{1/3}\right).$$

Remark 1 (On the minimum-cost assumption)**.** *The assumption $c_{\min} > 0$ is a convenient sufficient condition, not a necessary one. The same proof goes through whenever the residual uncertainty left by the burn-in phase can be absorbed into the cost-adjusted information term; for example, it is enough that*

$$\sigma\bar{\gamma}_T\sqrt{\frac{\bar{\gamma}_T}{T^{4/7}}} \lesssim \bar{\Gamma}_T(c)$$

up to logarithmic factors. This condition is satisfied in many nondegenerate GP-bandit regimes where information gain grows sub-polynomially and the cost-adjusted information gain is not asymptotically negligible relative to γ_T.

Proof Sketch. Let τ be the stopping time of the exploration phase. Throughout we have $\tilde{D}_t(x^*)/D_t(x^*) \geq 1$ and there-

fore $f(x^*) - f(x_\tau^\ell) \le u_\tau(x^*) - \ell_\tau(x_\tau^\ell) = \hat{\Delta}_\tau(x^*) \le \tilde{s}_\tau(x^*) \le \lambda$. This implies that the regret of exploitation can be bounded by $T\lambda$. For a labeled round $t \le \tau$, we get

$$r(x_t) + c(x_t) \le D_{t-1}(x_t) \le \frac{4\beta_T}{\lambda^2} \tilde{D}_{t-1}(x_t)\sigma^2_{t-1}(x_t).$$

Using the definition of the cost-adjusted information gain we can bound $\sum_{t\le\tau} r(x_t)+c(x_t) \le 8\beta_T(\sigma^2+\kappa^2)(\bar{\Gamma}_T(c)+\sigma_{\max}\bar{\gamma}_T)/\lambda^2$, where $\sigma_{\max}$ is a bound on the maximum posterior variance after burn-in. The burn-in stopping criterion, or the max-variance log-det bound if burn-in reaches $m_{\max}$, controls $\sigma_{\max}\bar{\gamma}_T$ by $\bar{\Gamma}_T(c)$ up to lower-order terms. The final bound is derived by balancing the burn-in, exploration, and exploitation terms through the theorem's choice of λ.

Example 1. *We show with a simple example that the cost-adjusted information gain can be small even when the maximum cost $c_{\max}$ and the average cost $\bar{c}$ are large. The high-level idea is that when the high-complexity regions of the action space are cheap and low-complexity regions are expensive then the cost-adjusted information gain remains moderate compared to the cost.*

Fix a horizon $T \in \mathbb{N}$ and let $\mathcal{X} \subseteq [0,1]$ be a finite action set containing two disjoint regions A and B with $\mathcal{X} = A \cup B$. For concreteness one may take a uniform grid $\mathcal{X} = \{i/K : i \in [K]\}$ and define $A \coloneqq \mathcal{X} \cap [0, 0.2]$ and $B \coloneqq \mathcal{X} \setminus A$. Let $\ell > 0$ and define the kernel

$$k(x,x') \coloneqq s(x)s(x') \exp\Big(-\frac{(x-x')^2}{2\ell^2}\Big),$$

where $s(x) \coloneqq \delta \mathbb{1}_A(x) + \mathbb{1}_B(x)$ and $0 < \delta < 1$. Let the observation noise proxy be $\sigma^2 > 0$. Define the costs by $c(x) \coloneqq M\mathbb{1}_A(x) + \epsilon\mathbb{1}_B(x)$ where $0 < \varepsilon < M$ and M can be arbitrarily large, so $c_{\max} = M$. Then, if $\varepsilon = 1/T$ and $\delta^2 = \sigma^2/(TM)$, the cost-adjusted information gain satisfies

$$\Gamma_T(c) \le \frac{1}{2}\log\Big(1+\sigma^{-2}\Big) + \frac{1}{2},$$

which is bounded by a constant independent of M. This shows that $\Gamma_T(c)$ can remain small even when $c_{\max}$ and $\bar{c} \ge M/K$ are large. A proof is given in Appendix B.6 and we investigate this example empirically in Figure 3 (right).

Comparison to Tucker et al. [2023]. The algorithms of Tucker et al. [2023] are developed for finite multi-armed bandits with independent arms and a constant observation cost. Consequently, they are not adaptive to heterogeneous cost functions: in particular, their sampling rules do not preferentially query cheaper arms over more expensive ones. As a result, their regret guarantees do not benefit from correlation structure and scale with the maximum cost $c_{\max}$ (their minimax bounds are of order $\mathcal{O}(c_{\max}^{1/3}|\mathcal{X}|^{2/3}T^{2/3})$). This dependence becomes problematic when $c_{\max} \gg c_{\min}$. By contrast, our analysis captures how correlation and heterogeneous costs can jointly reduce the effective problem complexity through $\Gamma_T(c)$. In particular, when expensive arms are well correlated with cheaper neighboring arms, `C3-GP` can avoid costly labels by exploiting this correlation structure and inferring their values from cheaper observations nearby. This mechanism is illustrated empirically in the cost-scaling experiment (see Figure 3), where the regret of `C3-GP` remains nearly constant as $c_{\max}$ (respectively, $\bar{c}$) increases, while the baselines' regret increase with $c_{\max}$. In the constant cost setting, we improve over the bounds of Tucker et al. [2023] with respect to the arm complexity. For finite independent arms, we achieve bounds that scale with $|\mathcal{X}|^{1/3}$ compared to their $|\mathcal{X}|^{2/3}$. For correlated arms, our bounds scale with the maximum information gain (through $\Gamma_T(c)$) which, for commonly used kernels, implies no dependence on the number of arms as long as the action domain is bounded.

Theorem 5 shows that Algorithm 1 is asymptotically optimal in the finite independent-arm setting, by matching our upper bound with a corresponding lower bound. In particular, it identifies $\Gamma_T(c)$ as the fundamental minimax complexity measure in the independent-arm regime. The same conclusion extends immediately to block-diagonal kernels, where each block corresponds to a cost-level set. Whether $\Gamma_T(c)$ (and thus the bound in Theorem 4) remains optimal for general kernels is an open question. Proving this would require controlling how much information observations in low-cost regions can reveal about rewards in high-cost regions. We conjecture that under suitable smoothness assumptions on the cost function c, such cross-region information transfer is limited: nearby actions are informative about one another, and smoothness of c implies that nearby actions have similar costs.

Theorem 5 (Minimax lower bound)**.** *Fix a finite action set $\mathcal{X}$ with $|\mathcal{X}| = K \ge 2$ and a cost function $c : \mathcal{X} \to \mathbb{R}_{\ge 0}$ with $0 < c_{\max} < \infty$. Let $\mathbb{S}$ be the class of all K-armed Gaussian bandit instances with rewards $\mathcal{N}(\mu(x), 1)$ and means $\mu(x) \in [0,1]$ for all $x \in \mathcal{X}$, together with the fixed cost function c. Then there exists a universal constant $C > 0$ such that for any algorithm $\mathcal{A}$,*

$$\sup_{\nu\in\mathbb{S}} \mathbb{E}_\nu\big[R_T(\mathcal{A})\big] \ge C \min\left\{\frac{\Gamma_T^{1/3}(c)}{\log^{1/3}(1+T)} T^{2/3},\ T\right\}.$$

4.2 RESULTS FOR `GP-C-LUCB`.

We now turn to `GP-C-LUCB` (inspired by Kalyanakrishnan et al. [2012]) and derive an instance-dependent (gap-dependent) regret bound. In contrast to `C3-GP`, which is designed for minimax guarantees, `GP-C-LUCB` is tailored to exploit favorable problem instances through the gap structure. The algorithm follows the two-phase principle from Lemma 1: it requests labels only while there remains a plausible challenger to the current incumbent, and once the confidence intervals are sufficiently separated,

it stops labeling and commits to the empirically best action. More concretely, at each labeled round the algorithm selects an *incumbent* $x_t^\ell \in \arg\max_x \ell_{t-1}(x)$ and a *challenger* $\bar{x}_t \in \arg\max_{x \neq x_t^\ell} u_{t-1}(x)$, and continues labeling only if the confidence gap $u_{t-1}(\bar{x}_t) - \ell_{t-1}(x_t^\ell)$ exceeds a threshold v. This is the natural GP analogue of LUCB-style elimination, adapted to the costly-observation setting. This design is useful for deriving gap-dependent guarantees, but it is less aggressively cost-aware than `C3-GP`: it only compares the incumbent with the most optimistic challenger and then samples the more uncertain of these two actions. Consequently, when a high-cost action remains plausible, `GP-C-LUCB` may continue sampling that action or its direct challenger, whereas `C3-GP` can often reduce uncertainty indirectly through cheaper correlated proxy actions.

The resulting regret bound is governed by the cost- and gap-adjusted information gain $\Gamma_T^{\text{gap}}(\Delta_c)$ introduced in Definition 3. This quantity emphasizes arms that are both statistically hard to distinguish (small gaps) and expensive to label (large costs), which are precisely the arms that dominate instance-dependent regret.

Theorem 6. *Assume $k(x,x) \leq \kappa^2$ for all $x \in \mathcal{X}$ and the confidence event $\mathcal{E}$ holds. Run Algorithm 2 with $v = (8\Gamma_T(c)\beta_T\sigma^2/T)^{1/3}$, and assume $\Delta_{\min} > 0$ and $T > 8\Gamma_T(c)\beta_T\sigma^2/(\Delta_{\min}^3)$. Then on $\mathcal{E}$,*

$$R_T \leq 32\beta_T(\sigma^2 + \kappa^2)\Gamma_T^{\text{gap}}(\Delta_c).$$

Theorem 6 shows that `GP-C-LUCB` adapts to instance difficulty through $\Gamma_T^{\text{gap}}(\Delta_c)$ rather than a worst-case quantity such as $c_{\max}\gamma_T$. In particular, arms that are clearly suboptimal contribute little, while the dominant terms come from arms with small gaps and/or high observation costs. This matches the intuition that costly observations matter most when they are needed to resolve difficult comparisons. In the finite independent-arm setting, $\Gamma_T^{\text{gap}}(\Delta_c)$ reduces (up to logarithmic factors) to a weighted sum of the cost-adjusted gap terms $\Delta_c(x)$, yielding the familiar logarithmic dependence on T and making the connection to classical instance-dependent bandit rates explicit. We next show that this dependence is unimprovable (up to constants and logarithmic factors) in the finite independent-arm setting.

Theorem 7 (Gap-dependent lower bound)**.** *Consider the set $\mathbb{S}$ of BwCRO instances with finite, independent arms and Gaussian rewards with variance σ^2. For any consistent policy $\mathcal{A}$ (that is, for every instance $\nu \in \mathbb{S}$ and every $a > 0$, $\mathbb{E}_\nu[R_T] = o(T^a)$), the expected regret (including costs) satisfies, for every instance $\nu \in \mathbb{S}$,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[R_T]}{\log(T)} \geq 2\sigma^2 \sum_{x\in\mathcal{X}} \Delta_c(x).$$

Theorem 7 identifies $\sum_x \Delta_c(x)$ as the fundamental instance complexity in the independent-arm regime. Combined with the upper bound for `GP-C-LUCB` (and its finite-arm specialization in Corollary 3), this shows that our cost- and gap-adjusted complexity measure captures the correct asymptotic scaling of the problem.

