A Minimalist Bayesian Framework for Stochastic Optimization

Kaizheng Wang[*]

This version: October 6, 2025

Abstract

The Bayesian paradigm offers principled tools for sequential decision-making under uncertainty, but its reliance on a probabilistic model for all parameters can hinder the incorporation of complex structural constraints. We introduce a minimalist Bayesian framework that places a prior only on the component of interest, such as the location of the optimum. Nuisance parameters are eliminated via profile likelihood, which naturally handles constraints. As a direct instantiation, we develop a MINimalist Thompson Sampling (MINTS) algorithm. Our framework accommodates structured problems, including continuum-armed Lipschitz bandits and dynamic pricing. It also provides a probabilistic lens on classical convex optimization algorithms such as the center of gravity and ellipsoid methods. We further analyze MINTS for multi-armed bandits and establish near-optimal regret guarantees.

Keywords: Stochastic optimization, Bayesian method, profile likelihood, regret analysis.

1 Introduction

Sequential decision-making under uncertainty is a ubiquitous challenge, where an agent repeatedly make decisions to optimize an unknown objective function based on limited, noisy feedback. Effective performance requires striking a delicate balance between exploiting actions that are believed to be optimal based on past data and exploring lesser-known actions to gather new information.

Several paradigms have emerged to address this exploration-exploitation tradeoff. The *explore-then-commit* (ETC) strategy is the most straightforward, dividing the time horizon into distinct exploration and exploitation phases. Yet, it can be difficult to decide an appropriate split. A more adaptive paradigm is the principle of *optimism in the face of uncertainty*, which constructs optimistic indices like upper confidence bounds to encourage exploration (Lai and Robbins, 1985; Auer et al., 2002). While theoretically sound, this approach often relies on problem-specific bonus calibration. A third one is the *Bayesian paradigm*, which treats the problem as a random instance and maintains a posterior distribution to quantify uncertainty (Thompson, 1933; Jones et al., 1998). The belief update powered by Bayes' rule provides a principled mechanism for automatically balancing exploration and exploitation. However, the standard Bayesian paradigm requires specifying a prior for all unknown parameters. This becomes a significant bottleneck when one wishes to encode rich structural knowledge, such as shape or smoothness constraints on the objective function, as it can be prohibitively difficult to design a tractable prior that is faithful to these constraints.

To resolve this dilemma, we introduce a minimalist Bayesian framework with significantly enhanced flexibility. Our approach allows the user to place a prior only on a low-dimensional component of interest, such as the location of the optimum. All other parameters are treated as nuisance

---

[*]Department of IEOR and Data Science Institute, Columbia University. Email: `kaizheng.wang@columbia.edu`.

and handled by the profile likelihood method (Barndorff-Nielsen and Cox, 1994). We can conveniently enforce structural constraints on them. The reduced-dimension prior and profile likelihood yield a generalized posterior for the component of interest, which then guides subsequent decisions.

Main contributions   Our contributions are threefold.

- (Framework) We develop a minimalist Bayesian framework for stochastic optimization that reasons about the parameter of interest without probabilistic modeling of all unknowns. The approach is lightweight and naturally accommodates structural constraints.

- (Algorithms and insights) We derive a MINimalist Thompson Sampling (MINTS) algorithm that directly updates and samples from the posterior of the optimum. We also instantiate the framework in complex structured problems including Lipschitz bandits and dynamic pricing, and provide novel probabilistic interpretations of classical convex optimization algorithms.

- (Theory) We analyze MINTS for multi-armed bandits and derive near-optimal regret bounds. The theoretical results rigorously justify the effectiveness of our new framework.

Related work   Our framework is closely related to Thompson sampling (TS) (Thompson, 1933), also known as posterior sampling or probability matching because the decision is drawn from the posterior distribution of the optimum (Scott, 2010; Russo et al., 2018). However, TS derives such posterior indirectly through a probabilistic model for the entire problem instance. For multi-armed bandits, this means placing a prior on the expected rewards of all arms. In contrast, MINTS directly specifies a prior on which arm is optimal. A related idea appears in online learning with full information, where one can place a prior on the optimum, exponentiate the empirical loss to form a "likelihood," and sample from a Gibbs posterior (Littlestone and Warmuth, 1989; Vovk, 1990; Cesa-Bianchi et al., 1997; Catoni, 2004; Bissiri et al., 2016). These methods require feedback for all decisions in each round, whereas we address the more challenging partial feedback setting.

Certain stochastic optimization problems permit Bayesian reasoning about the optimum without modeling the entire objective, such as probabilistic bisection search over an interval (Waeber et al., 2013). For general problems, several Bayesian algorithms choose actions to maximize information gain about the optimal solution, by increasing the mutual information between the next observation and the optimum (Villemonteix et al., 2009; Hennig and Schuler, 2012; Hernández-Lobato et al., 2014; Russo and Van Roy, 2018). These approaches still rely on full probabilistic models for belief updates. Closer in spirit to our work, Souza et al. (2021) combine a user-specified prior on the optimum with a standard Bayesian optimization model (e.g., Gaussian process) to form a pseudo-posterior, whereas we use profile likelihood to circumvent modeling of the nuisance parameters.

Finally, our framework provides a unified approach to Bayesian optimization under structural constraints, a domain that has largely been addressed case by case. In continuous black-box optimization, efforts have focused on designing Gaussian processes tailored to particular shape constraints (Swiler et al., 2020). In structured bandits, Van Parys and Golrezaei (2024) developed a computationally tractable frequentist algorithm with strong guarantees. It handles general convex constraints but requires the reward distributions of all arms to be supported on a common finite set. Bayesian counterparts with comparable flexibility and practicality are missing. Current approaches focus on specific settings, such as unimodal bandits on graphs (Paladino et al., 2017).