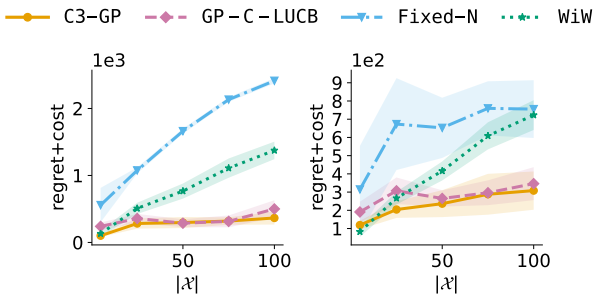

Figure 1: **Scaling with the number of arms.** Cumulative regret with observation cost versus number of arms $|\mathcal{X}|$ under constant, unit cost (left) and heterogeneous log-normal cost (right).

5 EXPERIMENTS

We evaluate our cost-aware methods against two baselines from Tucker et al. [2023]: `WiW` and `Fixed-N`. Both baselines were originally proposed for the constant-cost setting with finitely many independent arms. To apply them in our more general setting with heterogeneous costs and correlated rewards, we use conservative adaptations. In particular, we evaluate them using the maximum observation cost across arms. For `Fixed-N`, we additionally replace its base bandit subroutine with `GP-UCB` to enable operation in correlated-action (i.e. contextual arm) settings. Additional experiments and diagnostics are reported in Appendix A; these include `C3-GP` without the burn-in phase and a transfer baseline, `CostAware GP-UCB`, which observes every round and selects actions by maximizing the direct cost-normalized score $(\mu_{t-1}(x) + \sqrt{\beta_T}\sigma_{t-1}(x))/(c(x) + \epsilon)$.

Experimental setup. Unless otherwise specified, we use an arm set $\mathcal{X} = [50]$, horizon $T = 2000$, noise level $\sigma = 1$, and confidence parameter $\delta = 0.5$. Each arm $i \in \mathcal{X}$ has an associated context $z_i \in \mathbb{R}^2$, sampled once per run uniformly at random from a square of diameter 5 in $\mathbb{R}^2$. We then sample a reward function f from a Gaussian process prior with an RBF kernel (unit lengthscale and unit variance) over the arm contexts, and the reward obtained by pulling arm i is $y_t = f(z_i) + \varepsilon_t$. The observation cost is either *constant*, $c(i) = 1$, or *heterogeneous*, where the costs are sampled i.i.d. from a log-normal distribution with *actual* mean 0.5 and *actual* standard deviation 1.0. In all plots, shaded regions show ± 1 standard error around the mean over 30 independent runs. Our implementation uses Python with GPyTorch and NumPy [Gardner et al., 2018, Paszke et al., 2019, Harris et al., 2020] and is included in the supplement.

Synthetic experiments. We first study how cumulative regret (including observation costs) scales with the number of arms $|\mathcal{X}|$. Results for constant unit costs are shown in Figure 1 (left), and results for heterogeneous log-normal costs are shown in Figure 1 (right). Since `WiW` does not exploit context information (reward correlation across arms), its regret increases rapidly with $|\mathcal{X}|$, as expected. `Fixed-N` also performs poorly in this experiment, despite being evaluated in the constant-cost regime for which it was originally designed (and despite our context-aware GP-UCB adaptation of its base subroutine). We hypothesize that this is due to its fixed exploration horizon, which can be overly conservative in many instances. By contrast, both `C3-GP` and `GP-C-LUCB` exhibit only a mild increase in average regret as $|\mathcal{X}|$ grows. This is consistent with our theory: because each arm has an associated context in a fixed bounded region of $\mathbb{R}^2$, increasing $|\mathcal{X}|$ makes the action set denser and induces stronger reward correlations under the RBF kernel. Consequently, the effective problem complexity (as captured by $\Gamma_T(c)$ and $\Gamma_T^{\text{gap}}(\Delta_c)$) grows much more slowly than the number of arms.

In the second experiment, we examine how cumulative regret scales with the horizon T. We report results for constant unit costs in Figure 2 (left) and for i.i.d. log-normal costs in Figure 2 (right). All methods have theoretical upper bounds of order $T^{2/3}$, and empirically `WiW`, `GP-C-LUCB`, and `C3-GP` exhibit similar scaling trends. However, our methods achieve uniformly lower regret across the tested horizons. By contrast, `Fixed-N` scales less favorably with T in our experiments. We again attribute this to its non-adaptive exploration horizon, which can lead to over-exploration.

In the third experiment, we study the dependence of cumulative regret on the maximum observation cost $c_{\max}$. We consider two settings, shown in Figure 3 (left and right). In the left panel, all arms have unit observation cost except for a single high-cost arm. In the right panel, we use the correlated-cost construction from Example 1. This experiment most clearly illustrates the benefit of cost-aware label selection in `C3-GP`. As $c_{\max}$ (and correspondingly the average cost $\bar{c}$) increases, the regret of `C3-GP` remains nearly unchanged, whereas the regret of the competing methods increases with $c_{\max}$. The baselines are cost-agnostic in their label-selection rules, while `C3-GP` explicitly incorporates observation cost into its acquisition criterion. As a result, when an arm is substantially more expensive to label than nearby arms, `C3-GP` can often reduce costly queries by exploiting correlation and inferring its value from cheaper neighboring observations.

Real-world data. To evaluate our algorithms on real data, we construct a bandit problem from the Combined Cycle Power Plant dataset [Tfekci and Kaya, 2014]. We first reduce the four-dimensional feature space (T, AP, RH, V) to two dimensions using Principal Component Analysis (PCA),

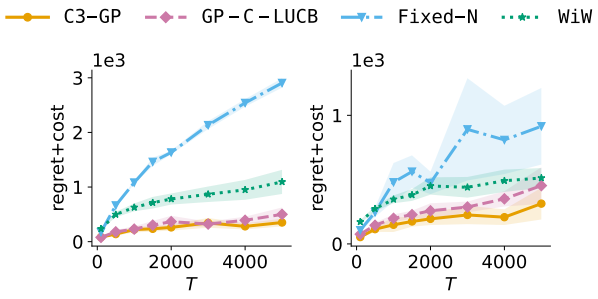

Figure 2: **Scaling with the horizon T.** Cumulative regret with observation cost versus horizon T under constant, unit cost (left) and heterogeneous log-normal cost (right).

retaining the first two principal components. We then discretize this two-dimensional space into 16 equal-sized bins, and treat each bin as one action. For each action (bin), the reward distribution is given by the net hourly electrical energy output (EP) values of the data points in that bin, so pulling an action yields a stochastic reward sampled from the corresponding bin. The context associated with each action is the average of the two retained principal components over the data points in that bin. Observation costs are sampled independently for each action from $\text{Unif}[1, 5]$ and redrawn for each run. The results in Figure 4 show that both of our algorithms outperform the baselines across all tested horizons. We also observe that `C3-GP` outperforms `GP-C-LUCB` in this setting. We attribute this to the cost-aware action-selection mechanism in `C3-GP`, which can improve estimates of costly arms by querying additional labels from nearby, less expensive arms—a capability that the baselines do not exploit.

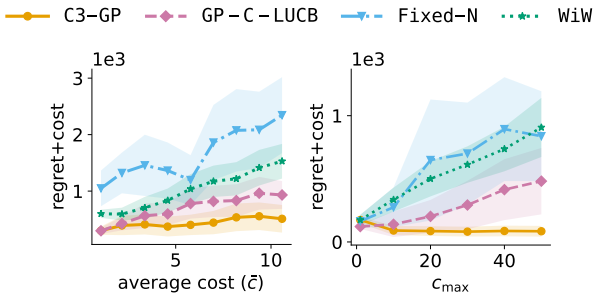

Figure 3: **Scaling with observation cost.** We consider two settings. **Left:** a reward function sampled from a GP with an RBF kernel, with unit observation costs for all arms except one high-cost arm. **Right:** the correlated-cost setting from Example 1.

6 RELATED WORK

BwCRO fits within the partial monitoring framework, where an agent learns from imperfect feedback instead of direct

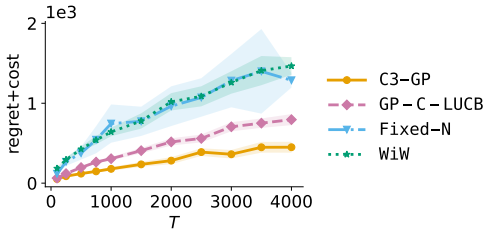

Figure 4: **Scaling with the horizon T.** Cumulative regret with observation cost versus horizon T for the Combined Cycle Power Plant dataset.

reward observations [Bartók et al., 2014, Lattimore and Szepesvári, 2020, 2019]. In many such problems (more specifically, the partial monitoring problems classified as "hard"), the need to take suboptimal actions for information gain leads to a higher regret rate of $\tilde{\Omega}(T^{2/3})$. Specifically, Seldin et al. [2014] investigate a setting where the agent in each step can decide to pay to observe noisy rewards of *any* actions (not just the action that was played). In contrast to partial monitoring, BwCRO has a more specific structure where the cost between the observed and unobserved actions is given by a cost function c, and we are interested in bounds depending on c.

It is also useful to contrast BwCRO with budgeted bandits and Bandits with Knapsacks (BwK). In budgeted bandits, a fixed resource budget limits the total arm pulls, and the goal is to maximize reward before depletion [Tran-Thanh, 2012, Ding et al., 2013, Xia et al., 2016]. BwCRO instead integrates observation cost directly into the regret: there is no hard budget cap, but each costly query reduces performance, so the learner must trade off information gain against cost. Similarly, BwK handles multiple resource constraints via knapsack budgets [Agrawal and Devanur, 2014, Badanidiyuru et al., 2018, Liu et al., 2022], whereas BwCRO can be viewed as a single-budget, infinite-knapsack variant whose cost is priced into the regret objective, requiring continuous query decisions rather than budget exhaustion.

We further relate to work on costly feature observations (a.k.a. active feature acquisition) [Zolghadr et al., 2013, Janisch et al., 2020, Atan et al., 2021, Ghoorchian et al., 2024]. The agent purchases partial or full access to the current feature vector x_t, typically paying per coordinate and sometimes operating under a per-round budget. Acquisition is often adaptive, interleaving feature queries with a stopping rule and a final action in order to trade off predictive/control accuracy against measurement cost. In that literature, the features (not the outcomes) are a priori unobserved, often assuming outcomes are revealed for free once an action is taken, whereas in our setting x_t is costless and the agent pays to observe outcomes.

There is also literature on costly probes [Elumar et al., 2024], in which, before pulling an arm, the decision-maker may pay a cost $c \geq 0$ to probe one of the K arms and observe its reward. Multi-fidelity bandits also model a cost–information trade-off, but there the learner observes a lower-cost approximation of the reward (typically biased/noisier yet informative) and can choose among several fidelities [Kandasamy et al., 2016, 2019, Wang et al., 2023]. In contrast, in our BwCRO formulation, declining to pay $c(x_t)$ yields missing feedback rather than a lower-fidelity signal, making observation acquisition—not fidelity choice—the central challenge.