Outline   The rest of the paper is organized as follows. Section 2 lays out the preliminaries of stochastic optimization. Section 3 introduces the minimalist Bayesian framework and the MINTS algorithm. Section 4 instantiates the framework on canonical problems. Section 5 demonstrates its

efficacy through numerical experiments. Section 6 conducts a theoretical analysis of MINTS on multi-armed bandits. Section 7 concludes the paper with a discussion of future directions.

Notation   We use the symbol $[n]$ as a shorthand for $\{1, 2, \cdots, n\}$ and $|\cdot|$ to denote the absolute value of a real number or cardinality of a set. For nonnegative sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \le Cb_n$.

2 Preliminaries

This section lays out the preliminaries for our analysis. We begin by formally defining the stochastic optimization framework and presenting several canonical examples that illustrate its scope. Then, we review the standard Bayesian paradigm for these problems and highlight the key difficulties in incorporating structural knowledge, which motivates the new framework developed in this paper.

2.1 Stochastic optimization

In the standard setup of stochastic optimization, an agent seeks to maximize an unknown objective function f over a decision set $\mathcal{X}$:

$$\max_{x \in \mathcal{X}} f(x). \tag{2.1}$$

The agent learns about the optimum by sequentially interacting with the environment. Starting with an empty dataset $\mathcal{D}_0 = \varnothing$, at each period $t \in \mathbb{Z}_+$, the agent selects a decision $x_t \in \mathcal{X}$ based on past data $\mathcal{D}_{t-1}$, receives randomized feedback ϕ_t from the environment, and updates the dataset to $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(x_t, \phi_t)\}$. The performance over T time periods is typically measured by either the cumulative regret $\sum_{t=1}^{T}[\max_{x \in \mathcal{X}} f(x) - f(x_t)]$, which sums the suboptimality of each action, or the simple regret $\max_{x \in \mathcal{X}} f(x) - f(x_T)$, which measures the quality of the final action taken. Below we present several common examples.

Example 2.1 (Multi-armed bandit)**.** *The decision set is a collection of $K \in \mathbb{Z}_+$ arms, i.e. $\mathcal{X} = [K]$. Each arm x is associated with a reward distribution $\mathcal{P}_x$ over $\mathbb{R}$, and the objective value $f(x)$ is the expected reward $\mathbb{E}_{Y \sim \mathcal{P}_x} Y$. Given x_t and $\mathcal{D}_{t-1}$, the feedback ϕ_t is a sample from $\mathcal{P}_{x_t}$.*

Example 2.2 (Lipschitz bandit (Kleinberg et al., 2008))**.** *The set $\mathcal{X}$ is equipped with a metric d. Each decision x is associated with a reward distribution $\mathcal{P}_x$ over $\mathbb{R}$, and the objective value $f(x)$ is the expected reward $\mathbb{E}_{Y \sim \mathcal{P}_x} Y$. In addition, there exists a constant $M > 0$ such that $|f(x) - f(x')| \le M \cdot \mathsf{d}(x, x')$ holds for all $x, x' \in \mathcal{X}$. Given x_t and $\mathcal{D}_{t-1}$, the feedback ϕ_t is a sample from $\mathcal{P}_{x_t}$.*

Example 2.3 (Dynamic pricing (Den Boer, 2015))**.** *The set $\mathcal{X} \subseteq (0, +\infty)$ consists of feasible prices for a product. Any price x induces a demand distribution $\mathcal{P}_x$ over $[0, +\infty)$. The objective value $f(x)$ is the expected revenue $x \cdot \mathbb{E}_{D \sim \mathcal{P}_x} D$. Given x_t and $\mathcal{D}_{t-1}$, the feedback ϕ_t is a sample from $\mathcal{P}_{x_t}$.*

Example 2.4 (Continuous optimization)**.** *The set $\mathcal{X}$ is a subset of a Euclidean space. The objective function f belongs to a known class of continuous functions on $\mathcal{X}$, such as convex functions with Lipschitz gradients. Given x_t and $\mathcal{D}_{t-1}$, the feedback ϕ_t may include the function value $f(x_t)$, the gradient $\nabla f(x_t)$, the Hessian $\nabla^2 f(x_t)$, or their noisy versions.*

As can be seen from the examples, the historical data only reveals incomplete and noisy information about f. To make informed decisions under uncertainty, the agent needs to quantify and update beliefs over time. The Bayesian paradigm offers a coherent framework for this task. We now discuss this approach and the key challenges it faces.

2.2 Bayesian approaches and their challenges

Consider a family of stochastic optimization problems $\{P_\theta\}_{\theta \in \Theta}$, indexed by a parameter θ in a potentially infinite-dimensional space Θ. The parameter θ specifies all unknown components of a problem instance, such as the objective function and feedback distributions. Any dataset $\mathcal{D}$ defines a likelihood function $\mathcal{L}(\cdot; \mathcal{D})$ over Θ.

The Bayesian paradigm treats the problem (2.1) as a random instance P_θ whose parameter θ is drawn from a prior distribution $\mathcal{Q}_0$ over the space Θ (Thompson, 1933; Jones et al., 1998; Shahriari et al., 2015; Frazier, 2018). This prior encodes the agent's initial beliefs, such as smoothness or sparsity of the objective function, based on domain knowledge. After t rounds of interaction, the agent obtains data $\mathcal{D}_t$ and follows a two-step procedure:

1. (Belief update) Derive the posterior distribution $\mathcal{Q}_t$ given data $\mathcal{D}_t$ using Bayes' theorem:

$$\frac{\mathrm{d}\mathcal{Q}_t}{\mathrm{d}\mathcal{Q}_0}(\theta) = \frac{\mathcal{L}(\theta; \mathcal{D}_t)}{\int_\Theta \mathcal{L}(\theta'; \mathcal{D}_t)\mathcal{Q}_0(\mathrm{d}\theta')}, \qquad \theta \in \Theta. \tag{2.2}$$

This posterior captures the agent's refined understanding of the problem.

2. (Decision-making) Choose the next decision x_{t+1} by optimizing a criterion based on $\mathcal{Q}_t$. Popular approaches include expected improvement (Močkus, 1974; Jones et al., 1998), knowledge gradient (Frazier et al., 2009), Thompson sampling (Thompson, 1933), Bayesian upper confidence bounds (Kaufmann et al., 2012), and information-directed sampling (Russo and Van Roy, 2018).

While elegant, the Bayesian framework requires specifying a probabilistic model for the *entire* problem instance. This becomes a significant bottleneck when the problem involves rich structural knowledge, as encoding complex constraints through priors can be difficult. In addition, maintaining and sampling from a high-dimensional posterior is computationally expensive.

We illustrate these challenges using the dynamic pricing problem in Example 2.3. Assume binary demand for simplicity: for any price x, the demand ϕ follows a Bernoulli distribution with parameter $\theta_x \in [0, 1]$. The objective is then $f(x) = x\theta_x$. In the prototypical model, a buyer at time t has a latent valuation v_t drawn from a distribution ρ over $[0, \infty)$, independently of the history and the posted price x_t. The buyer makes a purchase if and only if $x_t \leq v_t$, resulting in $\phi_t = \mathbf{1}(x_t \leq v_t)$ and

$$\theta_x = \mathbb{P}(v_t \geq x) = \rho([x, +\infty)), \qquad \forall x \in \mathcal{X}. \tag{2.3}$$

It is easily seen that ρ, θ and f serve as different parametrizations of the same demand model. Working with the θ-parametrization, we can write the likelihood of data $\mathcal{D}_t$:

$$\mathcal{L}(\theta; \mathcal{D}_t) = \prod_{i=1}^{t} \theta_{x_i}^{\phi_i}(1 - \theta_{x_i})^{1-\phi_i}. \tag{2.4}$$

A Bayesian algorithm would combine this likelihood with a prior $\mathcal{Q}_0$ on θ to obtain the posterior $\mathcal{Q}_t$ and then select the next price x_{t+1}. This is tractable for simple parametric classes (McLennan, 1984; Farias and Van Roy, 2010; Harrison et al., 2012). On the other hand, nonparametric models offer greater flexibility but run into the obstacle of structural constraints. The relation (2.3) immediately implies that θ is non-increasing. Furthermore, if ρ has a density bounded by $M > 0$, then θ is M-Lipschitz. Below we show in two cases the challenges arising from these constraints.

Case 1. Finite feasible set Suppose that $\mathcal{X}$ only consists of K prices $p_1 < \cdots < p_K$. Then, the function θ is represented by a vector $\boldsymbol{\theta} = (\theta_{p_1}, \cdots, \theta_{p_K})^\top$. The structural constraints confine $\boldsymbol{\theta}$ to the following convex set:

$$\{\boldsymbol{\theta} \in \mathbb{R}^K : \ 0 \leq \theta_K \leq \cdots \leq \theta_1 \leq 1 \text{ and } \theta_j - \theta_{j+1} \leq M(p_{j+1} - p_j), \ \forall j \in [K]\}. \tag{2.5}$$

It is unclear how to design a prior over this set that leads to a tractable posterior. The coupling between the parameters rules out simple product distributions, rendering standard Thompson sampling for Bernoulli bandits (Thompson, 1933) inapplicable. Even without the Lipschitz constraint, dealing with the monotonicity requires efforts. For instance, a simple reparametrization maps the monotonicity constraint to the probability simplex, where a Dirichlet prior can be used (Cope, 2007). However, this approach still requires posterior approximation and does not easily accommodate additional constraints like the Lipschitz condition.

Case 2. Continuous feasible set Suppose that $\mathcal{X}$ is an interval $[p_{\min}, p_{\max}]$. The structural constraints now define a function class:

$$\left\{h: \ [p_{\min}, p_{\max}] \to [0,1] \ \middle| \ -M \leq h'(x) \leq 0, \ \forall x \in [u,v]\right\}. \tag{2.6}$$

Specifying a tractable prior over this class is not straightforward. Conventional Gaussian processes for Bayesian optimization assign zero probability to a bounded class. Shape-constrained versions are only designed for certain structures (Swiler et al., 2020).

As the example shows, even simple structural constraints such as monotonicity and Lipschitz continuity pose significant challenges to the standard Bayesian paradigm. We will now introduce a more flexible framework to better incorporate prior knowledge.

3 A minimalist Bayesian framework

We develop a minimalist Bayesian framework that only specifies a prior for the key component of interest rather than the entire problem instance. This lightweight approach can be easily integrated with structural constraints. We start by modeling the optimum alone, illustrating the idea with a canonical example.

Example 3.1 (Multi-armed bandit with Gaussian rewards). *Let $K \in \mathbb{Z}_+$, $\Theta = \mathbb{R}^K$, and $\sigma > 0$. For any $\boldsymbol{\theta} \in \Theta$, denote by $P_{\boldsymbol{\theta}}$ the multi-armed bandit problem in Example 2.1 with reward distribution $\mathcal{P}_j = N(\theta_j, \sigma^2), \ \forall j \in [K]$. The likelihood function for a dataset $\mathcal{D}_t$ is*

$$\mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_t) = \prod_{i=1}^{t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\theta_{x_i} - \phi_i)^2}{2\sigma^2}\right). \tag{3.1}$$

We use a prior distribution $\mathcal{Q}_0$ over the decision space $[K]$ to represent our initial belief about which arm is optimal. The statement "Arm j is optimal" corresponds to the composite hypothesis $H_j: \ \boldsymbol{\theta} \in \Theta_j$, where

$$\Theta_j = \{\boldsymbol{v} \in \Theta : \ v_j \geq v_k, \ \forall k \in [K]\}. \tag{3.2}$$