7 CONCLUSION AND FUTURE WORK

We studied bandits with costly reward observations under action-dependent costs and correlated rewards. We showed that it is without loss of generality to consider two-phase policies: a label-requesting exploration phase followed by a no-label commitment phase. We then introduced two cost-aware complexity measures, $\Gamma_T(c)$ and $\Gamma_T^{\mathrm{gap}}(\Delta_c)$, and used them to analyze our GP-based algorithms `C3-GP` and `GP-C-LUCB`. In the finite independent-arm setting, we proved matching upper and lower bounds (up to constants/log factors), and our experiments confirm strong empirical performance, especially for `C3-GP` in heterogeneous-cost settings. An important direction for future work is to determine whether $\Gamma_T(c)$ is also minimax-optimal for general kernels, beyond the independent-arm case.

References

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

Onur Atan, Saeed Ghoorchian, Setareh Maghsudi, and Mihaela van der Schaar. Data-driven online recommender systems with costly information acquisition. *IEEE Transactions on Services Computing*, 16(1):235–245, 2021.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 232–238, 2013.

Eray Can Elumar, Cem Tekin, and Osman Yağan. Multi-armed bandits with costly probes. *IEEE Transactions on Information Theory*, 2024.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*, 2008.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. *Advances in Neural Information Processing Systems*, 29, 2016.

Saeed Ghoorchian, Evgenii Kortukov, and Setareh Maghsudi. Contextual multi-armed bandit with costly feature observation in non-stationary environments. *IEEE Open Journal of Signal Processing*, 5:820–830, 2024.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.

Jaromír Janisch, Tomáš Pevnỳ, and Viliam Lisỳ. Classification with costly features as a sequential decision-making problem. *Machine Learning*, 109(8):1587–1615, 2020.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

Kirthevasan Kandasamy, Gautam Dasarathy, Barnabas Poczos, and Jeff Schneider. The multi-fidelity multi-armed bandit. *Advances in neural information processing systems*, 29, 2016.

Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, and Barnabas Poczos. Multi-fidelity gaussian process bandit optimisation. *Journal of Artificial Intelligence Research*, 66:151–196, 2019.

David Krueger, Jan Leike, Owain Evans, and John Salvatier. Active reinforcement learning: Observing rewards at a cost. *arXiv preprint arXiv:2011.06709*, 2020.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Csaba Szepesvári. Cleaning up the neighborhood: A full classification for adversarial partial monitoring. In *Algorithmic Learning Theory*, pages 529–556. PMLR, 2019.

Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pages 2488–2515. PMLR, 2020.

Shang Liu, Jiashuo Jiang, and Xiaocheng Li. Non-stationary bandits with knapsacks. *Advances in Neural Information Processing Systems*, 35:16522–16532, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Sebastian Schulze and Owain Evans. Active reinforcement learning with monte-carlo tree search. *arXiv preprint arXiv:1803.04926*, 2018.

Yevgeny Seldin, Peter Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *International Conference on Machine Learning*, pages 280–287. PMLR, 2014.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Pnar Tfekci and Heysem Kaya. Combined Cycle Power Plant. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5002N.

Long Tran-Thanh. *Budget-limited multi-armed bandits*. PhD thesis, University of Southampton, 2012.

Aaron D Tucker, Caleb Biddulph, Claire Wang, and Thorsten Joachims. Bandits with costly reward observations. In *Uncertainty in Artificial Intelligence*, pages 2147–2156. PMLR, 2023.

Xuchuang Wang, Qingyun Wu, Wei Chen, and John Lui. Multi-fidelity multi-armed bandits revisited. *Advances in Neural Information Processing Systems*, 36:31570–31600, 2023.

Yingce Xia, Wenkui Ding, Xu-Dong Zhang, Nenghai Yu, and Tao Qin. Budgeted bandit problems with continuous random costs. In *Asian conference on machine learning*, pages 317–332. PMLR, 2016.

Navid Zolghadr, Gábor Bartók, Russell Greiner, András György, and Csaba Szepesvári. Online learning with costly features and labels. *Advances in neural information processing systems*, 26, 2013.

The Price of Knowledge: Optimal Algorithms for Costly Bandits (Supplementary Material)

Felix Schur[1] **Jesus Lago**[2] **Tanner Fiez**[2]

[1]ETH Zurich, Department of Mathematics
[2]Amazon

A ADDITIONAL EXPERIMENTS

Figures 5 and 6 repeat the main number-of-arms and horizon sweeps with the `CostAware GP-UCB` transfer baseline included. This heuristic observes every round and chooses the arm maximizing $(\mu_{t-1}(x) + \sqrt{\beta_T}\sigma_{t-1}(x))/(c(x) + \epsilon)$. We omit this baseline from the main figures because its larger regret compresses the scale of the other methods.

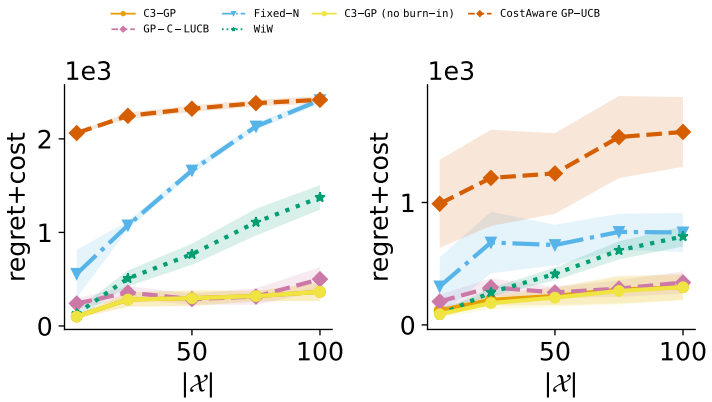

Figure 5: **Scaling with the number of arms, including `CostAware GP-UCB`.** Cumulative regret with observation cost versus number of arms $|\mathcal{X}|$ under constant, unit cost (left) and heterogeneous log-normal cost (right).

Table 1: **Runtime diagnostics.** Representative log-normal-cost experiment with $|\mathcal{X}| = 50$ and $T = 1000$. Times are average wall-clock seconds per run.

Algorithm	Avg. time
`C3-GP`	0.30s
`GP-C-LUCB`	0.13s
`Fixed-N`	1.23s
`WiW`	0.05s
`C3-GP` (no burn-in)	0.17s
`CostAware GP-UCB`	4.36s

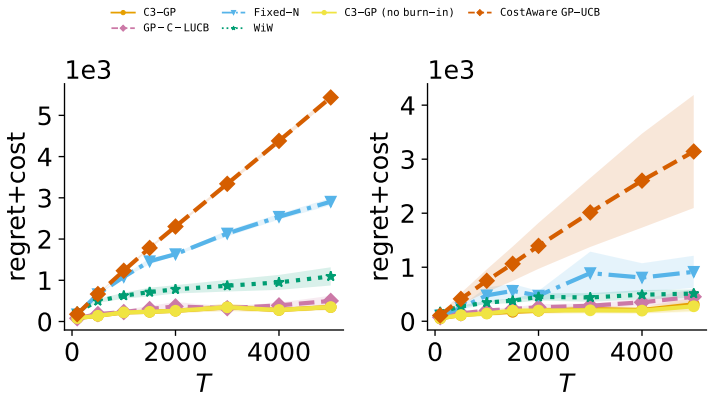

Figure 6: **Scaling with the horizon T, including `CostAware GP-UCB`.** Cumulative regret with observation cost versus horizon T under constant, unit cost (left) and heterogeneous log-normal cost (right).

B PROOFS

B.1 PROOF OF LEMMA 1

Fix $\mathcal{A} \in \tilde{\mathcal{A}}$. We construct $\mathcal{A}'$ without using the instance ν and then show the regret comparison for an arbitrary ν. Since ν is arbitrary, the result is uniform. Write the internal randomness of $\mathcal{A}$ as a seed ξ. Policy $\mathcal{A}'$ draws ξ once at time 1 and keeps it fixed.

Policy $\mathcal{A}'$ maintains a virtual copy of $\mathcal{A}$ driven by the same seed ξ. Let $\tilde{H}_k^{\text{full}}$ be the virtual full history after k virtual rounds. Virtual rounds are indexed by $k = 1, \dots, T$ and follow the original online order of $\mathcal{A}$. At each virtual round k, the virtual copy outputs

$$(\tilde{x}_k, \tilde{d}_k) = \mathcal{A}(\tilde{H}_{k-1}^{\text{full}}).$$

Policy $\mathcal{A}'$ builds a real-time labeled prefix by executing only the labeled virtual rounds, in their original order:

1. If $\tilde{d}_k = 1$, then in the next unused real round $\mathcal{A}'$ plays arm $\tilde{x}_k$, requests a label, observes $\tilde{y}_k$, and appends $(\tilde{x}_k, 1, \tilde{y}_k)$ to $\tilde{H}_k^{\text{full}}$.
2. If $\tilde{d}_k = 0$, then $\mathcal{A}'$ does not use a real round now. It appends $(\tilde{x}_k, 0, 0)$ to $\tilde{H}_k^{\text{full}}$ (no reward is observed), and records $\tilde{x}_k$ in a list of virtual unlabeled actions.

Because rewards are stationary and independent across time, and because unlabeled rounds yield no information, the virtual copy evolves exactly as $\mathcal{A}$ would evolve on any instance ν: the virtual full history $\tilde{H}_k^{\text{full}}$ has the same distribution as the actual full history of $\mathcal{A}$ after k rounds on ν. In particular, the labeled subsequence $(\tilde{x}_k, \tilde{y}_k)_{k \in L}$ has the same joint distribution as the labeled subsequence $(x_t, y_t)_{t \le T : d_t = 1}$ of $\mathcal{A}$ on ν, and the unlabeled action multiset $(\tilde{x}_k)_{k \in U}$ has the same distribution as $(x_t)_{t \le T : d_t = 0}$.

Let

$$L := \{k \in [T] \mid \tilde{d}_k = 1\}, \qquad U := \{k \in [T] \mid \tilde{d}_k = 0\},$$

and let $m := |L|$. The above construction uses exactly m real rounds and requests labels on all of them. Hence in real time $\mathcal{A}'$ requests labels only in a prefix of length m. Define the stopping time $\tau := m$.

If $m = T$, then $U = \emptyset$ and there is no suffix; the construction below is understood with $N = 0$ and no commitment is taken. Otherwise assume $m < T$. After simulating all virtual rounds $1, \dots, T$, all labeled rounds have been matched with real labeled interactions, and all remaining virtual rounds are unlabeled. Thus, before acting at real time $\tau + 1$, $\mathcal{A}'$ completes the virtual simulation and obtains the list of virtual unlabeled actions $(\tilde{x}_k)_{k \in U}$. Let $N := T - m = |U|$ be the number of unlabeled virtual rounds. Policy $\mathcal{A}'$ draws a single arm $\hat{x}$ uniformly from the multiset $(\tilde{x}_k)_{k \in U}$, i.e.,

$$\mathbb{P}(\hat{x} = x \mid (\tilde{x}_k)_{k \in U}) = \frac{1}{N} \sum_{k \in U} \mathbb{1}\{\tilde{x}_k = x\}.$$

It then plays $\hat{x}$ for all remaining real rounds $t > \tau$ with $d'_t = 0$. This makes $\mathcal{A}'$ two-phase and, after time τ, it repeats a single action.