The likelihood function $\mathcal{L}$ is defined for a point $\boldsymbol{\theta}$ rather than a set like Θ_j. To quantify the evidence for the composite hypothesis H_j, we turn to the profile likelihood method (Barndorff-Nielsen

and Cox, 1994). Define the profile likelihood of Arm j as the maximum likelihood achievable by any parameter vector consistent with H_j:

$$\bar{\mathcal{L}}(j; \mathcal{D}_t) = \max_{\boldsymbol{v} \in \Theta_j} \mathcal{L}(\boldsymbol{v}; \mathcal{D}_t). \tag{3.3}$$

This is equivalent to performing constrained maximum likelihood estimation of $\boldsymbol{\theta}$ over the set Θ_j. Finally, we mimic the Bayes' rule to derive a (generalized) posterior $\mathcal{Q}_t$ from the prior $\mathcal{Q}_0$ and the profile likelihood $\bar{\mathcal{L}}$:

$$\mathcal{Q}_t(j) = \frac{\bar{\mathcal{L}}(j; \mathcal{D}_t)\mathcal{Q}_0(j)}{\sum_{k=1}^{K} \bar{\mathcal{L}}(k; \mathcal{D}_t)\mathcal{Q}_0(k)}, \qquad j \in [K]. \tag{3.4}$$

It represents our updated belief about which arm is optimal, having integrated the evidence from data.

This approach is efficient, flexible, and general, as highlighted by the following remarks.

Remark 1 (Computational efficiency)**.** The posterior update is computationally tractable. Let $I_j = \{i \in [t] : x_i = j\}$ be the set of pulls for Arm j, and $\hat{\mu}_j = |I_j|^{-1} \sum_{i \in I_j} \phi_i$ be its empirical mean. The negative log-likelihood is a weighted sum-of-squares:

$$-\log \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_t) = \frac{1}{2\sigma^2} \sum_{j=1}^{K} |I_j|(\theta_j - \hat{\mu}_j)^2 + C,$$

where C is a constant. Denote by $L(\boldsymbol{\theta})$ the first term on the right-hand side. We have

$$\log \bar{\mathcal{L}}(j; \mathcal{D}_t) = -\min_{\boldsymbol{\theta} \in \Theta_j} L(\boldsymbol{\theta}) - C.$$

This minimization is a simple quadratic program over a convex polytope, which can be solved efficiently. The generalized posterior $\mathcal{Q}_t$ is then readily computed from these minimum values:

$$\mathcal{Q}_t(j) = \frac{e^{-\min_{\boldsymbol{\theta} \in \Theta_j} L(\boldsymbol{\theta})} \mathcal{Q}_0(j)}{\sum_{k=1}^{K} e^{-\min_{\boldsymbol{\theta} \in \Theta_k} L(\boldsymbol{\theta})} \mathcal{Q}_0(k)}.$$

When $K = 2$, an analytical expression is available. Suppose that $I_1, I_2 \neq \varnothing$ and $\hat{\mu}_1 \geq \hat{\mu}_2$. Let

$$\alpha = \frac{1}{2\sigma^2} \cdot \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{1/|I_1| + 1/|I_2|}.$$

We have

$$\frac{\mathcal{Q}_t(1)}{\mathcal{Q}_t(2)} = e^{\alpha} \frac{\mathcal{Q}_0(1)}{\mathcal{Q}_0(2)} \qquad \text{and} \qquad \mathcal{Q}_t(1) = 1 - \mathcal{Q}_t(2) = \frac{e^{\alpha} \mathcal{Q}_0(1)}{e^{\alpha} \mathcal{Q}_0(1) + \mathcal{Q}_0(2)}.$$

Remark 2 (Translation invariance)**.** The above analysis reveals the translation invariance of MINTS with Gaussian reward models: if all the reward distributions are simultaneously shifted by a constant, the generalized posterior for the optimal arm remains unchanged. In contrast, standard Bayesian modeling with a prior for all the mean rewards cannot have such property.

Remark 3 (Structured bandits)**.** Structural constraints on the parameter $\boldsymbol{\theta}$ can be seamlessly incorporated by restricting the parameter space Θ. For instance, adding the Lipschitz condition in Example 2.2 leads to

$$\Theta = \{\boldsymbol{v} \in \mathbb{R}^K : |v_i - v_j| \leq M \cdot \mathsf{d}(i, j), \ \forall i, j \in [K]\}. \tag{3.5}$$

The inference procedure outlined in (3.2), (3.3) and (3.4) remains exactly the same.

Remark 4 (Other reward distributions). The Gaussian assumption is for illustration. This procedure applies to any reward distribution with a tractable likelihood function, such as Bernoulli or other members of the exponential family.

The core logic of Example 3.1 can be abstracted into a general belief-updating algorithm for the location of the optimum. We formalize this in Algorithm 1.

Algorithm 1 Minimalist Bayesian inference for the optimum

Input: A family of stochastic optimization problems $\{P_\theta\}_{\theta \in \Theta}$ with decision space $\mathcal{X}$ and likelihood function $\mathcal{L}$. A prior distribution $\mathcal{Q}_0$ over $\mathcal{X}$. A dataset $\mathcal{D} = \{(x_i, \phi_i)\}_{i=1}^t$.

Step 1. Construct the profile likelihood function

$$\bar{\mathcal{L}}(x; \mathcal{D}) = \sup \left\{ \mathcal{L}(\theta; \mathcal{D}) : \ \theta \in \Theta \text{ and } f_\theta(x) = \max_{x' \in \mathcal{X}} f_\theta(x') \right\}, \qquad x \in \mathcal{X}, \tag{3.6}$$

where f_θ denotes the objective function in the problem instance P_θ.