Fix an arbitrary instance ν with mean function μ and let x^* be an optimal arm. Write $\Delta(x) = \mu(x^*) - \mu(x)$ and note $\Delta(x) \geq 0$ for all x. By the coupling, the labeled subsequences of $\mathcal{A}$ and $\mathcal{A}'$ agree in distribution, so

$$\mathbb{E}_\nu\Big[\sum_{k\in L}\big(\Delta(\tilde{x}_k) + c(\tilde{x}_k)\big)\Big] = \mathbb{E}_\nu\Big[\sum_{t:d_t=1}\big(\Delta(x_t) + c(x_t)\big)\Big].$$

For the suffix, condition on the realized unlabeled multiset $(\tilde{x}_k)_{k\in U}$. Policy $\mathcal{A}$ incurs unlabeled regret

$$\sum_{k\in U}\Delta(\tilde{x}_k).$$

Policy $\mathcal{A}'$ plays $\hat{x}$ for N rounds, so its conditional expected unlabeled regret is

$$N\mathbb{E}_\nu[\Delta(\hat{x}) \mid (\tilde{x}_k)_{k\in U}] = N \cdot \frac{1}{N}\sum_{k\in U}\Delta(\tilde{x}_k) = \sum_{k\in U}\Delta(\tilde{x}_k).$$

Thus the unlabeled part of the expected regret is identical for $\mathcal{A}$ and $\mathcal{A}'$. Adding the labeled and unlabeled contributions yields

$$\mathbb{E}_\nu[R_T(\mathcal{A}')] = \mathbb{E}_\nu[R_T(\mathcal{A})].$$

Since ν was arbitrary and $\mathcal{A}'$ was constructed without using ν, this equality (and hence the inequality) holds for all instances.

During the real labeled prefix, $\mathcal{A}'$ requests a label every round, so past d' values are deterministically equal to 1 and provide no additional information beyond the labeled pairs. All virtual d-values and unlabeled actions are generated internally from the labeled history and the seed ξ, and are never observed from the environment. The post-τ commitment is therefore a measurable function of (H_τ, ξ). Hence $\mathcal{A}'$ can be represented as a policy that observes only H_{t-1}.

B.2 PROOF OF LEMMA 9

Lemma 8 (Weighted variance via superlevel restricted information gain)**.** *Let $w : \mathcal{X} \to \mathbb{R}_+$ and $w_{\max} \coloneqq \max_{x\in\mathcal{X}} w(x)$. Assume $k(x,x) \leq \kappa^2$ for all x. For any (possibly adaptive) sequence $x_{1:T}$,*

$$\sum_{t=1}^{T} w(x_t)\sigma^2_{t-1}(x_t) \leq 2(\sigma^2+\kappa^2)\int_0^{w_{\max}} \gamma_T(\{x\in\mathcal{X} : w(x) \geq u\})\,du.$$

Proof. Using the layer-cake identity $w(x) = \int_0^{w(x)} du$,

$$\sum_{t=1}^{T} w(x_t)\sigma^2_{t-1}(x_t) = \int_0^{w_{\max}} \sum_{t:w(x_t)\geq u} \sigma^2_{t-1}(x_t)\,du.$$

Fix u and define $I(u) \coloneqq \{t : w(x_t) \geq u\}$ and $S(u) \coloneqq \{x : w(x) \geq u\}$. For $t \in I(u)$, let $\sigma^2_{t-1,I(u)}(x_t)$ denote the posterior variance at x_t computed using only the observations $\{(x_s, y_s) : s \in I(u) \cap [t-1]$. Removing observations cannot decrease posterior variances, hence

$$\sum_{t\in I(u)} \sigma^2_{t-1}(x_t) \leq \sum_{t\in I(u)} \sigma^2_{t-1,I(u)}(x_t).$$

For the subsequence indexed by $I(u)$, the log-det identity gives

$$\frac{1}{2}\log\det\Big(I + \sigma^{-2}K_{I(u)}\Big) = \frac{1}{2}\sum_{t\in I(u)}\log\Big(1+\sigma^{-2}\sigma^2_{t-1,I(u)}(x_t)\Big) \leq \gamma_T(S(u)).$$

Using $k(x,x) \leq \kappa^2$, we have $\sigma^2_{t-1,I(u)}(x_t) \leq \kappa^2$. For any $a \in [0,\kappa^2]$, the inequality $\log(1+a/\sigma^2) \geq a/(\sigma^2+a)$ implies

$$a \leq (\sigma^2+\kappa^2)\log\Big(1+\frac{a}{\sigma^2}\Big).$$

Applying this with $a = \sigma^2_{t-1,I(u)}(x_t)$ and summing yields

$$\sum_{t\in I(u)} \sigma^2_{t-1,I(u)}(x_t) \le (\sigma^2+\kappa^2) \sum_{t\in I(u)} \log\Big(1+\sigma^{-2}\sigma^2_{t-1,I(u)}(x_t)\Big) \le 2(\sigma^2+\kappa^2)\gamma_T(S(u)).$$

Integrating over u completes the proof. □

B.3 PROOF OF THEOREM 4

Lemma 9. *Let $w : \mathcal{X} \to \mathbb{R}_+$ and $w_{\max} \coloneqq \max_{x\in\mathcal{X}} w(x)$. Assume $k(x,x) \le \kappa^2$ for all x. For any (possibly adaptive) sequence $x_{1:T}$,*

$$\sum_{t=1}^{T} w(x_t)\sigma^2_{t-1}(x_t) \le 2(\sigma^2+\kappa^2)\Gamma_T(w).$$

Let $m_{\max} \coloneqq \lceil T^{4/7} \rceil$, and let $m \in \{0,\dots,m_{\max}\}$ be the (random) number of burn-in labels actually collected by the first loop of the algorithm. Define

$$\bar\sigma \coloneqq \max_{x\in\mathcal{X}} \sigma_m(x).$$

By construction, posterior standard deviations are nonincreasing in time, so for all $t \ge m+1$ and all $x \in \mathcal{X}$,

$$\sigma_{t-1}(x) \le \bar\sigma.$$

Let τ be the last labeled round of the entire algorithm (if the algorithm never stops requesting labels, set $\tau \coloneqq T$). We decompose

$$R_T = R_T^{\text{burn}} + R_T^{\text{main,explore}} + R_T^{\text{main,exploit}}.$$

During burn-in, every queried round incurs regret at most $2B + c_{\max}$, hence

$$R_T^{\text{burn}} \le m(2B+c_{\max}) \le m_{\max}(2B+c_{\max}).$$

Fix a labeled round $t \in \{m+1,\dots,\tau\}$.

Since the algorithm requests a label at round t

$$\max_{x\neq x_t^\ell} \tilde s_{t-1}(x) \ge \lambda.$$

Because $x_t^\ell \in \arg\max_x \ell_{t-1}(x)$, we have

$$\ell_{t-1}(x_t^\ell) \ge \ell_{t-1}(\bar x_t),$$

and therefore

$$\tilde s_{t-1}(\bar x_t) = \frac{\sqrt{\tilde D_{t-1}(\bar x_t)}}{\sqrt{D_{t-1}(\bar x_t)}}\Big(u_{t-1}(\bar x_t) - \ell_{t-1}(x_t^\ell)\Big) \le \frac{\sqrt{\tilde D_{t-1}(\bar x_t)}}{\sqrt{D_{t-1}(\bar x_t)}}\Big(u_{t-1}(\bar x_t) - \ell_{t-1}(\bar x_t)\Big) = s_{t-1}(\bar x_t).$$

Hence $s_{t-1}(\bar x_t) \ge \lambda$. Since we choose x_t maximizing s_{t-1} over $\{\bar x_t, x_t^\ell\}$,

$$s_{t-1}(x_t) \ge \lambda.$$

By the definition of s_{t-1},

$$s_{t-1}(x_t)^2 = \frac{\tilde D_{t-1}(x_t)}{D_{t-1}(x_t)}\big(u_{t-1}(x_t) - \ell_{t-1}(x_t)\big)^2 = \frac{\tilde D_{t-1}(x_t)}{D_{t-1}(x_t)} \cdot 4\beta_T\sigma^2_{t-1}(x_t),$$

so

$$D_{t-1}(x_t) \le \frac{4\beta_T}{\lambda^2}\tilde D_{t-1}(x_t)\sigma^2_{t-1}(x_t).$$

On the confidence event $\mathcal{E}$, we also have

$$r(x_t) + c(x_t) \le M_{t-1} - \ell_{t-1}(x_t) + c(x_t) = D_{t-1}(x_t),$$

because $f(x^*) \le M_{t-1}$ and $f(x_t) \ge \ell_{t-1}(x_t)$ on $\mathcal{E}$.

Therefore, for labeled main-phase rounds,

$$r(x_t) + c(x_t) \le \frac{4\beta_T}{\lambda^2} \tilde{D}_{t-1}(x_t)\sigma_{t-1}^2(x_t).$$

Summing over $t = m+1, \dots, \tau$ and extending by nonnegativity,

$$R_T^{\text{main,explore}} \le \frac{4\beta_T}{\lambda^2} \sum_{t=m+1}^{\tau} \tilde{D}_{t-1}(x_t)\sigma_{t-1}^2(x_t).$$

Now use that for all $t \ge m+1$ and all x,

$$\tilde{D}_{t-1}(x) = 2\sqrt{\beta_T}\big(\sigma_{t-1}(x) + \sigma_{t-1}(x_t^u)\big) + c(x) \le 4\sqrt{\beta_T}\,\bar{\sigma} + c(x).$$

Hence

$$R_T^{\text{main,explore}} \le \frac{4\beta_T}{\lambda^2} \sum_{t=m+1}^{\tau} \big(4\sqrt{\beta_T}\bar{\sigma} + c(x_t)\big)\sigma_{t-1}^2(x_t).$$

Applying Lemma 9 to $w \equiv 1$ and $w = c$ yields

$$\sum_{t=m+1}^{\tau} \sigma_{t-1}^2(x_t) \le \sum_{t=1}^{T} \sigma_{t-1}^2(x_t) \le 2(\sigma^2+\kappa^2)\bar{\gamma}_T, \qquad \sum_{t=1}^{T} c(x_t)\sigma_{t-1}^2(x_t) \le 2(\sigma^2+\kappa^2)\bar{\Gamma}_T(c).$$

Therefore

$$R_T^{\text{main,explore}} \le \frac{8\beta_T(\sigma^2+\kappa^2)}{\lambda^2}\Big(\bar{\Gamma}_T(c) + 4\sqrt{\beta_T}\,\bar{\sigma}\,\bar{\gamma}_T\Big).$$

Assume the algorithm stops at round $\tau+1$ (otherwise this term is zero). Let $x_{\text{stop}}^{\ell} := x_{\tau+1}^{\ell}$, which is computed from the posterior at time τ.

If $x^* = x_{\text{stop}}^{\ell}$, then the exploitation regret is zero. Otherwise, $x^* \ne x_{\text{stop}}^{\ell}$ is among the challengers and the stopping rule gives

$$\tilde{s}_\tau(x^*) < \lambda.$$

We claim $D_\tau(x^*) \le \tilde{D}_\tau(x^*)$ on $\mathcal{E}$, where

$$D_\tau(x^*) = M_\tau - \ell_\tau(x^*) + c(x^*), \qquad \tilde{D}_\tau(x^*) = 2\sqrt{\beta_T}\big(\sigma_\tau(x^*) + \sigma_\tau(x_\tau^u)\big) + c(x^*).$$

Indeed, with $x_\tau^u \in \arg\max_y u_\tau(y)$ and $M_\tau = u_\tau(x_\tau^u)$,

$$\begin{aligned} M_\tau - \ell_\tau(x^*) &= u_\tau(x_\tau^u) - \ell_\tau(x^*) \\ &= \mu_\tau(x_\tau^u) + \sqrt{\beta_T}\sigma_\tau(x_\tau^u) - \mu_\tau(x^*) + \sqrt{\beta_T}\sigma_\tau(x^*) \\ &\le 2\sqrt{\beta_T}\big(\sigma_\tau(x_\tau^u) + \sigma_\tau(x^*)\big), \end{aligned}$$

where we used $f(x_\tau^u) \le f(x^*)$ and the confidence event $\mathcal{E}$.

Thus

$$\frac{\sqrt{\tilde{D}_\tau(x^*)}}{\sqrt{D_\tau(x^*)}} \ge 1 \quad \Rightarrow \quad \hat{\Delta}_\tau(x^*) = u_\tau(x^*) - \ell_\tau(x_{\text{stop}}^{\ell}) \le \tilde{s}_\tau(x^*) < \lambda.$$

Since $f(x^*) \le u_\tau(x^*)$ and $f(x_{\text{stop}}^{\ell}) \ge \ell_\tau(x_{\text{stop}}^{\ell})$ on $\mathcal{E}$, we obtain

$$f(x^*) - f(x_{\text{stop}}^{\ell}) \le \lambda.$$

Therefore

$$R_T^{\text{main,exploit}} \le T\lambda.$$

There are two cases.

(i) Early burn-in stop: If the first loop breaks early, then by line 5 (evaluated at the posterior after m burn-in labels),

$$\bar{\sigma}\,\bar{\gamma}_T \le \bar{\Gamma}_T(c).$$

(ii) Burn-in reaches $m_{\max}$: Then burn-in queried maximizers of posterior variance for $m_{\max}$ rounds. Since posterior variances are nonincreasing, each burn-in query before termination has posterior variance at least $\bar{\sigma}^2$, and hence

$$m_{\max}\log\left(1+\frac{\bar{\sigma}^2}{\sigma^2}\right) \le \sum_{t=1}^{m_{\max}} \log\left(1+\frac{\sigma_{t-1}^2(x_t)}{\sigma^2}\right) \le 2\bar{\gamma}_T.$$

Thus

$$\bar{\sigma}^2 \le \sigma^2\Big(\exp\big(\tfrac{2\bar{\gamma}_T}{m_{\max}}\big)-1\Big),$$

and in particular (for large enough T, or after adjusting constants),

$$\bar{\sigma} \le \sigma\sqrt{\frac{2\bar{\gamma}_T}{m_{\max}}}.$$

Combining both cases,

$$\bar{\sigma}\,\bar{\gamma}_T \le \bar{\Gamma}_T(c) + \sigma\bar{\gamma}_T\sqrt{\frac{2\bar{\gamma}_T}{m_{\max}}}.$$

Putting everything together,

$$R_T \le m_{\max}(2B+c_{\max}) + \frac{8\beta_T(\sigma^2+\kappa^2)}{\lambda^2}\Big(\bar{\Gamma}_T(c) + 4\sqrt{\beta_T}\,\bar{\sigma}\,\bar{\gamma}_T\Big) + T\lambda.$$

Using the algorithm's choice

$$\lambda = 2\left(\frac{\beta_T(\sigma^2+\kappa^2)\big(\bar{\sigma}\bar{\gamma}_T + \bar{\Gamma}_T(c)\big)}{T}\right)^{1/3},$$

Since the upper bounds are chosen consistently, $\bar{\Gamma}_T(c) \ge c_{\min}\bar{\gamma}_T$; therefore the display above and the choice $m_{\max} = \lceil T^{4/7}\rceil$ make the residual variance term from case (ii) lower order (up to polylogarithmic factors). The displayed choice of λ then yields the advertised $T^{2/3}$ behavior.

B.4 PROOF OF COROLLARY 2

Let $S \subseteq \mathcal{X}$ with $|S| = j$. For the independent (delta) kernel, any query sequence $x_{1:T} \in S^T$ is fully characterized by the pull counts $n_i := \sum_{t=1}^{T} \mathbb{1}\{x_t = i\}$ for $i \in S$, with $\sum_{i\in S} n_i = T$.

For this kernel, the Gram matrix $K_T(x_{1:T})$ has one nonzero eigenvalue n_i for each arm $i \in S$ with $n_i > 0$ (and the remaining eigenvalues are zero). Therefore

$$\frac{1}{2}\log\det\Big(I_T + \sigma^{-2}K_T(x_{1:T})\Big) = \frac{1}{2}\sum_{i\in S}\log\Big(1+\frac{n_i}{\sigma^2}\Big).$$

Hence

$$\gamma_T(S) = \max_{\substack{n_i\in\mathbb{N}_0 \\ \sum_{i\in S} n_i = T}} \frac{1}{2}\sum_{i\in S}\log\Big(1+\frac{n_i}{\sigma^2}\Big).$$

Using concavity of $u \mapsto \log(1 + u/\sigma^2)$ (equivalently, Jensen's inequality),

$$\frac{1}{2}\sum_{i\in S}\log\Big(1+\frac{n_i}{\sigma^2}\Big) \le \frac{j}{2}\log\Big(1+\frac{1}{j\sigma^2}\sum_{i\in S} n_i\Big) = \frac{j}{2}\log\Big(1+\frac{T}{j\sigma^2}\Big) \le \frac{j}{2}\log\Big(1+\frac{T}{\sigma^2}\Big).$$

Therefore, for every finite S,

$$\gamma_T(S) \le \frac{|S|}{2}\log\Big(1+\frac{T}{\sigma^2}\Big).$$

Now define the superlevel sets

$$S_v \coloneqq \{i \in \mathcal{X} : c(i) \ge v\}.$$

By Definition 1,

$$\Gamma_T(c) = \int_0^{c_{\max}} \gamma_T(S_v)\,dv \le \frac{1}{2}\log\Big(1+\frac{T}{\sigma^2}\Big)\int_0^{c_{\max}} |S_v|\,dv.$$

Finally, using the layer-cake identity for finite sums,

$$\int_0^{c_{\max}} |S_v|\,dv = \int_0^{c_{\max}} \sum_{i=1}^{K} \mathbb{1}\{c(i) \ge v\}\,dv = \sum_{i=1}^{K}\int_0^{c_{\max}} \mathbb{1}\{c(i) \ge v\}\,dv = \sum_{i=1}^{K} c(i).$$

Combining the displays gives

$$\Gamma_T(c) \le \frac{1}{2}\log\Big(1+\frac{T}{\sigma^2}\Big)\sum_{i=1}^{K} c(i),$$

as claimed.

B.5 PROOF OF THEOREM 5

Write $\bar{c} \coloneqq \sum_{x\in\mathcal{X}} c(x)$ and $c_{\max} \coloneqq \max_{x\in\mathcal{X}} c(x)$. Fix any $\Delta \in (0, 1/2]$ to be chosen later.

For an algorithm $\mathcal{A}$, let x_t be the arm played at round t and $d_t \in \{0, 1\}$ indicate whether the reward is observed. Define, for each $x \in \mathcal{X}$,

$$L_T(x) \coloneqq \sum_{t=1}^{T} \mathbb{1}\{x_t = x,\ d_t = 1\}, \qquad P_T(x) \coloneqq \sum_{t=1}^{T} \mathbb{1}\{x_t = x\}.$$

Let $x^\star$ be an optimal arm under the instance in question, and recall

$$R_T = \sum_{t=1}^{T} \big(\mu(x^\star) - \mu(x_t)\big) + \sum_{t=1}^{T} c(x_t)\mathbb{1}\{d_t = 1\}.$$

We prove the following stronger statement: there exists a universal constant $C_0 > 0$ such that

$$\sup_{\nu\in\mathbb{S}} \mathbb{E}_\nu[R_T] \ge C_0 \min\Big\{\frac{\bar{c}}{\Delta^2},\ T\Delta\Big\}.$$

Choosing $\Delta \coloneqq \min\{1/2, (\bar{c}/T)^{1/3}\}$ then yields $\min\{\bar{c}/\Delta^2,\ T\Delta\} \ge c_1 \min\{\bar{c}^{1/3}T^{2/3},\ T\}$ for a universal constant $c_1 > 0$, which proves the theorem.

Let $x_{\max} \in \arg\max_{x\in\mathcal{X}} c(x)$, so $c(x_{\max}) = c_{\max}$. Consider two instances ν_+ and ν_- defined by

$$\mu_+(x_{\max}) \coloneqq \frac{1}{2} + \Delta, \qquad \mu_+(x) \coloneqq \frac{1}{2}\ \ (x \ne x_{\max}),$$

and

$$\mu_-(x_{\max}) \coloneqq \frac{1}{2} - \Delta, \qquad \mu_-(x) \coloneqq \frac{1}{2}\ \ (x \ne x_{\max}).$$

Let $\mathbb{P}_+$ and $\mathbb{P}_-$ denote the induced distributions over the interaction transcript up to time T. Define the event

$$E \coloneqq \{P_T(x_{\max}) \ge T/2\}.$$

On ν_+, every play of an arm $x \neq x_{\max}$ incurs reward regret Δ, hence on E^c the reward regret is at least $(T/2)\Delta$. On ν_-, every play of $x_{\max}$ incurs reward regret Δ, hence on E the reward regret is at least $(T/2)\Delta$. Therefore,

$$\mathbb{E}_+[R_T] + \mathbb{E}_-[R_T] \geq \frac{T\Delta}{2}\big(\mathbb{P}_+(E^c) + \mathbb{P}_-(E)\big).$$

By the Bretagnolle–Huber inequality, for any event A, $\mathbb{P}_+(A^c) + \mathbb{P}_-(A) \geq \frac{1}{2}\exp(-\mathrm{KL}(\mathbb{P}_+, \mathbb{P}_-))$. Applying this with $A = E$ gives

$$\mathbb{P}_+(E^c) + \mathbb{P}_-(E) \geq \frac{1}{2}\exp\big(-\mathrm{KL}(\mathbb{P}_+, \mathbb{P}_-)\big).$$

Only labeled pulls of $x_{\max}$ produce observations whose distribution differs between ν_+ and ν_-. For Gaussian rewards with unit variance,

$$\mathrm{KL}\big(\mathcal{N}(\tfrac{1}{2}+\Delta, 1), \mathcal{N}(\tfrac{1}{2}-\Delta, 1)\big) = 2\Delta^2.$$

Using the chain rule for KL under adaptive sampling with partial observations,

$$\mathrm{KL}(\mathbb{P}_+, \mathbb{P}_-) = 2\Delta^2\,\mathbb{E}_+[L_T(x_{\max})].$$

Now consider two subcases.