Step 2. Derive a generalized posterior distribution $\mathcal{Q}_t$ by reweighting the prior $\mathcal{Q}_0$:

$$\frac{\mathrm{d}\mathcal{Q}_t}{\mathrm{d}\mathcal{Q}_0}(x) = \frac{\bar{\mathcal{L}}(x; \mathcal{D}_t)}{\int_{\mathcal{X}} \bar{\mathcal{L}}(x'; \mathcal{D}_t) \mathcal{Q}_0(\mathrm{d}x')}, \qquad x \in \mathcal{X}. \tag{3.7}$$

Output: $\mathcal{Q}_t$.

Algorithm 1 provides a modular inference engine. It can be paired with any decision-making rule that operates on a belief distribution over the optimal action. One natural choice is a minimalist version of Thompson Sampling, presented in Algorithm 2. It is conceptually simpler than standard Thompson Sampling, which starts with a prior for the full parameter θ and then draws a decision from the posterior of the optimum in each iterate. Our approach only requires a prior for the optimum.

Algorithm 2 MINimalist Thompson Sampling (MINTS)

Input: A family of stochastic optimization problems $\{P_\theta\}_{\theta \in \Theta}$ with decision space $\mathcal{X}$ and likelihood function $\mathcal{L}$. A prior distribution $\mathcal{Q}_0$ over $\mathcal{X}$.

Let $\mathcal{D}_0 = \varnothing$.

For $t = 1, 2, \cdots$**:**

    Sample a decision x_t from $\mathcal{Q}_{t-1}$ and receive feedback ϕ_t.

    Run Algorithm 1 on the updated dataset $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(x_t, \phi_t)\}$ to get $\mathcal{Q}_t$.

Output: A sequence of decisions $\{x_t\}_{t=1}^\infty$.

In addition to the optimum, we may also have unknown structural parameters to deal with, such as the noise level σ in Example 3.1 or the Lipschitz constant M in Example 2.2. We can use a prior distribution to jointly model the optimum and those parameters. In general, let γ represent the key components that we wish to model, and Γ be the space it lives in. For instance, when γ encodes the optimum and the Lipshitz constant M, we have $\Gamma = \mathcal{X} \times [0, +\infty)$. Denote by Θ_γ the collection of θ's whose associated instance P_θ has key components γ. The full parameter space Θ is covered by the subsets $\{\Theta_\gamma\}_{\gamma \in \Gamma}$. Let $\mathcal{Q}_0$ be a prior distribution for the key components. Given data $\mathcal{D}_t$, the profile likelihood is

$$\bar{\mathcal{L}}(\gamma; \mathcal{D}_t) = \sup_{\theta \in \Theta_\gamma} \mathcal{L}(\theta; \mathcal{D}_t).$$

Then, we obtain the generalized posterior distribution $\mathcal{Q}_t$ through

$$\frac{\mathrm{d}\mathcal{Q}_t}{\mathrm{d}\mathcal{Q}_0}(\gamma) = \frac{\bar{\mathcal{L}}(\gamma; \mathcal{D}_t)}{\int_{\mathcal{X}} \bar{\mathcal{L}}(\gamma'; \mathcal{D}_t)\mathcal{Q}_0(\mathrm{d}\gamma')}, \qquad \gamma \in \Gamma.$$

Algorithm 1 is a special case where the key component γ is just the optimum. In the other extreme, when the key component is the whole parameter θ, each Θ_γ becomes a singleton, and we obtain the standard Bayesian procedure.

4 Examples and new insights

In this section, we demonstrate the versatility of the minimalist Bayesian framework by applying it to several fundamental problems. We show that it not only handles complex settings like continuum-armed Lipschitz bandits but also provides novel probabilistic interpretations of classical algorithms in convex optimization.

4.1 Continuum-armed Lipschitz bandit

Consider a continuum-armed Lipschitz bandit whose decision space $\mathcal{X}$ is the d-dimensional unit cube $[0,1]^d$. The parameter space Θ consists of all M-Lipschitz functions on $\mathcal{X}$, where $M > 0$ is a known constant.

Noiseless rewards First, assume the feedback ϕ_t is the exact function value, $f(x_t)$. In this setting, the likelihood $\mathcal{L}(h; \mathcal{D}_t)$ is binary: it is 1 if h is consistent with all observations, i.e. $h(x_i) = \phi_i$ for all $i \in [t]$. Consequently, the profile likelihood $\bar{\mathcal{L}}(x; \mathcal{D}_t)$ for an action x being optimal is also binary: it is 1 if there exists at least one M-Lipschitz function that interpolates the data and attains its maximum at x. The following lemma shows that the existence check is a convex feasibility problem. See Appendix A.1 for the proof.

Lemma 4.1. *The profile likelihood* $\bar{\mathcal{L}}(x; \mathcal{D}_t)$ *is* 1 *if the convex polytope*

$$S(x) = \{(v, v_1, \cdots, v_t): \ v_i \leq v \leq v_i + M\|x - x_i\|_2, \ \forall i \ \text{and} \ |v_i - v_j| \leq M\|x_i - x_j\|_2, \ \forall i,j\}$$

is non-empty, and 0 *otherwise.*

Based on that, we have

$$\frac{\mathrm{d}\mathcal{Q}_t}{\mathrm{d}\mathcal{Q}_0}(x) = \frac{\mathbf{1}(S(x) \neq \varnothing)}{\int_{\mathcal{X}} \mathbf{1}(S(x') \neq \varnothing)\mathcal{Q}_0(\mathrm{d}x')}.$$

The result leads to a rejection sampling algorithm for sampling from the generalized posterior $\mathcal{Q}_t$: draw a candidate optimum x' from $\mathcal{Q}_0$, and accept it if $S(x') \neq \varnothing$.