If $\mathbb{E}_+[L_T(x_{\max})] \geq \frac{1}{4\Delta^2}$, then

$$\mathbb{E}_+[R_T] \geq c(x_{\max})\,\mathbb{E}_+[L_T(x_{\max})] \geq \frac{c_{\max}}{4\Delta^2} \geq \frac{\bar{c}}{16\Delta^2},$$

where we used $c_{\max} \geq \bar{c}/4$.

Otherwise $\mathbb{E}_+[L_T(x_{\max})] < \frac{1}{4\Delta^2}$, so $\mathrm{KL}(\mathbb{P}_+, \mathbb{P}_-) < \frac{1}{2}$ and hence

$$\mathbb{P}_+(E^c) + \mathbb{P}_-(E) \geq \frac{1}{2}e^{-1/2}.$$

Thus,

$$\max\{\mathbb{E}_+[R_T], \mathbb{E}_-[R_T]\} \geq \frac{1}{2}\cdot\frac{T\Delta}{2}\cdot\frac{1}{2}e^{-1/2} = \frac{e^{-1/2}}{8}\,T\Delta.$$

Combining the two subcases, in the regime $c_{\max} \geq \bar{c}/4$ we have shown

$$\sup_{\nu\in\mathbb{S}} \mathbb{E}_\nu[R_T] \geq C_1 \min\left\{\frac{\bar{c}}{\Delta^2},\, T\Delta\right\}$$

for a universal constant $C_1 > 0$.

Let ν_0 be the instance with $\mu_0(x) = 0$ for all x. For each $x \in \mathcal{X}$, let ν_x be the instance with $\mu_x(x) = \Delta$ and $\mu_x(z) = 0$ for $z \neq x$. Let $\mathbb{P}_0$ and $\mathbb{P}_x$ be the induced transcript distributions.

Let $x^\star \in \arg\max_{z\in\mathcal{X}} P_T(z)$ be the most-played arm (ties broken measurably), and define events

$$E_x \coloneqq \{x^\star = x\}.$$

Under ν_x, arm x is uniquely optimal with gap Δ. On E_x^c, we have $P_T(x) \leq T/2$, hence at least $T/2$ plays are of suboptimal arms, so the reward regret is at least $(T/2)\Delta$. Therefore,

$$\mathbb{E}_x[R_T] \geq \frac{T\Delta}{2}\,\mathbb{P}_x(E_x^c).$$

Next, compute the KL divergences from ν_0 to ν_x. Only labeled pulls of arm x produce observations whose distribution differs between ν_0 and ν_x. Since $\mathrm{KL}(\mathcal{N}(0,1), \mathcal{N}(\Delta,1)) = \Delta^2/2$, we have

$$\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_x) = \frac{\Delta^2}{2}\,\mathbb{E}_0[L_T(x)].$$

Apply Bretagnolle–Huber with the event E_x:

$$\mathbb{P}_0(E_x) + \mathbb{P}_x(E_x^c) \geq \frac{1}{2} \exp\big(-\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_x)\big).$$

Multiply by $p(x) \coloneqq c(x)/\bar{c}$ and sum over $x \in \mathcal{X}$:

$$\sum_x p(x)\mathbb{P}_x(E_x^c) \geq \frac{1}{2} \sum_x p(x) \exp\big(-\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_x)\big) - \sum_x p(x)\mathbb{P}_0(E_x).$$

Since E_x form a partition, $\sum_x \mathbb{P}_0(E_x) = 1$, hence

$$\sum_x p(x)\mathbb{P}_0(E_x) \leq \max_x p(x) = \frac{c_{\max}}{\bar{c}} < \frac{1}{4}.$$

Using Jensen's inequality and convexity of $u \mapsto e^{-u}$,

$$\sum_x p(x) \exp\big(-\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_x)\big) \geq \exp\left(-\sum_x p(x)\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_x)\right).$$

Moreover,

$$\begin{aligned}\sum_x p(x)\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_x) &= \sum_x \frac{c(x)}{\bar{c}} \cdot \frac{\Delta^2}{2} \mathbb{E}_0[L_T(x)] \\ &= \frac{\Delta^2}{2\bar{c}}\, \mathbb{E}_0\left[\sum_x c(x) L_T(x)\right] = \frac{\Delta^2}{2\bar{c}}\, \mathbb{E}_0[R_T],\end{aligned}$$

where the last equality uses that under ν_0 the reward regret is identically zero.

Now split into two cases.

If $\mathbb{E}_0[R_T] \geq \bar{c}/\Delta^2$, then

$$\sup_{\nu \in \mathbb{S}} \mathbb{E}_\nu[R_T] \geq \mathbb{E}_0[R_T] \geq \frac{\bar{c}}{\Delta^2}.$$

Otherwise $\mathbb{E}_0[R_T] < \bar{c}/\Delta^2$, so $\sum_x p(x)\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_x) < 1/2$ and hence

$$\sum_x p(x)\mathbb{P}_x(E_x^c) \geq \frac{1}{2} e^{-1/2} - \frac{1}{4} \coloneqq \alpha,$$

where $\alpha > 0$ is a universal constant. Since $\sum_x p(x)\mathbb{P}_x(E_x^c) \leq \max_x \mathbb{P}_x(E_x^c)$, there exists $x \in \mathcal{X}$ such that $\mathbb{P}_x(E_x^c) \geq \alpha$. For this x,

$$\sup_{\nu \in \mathbb{S}} \mathbb{E}_\nu[R_T] \geq \mathbb{E}_x[R_T] \geq \frac{T\Delta}{2}\alpha.$$

Combining the two cases yields, when $c_{\max} < \bar{c}/4$,

$$\sup_{\nu \in \mathbb{S}} \mathbb{E}_\nu[R_T] \geq C_2 \min\left\{\frac{\bar{c}}{\Delta^2},\ T\Delta\right\}$$

for a universal constant $C_2 > 0$.

By Corollary 2 (for the independent-arm kernel with unit reward variance), we have $\Gamma_T(c) \leq \frac{1}{2}\log(1+T)\bar{c}$, and therefore $\bar{c} \geq \frac{2\Gamma_T(c)}{\log(1+T)}$. Substituting this into the bound proved above and absorbing the factor $2^{1/3}$ into the universal constant yields

$$\sup_{\nu \in \mathbb{S}} \mathbb{E}_\nu[R_T] \geq C' \min\left\{\left(\frac{\Gamma_T(c)}{\log(1+T)}\right)^{1/3} T^{2/3},\ T\right\},$$

which proves the theorem.

B.6 PROOF OF EXAMPLE 1

Since the cost takes only two values, for every $v \in [0, M]$ the superlevel sets satisfy

$$\{x \in \mathcal{X} : c(x) \geq v\} = \begin{cases} \mathcal{X}, & 0 \leq v \leq \varepsilon, \\ A, & \varepsilon < v \leq M. \end{cases}$$

Therefore

$$\Gamma_T(c) = \int_0^M \gamma_T(\{c \geq v\})\, dv = \varepsilon\, \gamma_T(\mathcal{X}) + (M - \varepsilon)\, \gamma_T(A).$$

For any sequence $x_{1:T} \in \mathcal{X}^T$, let K_T be the corresponding kernel matrix. Then $I_T + \sigma^{-2} K_T$ is positive semidefinite, and Hadamard's inequality gives

$$\det(I_T + \sigma^{-2} K_T) \leq \prod_{t=1}^{T} \big(1 + \sigma^{-2} k(x_t, x_t)\big).$$

Since $k(x, x) = s(x)^2 \leq 1$ for all $x \in \mathcal{X}$, we obtain

$$\det(I_T + \sigma^{-2} K_T) \leq (1 + \sigma^{-2})^T,$$

and hence

$$\gamma_T(\mathcal{X}) = \sup_{x_{1:T} \in \mathcal{X}^T} \frac{1}{2} \log \det(I_T + \sigma^{-2} K_T) \leq \frac{T}{2} \log(1 + \sigma^{-2}).$$

Fix any sequence $x_{1:T} \in A^T$ and let $K_{T,A}$ be the corresponding kernel matrix. Using again Hadamard's inequality,

$$\det(I_T + \sigma^{-2} K_{T,A}) \leq \prod_{t=1}^{T} \big(1 + \sigma^{-2} k(x_t, x_t)\big).$$

For $x_t \in A$ we have $k(x_t, x_t) = \delta^2$, so

$$\det(I_T + \sigma^{-2} K_{T,A}) \leq (1 + \sigma^{-2} \delta^2)^T,$$

and thus

$$\gamma_T(A) \leq \frac{T}{2} \log(1 + \sigma^{-2} \delta^2) \leq \frac{T}{2} \cdot \sigma^{-2} \delta^2,$$

where we used $\log(1 + u) \leq u$. Set $\varepsilon = 1/T$ and $\delta^2 = \sigma^2/(TM)$. Then

$$\gamma_T(A) \leq \frac{T}{2} \cdot \sigma^{-2} \cdot \frac{\sigma^2}{TM} = \frac{1}{2M}.$$

Combining with Step 1 yields

$$\begin{aligned} \Gamma_T(c) &= \varepsilon\, \gamma_T(\mathcal{X}) + (M - \varepsilon)\, \gamma_T(A) \\ &\leq \frac{1}{T} \cdot \frac{T}{2} \log(1 + \sigma^{-2}) + M \cdot \frac{1}{2M} \\ &= \frac{1}{2} \log(1 + \sigma^{-2}) + \frac{1}{2}, \end{aligned}$$

which is independent of T and M even though $c_{\max} = M$ can be arbitrarily large.

B.7 PROOF OF THEOREM 6

Work on the confidence event $\mathcal{E}$ and let

$$w \coloneqq \Delta_c.$$

Let τ be the last labeled round (if the algorithm never stops requesting labels before the horizon, set $\tau \coloneqq T$).

We first use the theorem assumption on v. Since

$$v^3 = \frac{8\Gamma_T(c)\beta_T\sigma^2}{T}$$

and

$$T > \frac{8\Gamma_T(c)\beta_T\sigma^2}{\Delta_{\min}^3},$$

we have

$$v < \Delta_{\min}.$$

If $\tau = T$, there is no unlabeled suffix and nothing to prove. Assume $\tau < T$, and let $t \coloneqq \tau + 1$ be the stopping round. The algorithm stops because

$$u_{t-1}(\bar{x}_t) - \ell_{t-1}(x_t^\ell) \le v.$$

We claim that $x_t^\ell = x^*$.

Assume for contradiction that $x_t^\ell \neq x^*$. Since $\bar{x}_t \in \arg\max_{x \in \mathcal{X} \setminus \{x_t^\ell\}} u_{t-1}(x)$ and $x^* \in \mathcal{X} \setminus \{x_t^\ell\}$, we have

$$u_{t-1}(\bar{x}_t) \ge u_{t-1}(x^*).$$

On $\mathcal{E}$,

$$u_{t-1}(x^*) \ge f(x^*) \quad \text{and} \quad \ell_{t-1}(x_t^\ell) \le f(x_t^\ell),$$

so

$$u_{t-1}(\bar{x}_t) - \ell_{t-1}(x_t^\ell) \ge u_{t-1}(x^*) - \ell_{t-1}(x_t^\ell) \ge f(x^*) - f(x_t^\ell) = \Delta(x_t^\ell) \ge \Delta_{\min} > v,$$

which contradicts the stopping condition. Therefore $x_t^\ell = x^*$.

Hence after stopping the algorithm commits to an optimal action, so the exploitation regret is zero. Therefore

$$R_T = \sum_{t=1}^{\tau} \big(\Delta(x_t) + c(x_t)\big).$$

Fix a labeled round $t \le \tau$ and suppose $x_t \neq x^*$. We show

$$\Delta(x_t) \le 4\sqrt{\beta_T}\sigma_{t-1}(x_t).$$

There are two cases.

Case 1: $x_t = \bar{x}_t$. If $x_t^\ell \neq x^*$, then $x^* \in \mathcal{X} \setminus \{x_t^\ell\}$ and by definition of $\bar{x}_t$,

$$u_{t-1}(\bar{x}_t) \ge u_{t-1}(x^*).$$

Thus on $\mathcal{E}$,

$$\Delta(x_t) = f(x^*) - f(\bar{x}_t) \le u_{t-1}(x^*) - \ell_{t-1}(\bar{x}_t) \le u_{t-1}(\bar{x}_t) - \ell_{t-1}(\bar{x}_t) = 2\sqrt{\beta_T}\sigma_{t-1}(\bar{x}_t) \le 4\sqrt{\beta_T}\sigma_{t-1}(x_t).$$

If $x_t^\ell = x^*$, then $x_t = \bar{x}_t \neq x^*$. Since the algorithm chooses x_t as the larger-variance arm in $\{x_t^\ell, \bar{x}_t\}$, we have

$$\sigma_{t-1}(x_t^\ell) \le \sigma_{t-1}(\bar{x}_t) = \sigma_{t-1}(x_t).$$

On $\mathcal{E}$,

$$\Delta(x_t) = f(x^*) - f(\bar{x}_t) \le u_{t-1}(x^*) - \ell_{t-1}(\bar{x}_t) \le 2\sqrt{\beta_T}\big(\sigma_{t-1}(x^*) + \sigma_{t-1}(\bar{x}_t)\big) \le 4\sqrt{\beta_T}\sigma_{t-1}(x_t).$$

Case 2: $x_t = x_t^\ell$. Since $x_t \neq x^*$, we have $x_t^\ell \neq x^*$, hence $x^* \in \mathcal{X} \setminus \{x_t^\ell\}$ and

$$u_{t-1}(\bar{x}_t) \geq u_{t-1}(x^*).$$

Also, because the algorithm chose $x_t = x_t^\ell$ over $\bar{x}_t$, we have

$$\sigma_{t-1}(\bar{x}_t) \leq \sigma_{t-1}(x_t^\ell) = \sigma_{t-1}(x_t).$$

Therefore on $\mathcal{E}$,

$$\begin{aligned}
\Delta(x_t) &= f(x^*) - f(x_t^\ell) \\
&\leq u_{t-1}(x^*) - \ell_{t-1}(x_t^\ell) \\
&\leq u_{t-1}(\bar{x}_t) - \ell_{t-1}(x_t^\ell) \\
&\leq u_{t-1}(\bar{x}_t) - \ell_{t-1}(\bar{x}_t) + \ell_{t-1}(\bar{x}_t) - \ell_{t-1}(x_t^\ell) \\
&\leq u_{t-1}(\bar{x}_t) - \ell_{t-1}(\bar{x}_t) + u_{t-1}(x_t^\ell) - \ell_{t-1}(x_t^\ell) \\
&= 2\sqrt{\beta_T}\big(\sigma_{t-1}(\bar{x}_t) + \sigma_{t-1}(x_t^\ell)\big) \\
&\leq 4\sqrt{\beta_T}\sigma_{t-1}(x_t).
\end{aligned}$$

This proves the claimed bound for every labeled round with $x_t \neq x^*$.

Fix a labeled round $t \leq \tau$.

If $x_t \neq x^*$, then by definition of $w = \Delta_c$,

$$w(x_t) = \frac{\Delta(x_t) + c(x_t)}{\Delta(x_t)^2}.$$

Hence

$$\Delta(x_t) + c(x_t) = w(x_t)\Delta(x_t)^2 \leq 16\beta_T w(x_t)\sigma_{t-1}^2(x_t).$$

If $x_t = x^*$, then the per-round contribution is $c(x^*)$. Since $t \leq \tau$, the algorithm requested a label at round t, so

$$u_{t-1}(\bar{x}_t) - \ell_{t-1}(x_t^\ell) > v.$$

Here $x_t = x^*$ implies $x_t^\ell = x^*$, so

$$u_{t-1}(\bar{x}_t) - \ell_{t-1}(x^*) > v.$$

Also on $\mathcal{E}$,

$$\Delta(\bar{x}_t) = f(x^*) - f(\bar{x}_t) \leq u_{t-1}(x^*) - \ell_{t-1}(\bar{x}_t).$$

Using

$$u_{t-1}(\bar{x}_t) - \ell_{t-1}(x^*) \geq 0$$

(which follows from the strict inequality above), we get

$$\mu_{t-1}(x^*) - \mu_{t-1}(\bar{x}_t) \leq \sqrt{\beta_T}\big(\sigma_{t-1}(x^*) + \sigma_{t-1}(\bar{x}_t)\big),$$

and therefore

$$u_{t-1}(x^*) - \ell_{t-1}(\bar{x}_t) \leq 2\sqrt{\beta_T}\big(\sigma_{t-1}(x^*) + \sigma_{t-1}(\bar{x}_t)\big).$$

Since $\bar{x}_t \neq x^*$, we have $\Delta(\bar{x}_t) \geq \Delta_{\min}$, so

$$\Delta_{\min} \leq 2\sqrt{\beta_T}\big(\sigma_{t-1}(x^*) + \sigma_{t-1}(\bar{x}_t)\big).$$

Because the algorithm chose $x_t = x^*$ over $\bar{x}_t$, we have

$$\sigma_{t-1}(\bar{x}_t) \leq \sigma_{t-1}(x^*),$$

hence

$$\Delta_{\min} \leq 4\sqrt{\beta_T}\sigma_{t-1}(x^*).$$

Using $w(x^*) = c(x^*)/\Delta_{\min}^2$, we conclude

$$c(x^*) = w(x^*)\Delta_{\min}^2 \le 16\beta_T w(x^*)\sigma_{t-1}^2(x^*) = 16\beta_T w(x_t)\sigma_{t-1}^2(x_t).$$

Thus for every labeled round $t \le \tau$,

$$\Delta(x_t) + c(x_t) \le 16\beta_T w(x_t)\sigma_{t-1}^2(x_t).$$

Summing over $t = 1, \dots, \tau$ gives

$$R_T = \sum_{t=1}^{\tau} \big(\Delta(x_t) + c(x_t)\big) \le 16\beta_T \sum_{t=1}^{\tau} w(x_t)\sigma_{t-1}^2(x_t).$$

Apply Lemma 9 to the labeled prefix $x_{1:\tau}$ (which is a valid adaptive sequence of observed points) to obtain

$$\sum_{t=1}^{\tau} w(x_t)\sigma_{t-1}^2(x_t) \le 2(\sigma^2 + \kappa^2) \int_0^{w_{\max}} \gamma_\tau(\{x \in \mathcal{X} : w(x) \ge u\})du,$$

where $w_{\max} \coloneqq \max_{x\in\mathcal{X}} w(x)$.

Since $\tau \le T$ and $\gamma_\tau(S) \le \gamma_T(S)$ for every $S \subseteq \mathcal{X}$,

$$\int_0^{w_{\max}} \gamma_\tau(\{x : w(x) \ge u\})du \le \int_0^{w_{\max}} \gamma_T(\{x : w(x) \ge u\})du = \Gamma_T^{\mathrm{gap}}(\Delta_c).$$

Combining the displays,

$$R_T \le 16\beta_T \cdot 2(\sigma^2 + \kappa^2)\Gamma_T^{\mathrm{gap}}(\Delta_c) = 32\beta_T(\sigma^2 + \kappa^2)\Gamma_T^{\mathrm{gap}}(\Delta_c).$$

This proves the theorem.

B.8 PROOF OF COROLLARY 3

Let $w(x) \coloneqq \Delta_c(x)$ and $w_{\max} \coloneqq \max_{x\in\mathcal{X}} w(x)$. For $u \in [0, w_{\max}]$, define

$$S_u \coloneqq \{x \in \mathcal{X} : w(x) \ge u\}, \qquad j(u) \coloneqq |S_u|.$$

By Definition 3,

$$\Gamma_T^{\mathrm{gap}}(\Delta_c) = \int_0^{w_{\max}} \gamma_T(S_u)\, du.$$

Fix any subset $S \subseteq \mathcal{X}$ with $|S| = j$. For the independent (delta) kernel, if arm $i \in S$ is queried n_i times with $\sum_{i=1}^{j} n_i = T$, then the corresponding kernel matrix is block-diagonal by arm counts, and

$$\gamma_T(S) = \frac{1}{2}\sum_{i=1}^{j} \log\Big(1 + \frac{n_i}{\sigma^2}\Big).$$

Since $a \mapsto \log(1 + a/\sigma^2)$ is concave, Jensen's inequality gives

$$\gamma_T(S) \le \frac{j}{2}\log\Big(1 + \frac{1}{j}\sum_{i=1}^{j}\frac{n_i}{\sigma^2}\Big) = \frac{j}{2}\log\Big(1 + \frac{T}{j\sigma^2}\Big) \le \frac{j}{2}\log\Big(1 + \frac{T}{\sigma^2}\Big).$$

Applying this with $S = S_u$ yields

$$\gamma_T(S_u) \le \frac{j(u)}{2}\log\Big(1 + \frac{T}{\sigma^2}\Big) \qquad \text{for all } u \in [0, w_{\max}].$$

Therefore

$$\Gamma_T^{\mathrm{gap}}(\Delta_c) = \int_0^{w_{\max}} \gamma_T(S_u)\, du \\ \le \frac{1}{2}\log\Big(1+\frac{T}{\sigma^2}\Big)\int_0^{w_{\max}} j(u)\, du.$$

It remains to evaluate the integral. Using the layer-cake identity on the finite set $\mathcal{X}$,

$$\int_0^{w_{\max}} j(u)\, du = \int_0^{w_{\max}} \sum_{x\in\mathcal{X}} \mathbb{1}\{w(x)\ge u\}\, du \\ = \sum_{x\in\mathcal{X}} \int_0^{w_{\max}} \mathbb{1}\{w(x)\ge u\}\, du \\ = \sum_{x\in\mathcal{X}} w(x) = \sum_{x\in\mathcal{X}} \Delta_c(x).$$

Combining the displays proves

$$\Gamma_T^{\mathrm{gap}}(\Delta_c) \le \frac{1}{2}\log\Big(1+\frac{T}{\sigma^2}\Big)\sum_{x\in\mathcal{X}} \Delta_c(x).$$

B.9 PROOF OF THEOREM 7

Fix an instance $\nu \in \mathbb{S}$ with a unique optimal arm x^* and let

$$\mu^* \coloneqq \mu(x^*), \qquad \Delta(x) \coloneqq \mu^* - \mu(x), \qquad \Delta_{\min} \coloneqq \min_{x\neq x^*} \Delta(x) > 0.$$

For each arm x, define the number of labeled pulls and total pulls up to time T by

$$L_T(x) \coloneqq \sum_{t=1}^{T} \mathbb{1}\{x_t = x,\ d_t = 1\}, \qquad P_T(x) \coloneqq \sum_{t=1}^{T} \mathbb{1}\{x_t = x\}.$$

The regret decomposes as

$$R_T = \sum_{x\in\mathcal{X}} c(x) L_T(x) + \sum_{x\neq x^*} \Delta(x) P_T(x).$$

Since $P_T(x) \ge L_T(x)$ for all x, we have the pointwise lower bound

$$R_T \ge c(x^*) L_T(x^*) + \sum_{x\neq x^*} (\Delta(x) + c(x)) L_T(x). \tag{1}$$

It therefore suffices to lower bound $\mathbb{E}_\nu[L_T(x)]$ for every arm x.