Gaussian rewards Suppose that the feedback ϕ_t is the function value at x_t contaminated by random noise from $N(0, \sigma^2)$ with known $\sigma > 0$. The likelihood and profile likelihood are

$$\mathcal{L}(h; \mathcal{D}_t) = \prod_{i=1}^{t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[h(x_i) - \phi_i]^2}{2\sigma^2}\right),$$

$$\bar{\mathcal{L}}(x; \mathcal{D}_t) = \sup\left\{\mathcal{L}(h; \mathcal{D}_t)\Big| \ h \in \Theta \ \text{and} \ h(x) = \max_{x' \in [0,1]^d} h(x')\right\}.$$

The following lemma shows that for any $x \in \mathcal{X}$, $\bar{\mathcal{L}}(x; \mathcal{D}_t)$ can be computed up to an additive constant by solving a finite-dimensional convex program. The proof is deferred to Appendix A.2.

Lemma 4.2. *For any $x \in [0,1]^d$, let $S(x)$ be the set defined in Lemma 4.1, and $V(x)$ be the optimal value of the following convex program:*

$$\min_{(v,v_1,\cdots,v_t)\in S(x)} \left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{t} (v_i - \phi_i)^2 \right\}. \tag{4.1}$$

We have $-\log \bar{\mathcal{L}}(x;\mathcal{D}_t) = V(x) + C$ for a constant C.

Define $V_{\min} = \min_{x \in \mathcal{X}} V(x)$, which can be computed by solving (4.1) without the constraints $v_i \le v \le v_i + M\|x - x_i\|_2$ for $i \in [t]$. The facts $V_{\min} - V(x) \le 0$ and

$$\frac{\mathrm{d}\mathcal{Q}_t}{\mathrm{d}\mathcal{Q}_0}(x) = \frac{e^{V_{\min}-V(x)}}{\int_{\mathcal{X}} e^{V_{\min}-V(x')}\mathcal{Q}_0(\mathrm{d}x')}$$

lead to a rejection sampling scheme for drawing from $\mathcal{Q}_t$: sample x' from $\mathcal{Q}_0$ and accept it with probability $e^{V_{\min}-V(x')}$.

4.2 Dynamic pricing

Next, we revisit the dynamic pricing problem discussed in Section 2.2. The demand is binary, the full parameter θ encodes the purchase probability at each price, and the likelihood function is defined in (2.4).

Suppose the decision space $\mathcal{X}$ consists of K feasible prices $p_1 < \cdots < p_K$. The parameter space Θ is given by (2.5). Let $I_j = \{i \in [t] : x_i = p_j\}$ record the times when price p_j was chosen, $\hat{\mu}_j = |I_j|^{-1} \sum_{i \in I_j} \phi_i$ be the empirical mean demand, and Θ_j be defined as in (3.2). Note that Θ_j is a convex polytope and the log-likelihood is a concave function:

$$\log \mathcal{L}(\boldsymbol{\theta};\mathcal{D}_t) = \sum_{j=1}^{K} |I_j| \cdot [\hat{\mu}_j \log \theta_j + (1 - \hat{\mu}_j)\log(1 - \theta_j)]. \tag{4.2}$$

Then, the profile likelihood $\bar{\mathcal{L}}(p_j, \mathcal{D}_t) = \sup_{\theta \in \Theta_j} \mathcal{L}(\boldsymbol{\theta};\mathcal{D}_t)$ can be computed through convex optimization, which further yields the generalized posterior $\mathcal{Q}_t$.

We can add more structural constraints based on domain knowledge. As long as the parameter space Θ remains convex, the computation of $\mathcal{Q}_t$ is tractable. In contrast, the constraints could render standard Bayesian approaches difficult.

4.3 First-order convex optimization

Perhaps surprisingly, our framework provides probabilistic interpretations of classical cutting-plane methods in first-order convex optimization (Nesterov, 2018, Section 3.2.8). Consider minimizing an unknown convex function f over a convex body $\mathcal{X} \subseteq \mathbb{R}^d$, where at each query point x_t, we receive a subgradient $\phi_t \in \partial f(x_t)$. We first offer a minimalist Bayesian perspective for the *center of gravity method*.

Example 4.1 (Center of gravity method)**.** *Denote by $\mathcal{Q}_0$ the uniform distribution over $\mathcal{X}$, which is a natural prior for the unknown optimum. Suppose that for any $t \in \mathbb{N}$, we apply Algorithm 1 to the dataset $\mathcal{D}_t = \{(x_i, \phi_i)\}_{i=1}^{t}$ to derive a generalized posterior $\mathcal{Q}_t$, and use its mean as the next decision x_{t+1}. We have the following results. See Appendix A.3 for their proof.*

1. *$\mathcal{Q}_t$ is the uniform distribution over the set $\{x \in \mathcal{X} : \phi_i^\top(x - x_i) \le 0, \ \forall i \in [t]\}$, which is the intersection of $\mathcal{X}$ and t hyperplanes;*

2. *The procedure is equivalent to the center of gravity method.*

Therefore, the center of gravity method can be viewed as a (minimalist) Bayesian approach that uses the (generalized) posterior mean to propose the next query in each iterate. Each query yields a hyperplane that reduces the current search space to one side of it. When t is large, $\mathcal{Q}_t$ becomes complex, making it hard to compute the mean. Indeed, the center of gravity method is known to be computationally expensive.

Meanwhile, Bayesian statistics offers two powerful approaches for handling sophisticated posteriors. One is Markov chain Monte Carlo (MCMC), which designs a Markov chain with the posterior as the limiting distribution; the other is variational Bayes, which approximates the posterior using a family of simple distributions that are easy to sample from. Interestingly, both ideas lead to well-known optimization algorithms developed as tractable approximations of the center of gravity method. Bertsimas and Vempala (2004) designed an MCMC algorithm for estimating the center of gravity based on a random walk. On the other hand, below we give a variational Bayes interpretation of the celebrated *ellipsoid method*.