For two distributions P, Q on the same measurable space and any event A,

$$\mathrm{KL}(P,Q) \ge \mathrm{kl}(P(A), Q(A)),$$

where kl is the binary relative entropy. Also, for $p, q \in [0,1]$,

$$\mathrm{kl}(p,q) = p\log\frac{1}{q} + (1-p)\log\frac{1}{1-q} - \big(-p\log p - (1-p)\log(1-p)\big) \ge p\log\frac{1}{q} - \log 2.$$

We will repeatedly use the following consequence of consistency: if under some instance $\bar{\nu}$ an event B_T implies $R_T \ge gT$ for a constant $g > 0$, then for every $\eta \in (0,1)$,

$$\mathbb{P}_{\bar{\nu}}(B_T) \le \frac{\mathbb{E}_{\bar{\nu}}[R_T]}{gT} = o(T^{-\eta}),$$

because consistency gives $\mathbb{E}_{\bar{\nu}}[R_T] = o(T^{1-\eta})$.

Fix $x \neq x^*$ and fix $\epsilon > 0$. Let $\nu^{(x,\epsilon)}$ be the instance obtained from ν by changing only arm x to

$$\mu^{(x,\epsilon)}(x) \coloneqq \mu^* + \epsilon,$$

and leaving all other arm means unchanged. Under $\nu^{(x,\epsilon)}$, arm x is uniquely optimal.

Define the event

$$A_T^{x,\epsilon} \coloneqq \{P_T(x) < T/2\}.$$

Under the true instance ν, on $(A_T^{x,\epsilon})^c$ the algorithm plays the suboptimal arm x at least $T/2$ times, so

$$R_T \geq \frac{T}{2}\Delta(x).$$

Hence for every $\eta \in (0,1)$,

$$\mathbb{P}_\nu((A_T^{x,\epsilon})^c) = o(T^{-\eta}), \qquad \mathbb{P}_\nu(A_T^{x,\epsilon}) \to 1.$$

Under the alternative instance $\nu^{(x,\epsilon)}$, on $A_T^{x,\epsilon}$ the algorithm plays arms other than the optimal arm x at least $T/2$ times. Every such pull incurs reward regret at least ϵ, so

$$R_T \geq \frac{T}{2}\epsilon \qquad \text{on } A_T^{x,\epsilon}.$$

Therefore for every $\eta \in (0,1)$,

$$\mathbb{P}_{\nu^{(x,\epsilon)}}(A_T^{x,\epsilon}) = o(T^{-\eta}).$$

Let $\mathbb{P}_\nu^T$ and $\mathbb{P}_{\nu^{(x,\epsilon)}}^T$ denote the laws of the full interaction transcript up to time T. Applying the previous inequalities with $A = A_T^{x,\epsilon}$ gives, for every $\eta \in (0,1)$,

$$\begin{aligned}\frac{\mathrm{KL}(\mathbb{P}_\nu^T, \mathbb{P}_{\nu^{(x,\epsilon)}}^T)}{\log T} &\geq \frac{\mathrm{kl}(\mathbb{P}_\nu(A_T^{x,\epsilon}), \mathbb{P}_{\nu^{(x,\epsilon)}}(A_T^{x,\epsilon}))}{\log T} \\ &\geq \frac{\mathbb{P}_\nu(A_T^{x,\epsilon}) \log \frac{1}{\mathbb{P}_{\nu^{(x,\epsilon)}}(A_T^{x,\epsilon})} - \log 2}{\log T}.\end{aligned}$$

Since $\mathbb{P}_\nu(A_T^{x,\epsilon}) \to 1$ and $\mathbb{P}_{\nu^{(x,\epsilon)}}(A_T^{x,\epsilon}) = o(T^{-\eta})$, we obtain

$$\liminf_{T\to\infty} \frac{\mathrm{KL}(\mathbb{P}_\nu^T, \mathbb{P}_{\nu^{(x,\epsilon)}}^T)}{\log T} \geq \eta.$$

As this holds for every $\eta \in (0,1)$,

$$\liminf_{T\to\infty} \frac{\mathrm{KL}(\mathbb{P}_\nu^T, \mathbb{P}_{\nu^{(x,\epsilon)}}^T)}{\log T} \geq 1. \tag{2}$$

Only labeled observations from arm x differ between ν and $\nu^{(x,\epsilon)}$, so by the chain rule for KL under adaptive sampling,

$$\mathrm{KL}(\mathbb{P}_\nu^T, \mathbb{P}_{\nu^{(x,\epsilon)}}^T) = \mathbb{E}_\nu[L_T(x)]\mathrm{KL}\big(\mathcal{N}(\mu(x),\sigma^2), \mathcal{N}(\mu^* + \epsilon, \sigma^2)\big).$$

For Gaussian rewards with common variance σ^2,

$$\mathrm{KL}\big(\mathcal{N}(\mu(x),\sigma^2), \mathcal{N}(\mu^* + \epsilon, \sigma^2)\big) = \frac{(\Delta(x)+\epsilon)^2}{2\sigma^2}.$$

Combining with (2),

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[L_T(x)]}{\log T} \geq \frac{2\sigma^2}{(\Delta(x)+\epsilon)^2}. \tag{3}$$

Letting $\epsilon \downarrow 0$ yields

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[L_T(x)]}{\log T} \geq \frac{2\sigma^2}{\Delta(x)^2} \qquad \text{for all } x \neq x^*. \tag{4}$$

Fix $\epsilon > 0$ and choose any arm $x_{\min}$ with $\Delta(x_{\min}) = \Delta_{\min}$. Define an alternative instance $\nu^{(\dagger,\epsilon)}$ by changing only arm x^* to

$$\mu^{(\dagger,\epsilon)}(x^*) \coloneqq \mu^* - (\Delta_{\min} + \epsilon),$$

and leaving all other arm means unchanged. Under $\nu^{(\dagger,\epsilon)}$, the arm x^* is suboptimal, and $x_{\min}$ is optimal.

Define the event

$$A_T^{*,\epsilon} \coloneqq \{P_T(x^*) \geq T/2\}.$$

Under the true instance ν, on $(A_T^{*,\epsilon})^c$ the algorithm plays suboptimal arms at least $T/2$ times, and each such pull has reward regret at least $\Delta_{\min}$, so

$$R_T \geq \frac{T}{2}\Delta_{\min}.$$

Hence for every $\eta \in (0,1)$,

$$\mathbb{P}_\nu((A_T^{*,\epsilon})^c) = o(T^{-\eta}), \qquad \mathbb{P}_\nu(A_T^{*,\epsilon}) \to 1.$$

Under $\nu^{(\dagger,\epsilon)}$, on $A_T^{*,\epsilon}$ the algorithm plays the suboptimal arm x^* at least $T/2$ times. Its reward gap is at least ϵ (compared with $x_{\min}$), so

$$R_T \geq \frac{T}{2}\epsilon \qquad \text{on } A_T^{*,\epsilon}.$$

Thus for every $\eta \in (0,1)$,

$$\mathbb{P}_{\nu^{(\dagger,\epsilon)}}(A_T^{*,\epsilon}) = o(T^{-\eta}).$$

Exactly as above, data processing and the binary KL lower bound imply

$$\liminf_{T\to\infty} \frac{\mathrm{KL}(\mathbb{P}_\nu^T, \mathbb{P}_{\nu^{(\dagger,\epsilon)}}^T)}{\log T} \geq 1. \tag{5}$$

Only labeled observations from arm x^* differ between ν and $\nu^{(\dagger,\epsilon)}$, hence

$$\mathrm{KL}(\mathbb{P}_\nu^T, \mathbb{P}_{\nu^{(\dagger,\epsilon)}}^T) = \mathbb{E}_\nu[L_T(x^*)]\mathrm{KL}\big(\mathcal{N}(\mu^*,\sigma^2), \mathcal{N}(\mu^* - (\Delta_{\min}+\epsilon),\sigma^2)\big) = \mathbb{E}_\nu[L_T(x^*)]\frac{(\Delta_{\min}+\epsilon)^2}{2\sigma^2}.$$

Combining with (5),

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[L_T(x^*)]}{\log T} \geq \frac{2\sigma^2}{(\Delta_{\min}+\epsilon)^2}. \tag{6}$$

Letting $\epsilon \downarrow 0$ gives

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[L_T(x^*)]}{\log T} \geq \frac{2\sigma^2}{\Delta_{\min}^2}. \tag{7}$$

From (1),

$$\frac{\mathbb{E}_\nu[R_T]}{\log T} \geq c(x^*)\frac{\mathbb{E}_\nu[L_T(x^*)]}{\log T} + \sum_{x\neq x^*}(\Delta(x)+c(x))\frac{\mathbb{E}_\nu[L_T(x)]}{\log T}.$$

Taking $\liminf_{T\to\infty}$ and using that all terms are nonnegative,

$$\begin{aligned}\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[R_T]}{\log T} &\geq c(x^*)\liminf_{T\to\infty}\frac{\mathbb{E}_\nu[L_T(x^*)]}{\log T} + \sum_{x\neq x^*}(\Delta(x)+c(x))\liminf_{T\to\infty}\frac{\mathbb{E}_\nu[L_T(x)]}{\log T} \\ &\geq c(x^*)\frac{2\sigma^2}{\Delta_{\min}^2} + \sum_{x\neq x^*}(\Delta(x)+c(x))\frac{2\sigma^2}{\Delta(x)^2}.\end{aligned}$$

By the definition of Δ_c,

$$\Delta_c(x^*) = \frac{c(x^*)}{\Delta_{\min}^2}, \qquad \Delta_c(x) = \frac{\Delta(x)+c(x)}{\Delta(x)^2} \quad \text{for } x \neq x^*.$$

Therefore

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[R_T]}{\log T} \geq 2\sigma^2 \sum_{x\in\mathcal{X}} \Delta_c(x).$$

This proves the theorem.