Example 4.2 (Ellipsoid method)**.** *For any vector $\boldsymbol{c} \in \mathbb{R}^d$ and positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, define an ellipsoid $E(\boldsymbol{c}, \boldsymbol{A}) = \{\boldsymbol{v} \in \mathbb{R}^d : (\boldsymbol{v} - \boldsymbol{c})^\top \boldsymbol{A}(\boldsymbol{v} - \boldsymbol{c}) \leq 1\}$ and denote by $\mathcal{Q}(\boldsymbol{c}, \boldsymbol{A})$ be the uniform distribution over that. Let Φ be the parametric distribution family $\{\mathcal{Q}(\boldsymbol{c}, \boldsymbol{A})\}_{\boldsymbol{c} \in \mathbb{R}^d, \boldsymbol{A} \succ 0}$.*

Suppose that $\mathcal{X}$ is an ellipsoid $E(\boldsymbol{c}_0, \boldsymbol{A}_0)$, and let the associated uniform distribution $\mathcal{Q}_0 = \mathcal{Q}(\boldsymbol{c}_0, \boldsymbol{A}_0)$ be our prior for the optimum. At any time $t \in \mathbb{N}$, our belief is characterized by a distribution $\mathcal{Q}_t = \mathcal{Q}(\boldsymbol{c}_t, \boldsymbol{A}_t) \in \Phi$. Consider the following updating rule motivated by the assumed density filtering method (Minka, 2001):

- *use the mean of $\mathcal{Q}_t$ as the next decision (i.e. $x_{t+1} = \boldsymbol{c}_t$) and receive feedback $\phi_{t+1} \in \partial f(x_{t+1})$;*

- *run Algorithm 1 with prior $\mathcal{Q}_t$ and data $\{(x_{t+1}, \phi_{t+1})\}$ to obtain a generalized posterior $\bar{\mathcal{Q}}_{t+1}$;*

- *find its best approximation in Φ that minimizes the forward Kullback-Leibler divergence:*

$$\mathcal{Q}_{t+1} \in \underset{\mathcal{Q} \in \Phi}{\operatorname{argmin}} D_{\mathrm{KL}}(\bar{\mathcal{Q}}_{t+1} \| \mathcal{Q}).$$

The procedure generates a sequence of uniform distributions over ellipsoids to capture our belief about the optimum. It is equivalent to the ellipsoid method in convex optimization, which has closed-form updates. See Appendix A.4 for a proof.

5   Numerical experiments

Having established the minimalist Bayesian framework, we now demonstrate its efficacy through numerical experiments. The Python code for reproducing the results is available at https://github.com/kw2934/MINTS.

Our first example is dynamic pricing with binary demand, as discussed in Example 2.3, Section 2.2 and Section 4.2. The decision set $\mathcal{X}$ is $\{j/20 : j \in [19]\}$, with $K = 19$ candidate prices from 0.05 to 0.95. The latent valuation distribution ρ is the uniform distribution over the unit interval $[0, 1]$. This determines the expected demand $\theta_x = 1 - x$ and expected revenue $f(x) = x(1 - x)$. The objective function is maximized at $x^* = 1/2$. After t rounds, our data $\mathcal{D}_t$ consists of chosen prices $\{x_i\}_{i=1}^t$ and realized demands $\{\phi_i\}_{i=1}^t$. We test two versions of MINTS (Algorithm 2) and two multi-armed bandit algorithms over $T = 5000$ rounds:
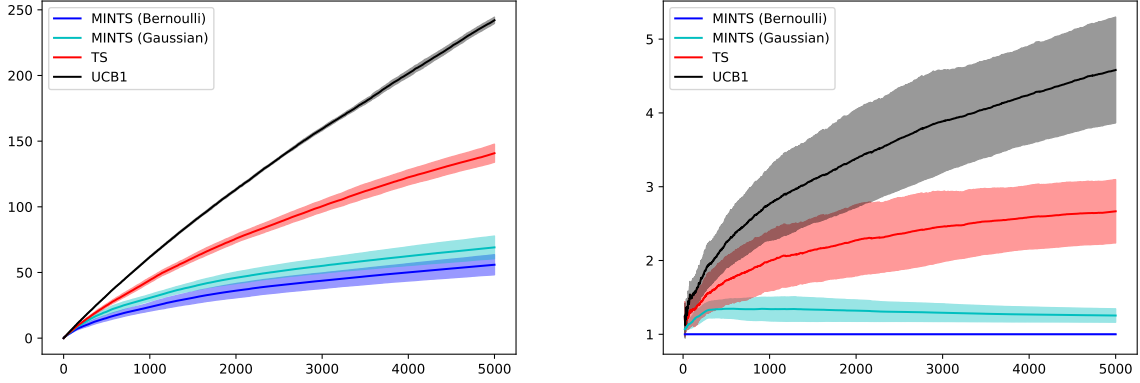
Figure 1: Dynamic pricing experiments. x-axis: time. y-axis: regret (left panel) and relative regret (right panel). Blue: MINTS with Bernoulli likelihood. Cyan: MINTS with Gaussian likelihood. Red: Thompson sampling. Black: UCB1.

1. MINTS with Bernoulli likelihood (2.4), which uses the correct demand model. The parameter space is the one in (2.5) with $M = 1$.

2. MINTS with Gaussian likelihood (3.1), which incorrectly models the conditional demand distribution given price x as $N(\theta_x, \sigma^2)$. We set $\sigma = 1/2$ because the sub-Gaussian variance proxy of the binary demand is bounded by $1/2$ (Hoeffding, 1994). The parameter space is the one in (2.5) with $M = 1$.

3. Thompson sampling with Gaussian likelihood and prior (Russo et al., 2018). It treats the pricing problem as a multi-armed bandit and ignores structural constraints. In the t-th round, the reward is the revenue $x_t \phi_t$ rather than the demand ϕ_t. The conditional reward (revenue) distribution given the j-th candidate price is modeled as $N(\mu_j, \sigma^2)$ with $\sigma = 1/2$. The mean parameters $\{\mu_j\}_{j=1}^K$ are assumed to be independently drawn from a prior distribution $N(0, 1)$.

4. The UCB1 algorithm (Auer et al., 2002). Similar to the above, this is also a multi-armed bandit algorithm that directly works with revenues. It constructs optimistic estimates of conditional mean rewards using upper confidence bounds, and chooses the decision with the highest estimate.

We run 100 independent simulations and in each of them, we compute the cumulative regret $\sum_{i=1}^{t}[f(x^*) - f(x_i)]$ over $t \in [T]$ for every algorithm. The results are summarized in Figure 1. The left panel shows the regrets of different algorithms. The right panel plots the relative regrets normalized by that of MINTS with Bernoulli likelihood (skipping the first 20 rounds). In both panels, each solid curve corresponds to the empirical mean value, and the thin shaded areas around that give 95% confidence bounds on the expected values (empirical mean $\pm$ 1.96 times the standard error). The two MINTS algorithms significantly outperform bandit algorithms, and the relative advantage grows with time. This highlights the benefit of incorporating structural knowledge. We also note that using a misspecified demand model only has minor impact on the performance of MINTS.

Our second numerical example concerns unstructured multi-armed bandit with Gaussian rewards (Example 3.1). There are $K = 5$ arms, and the reward distribution of Arm j is $N(\theta_j, 1)$ with $\theta_j = j/5$. We compare the following algorithms over $T = 5000$ rounds:
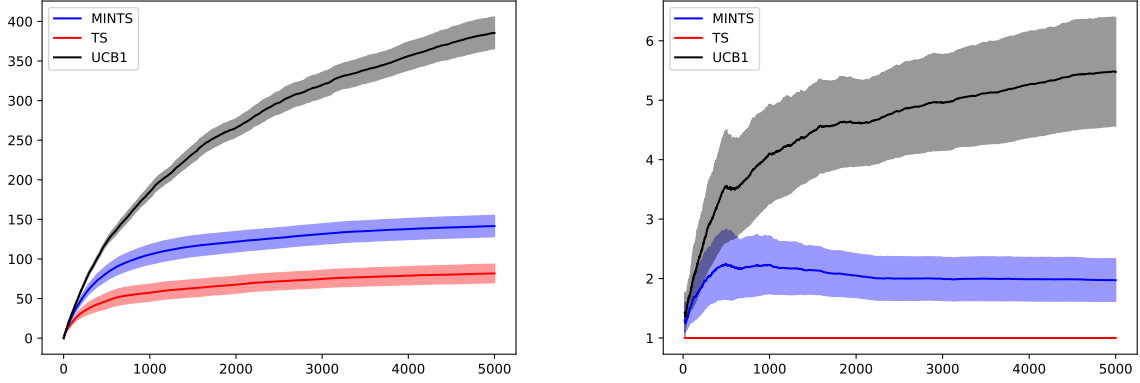
Figure 2: Multi-armed bandit experiments. x-axis: time. y-axis: regret (left panel) and relative regret (right panel). Blue: MINTS. Red: Thompson sampling. Black: UCB1.

1. MINTS with Gaussian likelihood (3.1) and $\sigma = 1$. The parameter space is $\mathbb{R}^K$.

2. Thompson sampling with Gaussian likelihood and prior (Russo et al., 2018). The prior distribution is $N(0, 1)$.

3. The UCB1 algorithm (Auer et al., 2002).

Figure 2 summarizes the results over 100 independent simulations, showing the absolute and relative (compared to Thompson sampling) regrets of all algorithms. We see that MINTS outperforms UCB1 and is comparable to Thompson sampling by a factor of around 2. The latter can be understood as the price of minimalist modeling. Targeting the main parameter while profiling out the nuisance component inevitably loses information. Fortunately, the cost of flexibility is not prohibitive and can be compensated by major gains in structured problems.

6 Theoretical analysis for multi-armed bandits

In this section, we will analyze MINTS (Algorithm 2) for the multi-armed bandit problem in Example 2.1 and derive near-optimal guarantees. We use μ_j to refer to the expected reward of Arm j, and measure the performance through the *regret* defined below.

Definition 6.1 (Regret). *For any $T \in \mathbb{Z}_+$, the regret of a decision sequence $\{x_t\}_{t=1}^T$ is*

$$\mathcal{R}(T) = \sum_{t=1}^{T} \Big(\max_{j \in [K]} \mu_j - \mu_{x_t} \Big).$$

To implement MINTS, we model each reward distribution $\mathcal{P}_j$ as a Gaussian distribution $N(\mu_j, \sigma^2)$ with unknown mean μ_j and known standard deviation $\sigma > 0$. Hence, the bandit problem is modeled as a parametrized instance $P_{\boldsymbol{\mu}}$ in Example 3.1 with unknown $\boldsymbol{\mu} \in \mathbb{R}^K$, and the likelihood function $\mathcal{L}$ is given by (3.1). The parametric model is merely a tool for algorithm design; our theoretical guarantees will not be restricted to Gaussian rewards. As we will show shortly, the algorithm performs well for reward distributions satisfying the following light tail condition.

Assumption 6.1 (Sub-Gaussian reward). *The reward distributions* $\{\mathcal{P}_j\}_{j=1}^K$ *are 1-sub-Gaussian:*

$$\mathbb{E}_{y \sim \mathcal{P}_j} e^{\lambda(y-\mu_j)} \le e^{\lambda^2/2}, \qquad \forall \lambda \in \mathbb{R}.$$

Assumption 6.1 is standard for bandit studies (Lattimore and Szepesvári, 2020). It holds for many common distributions with sufficiently fast tail decay, including any Gaussian distribution with variance bounded by 1, or distributions supported on an interval of width 2 (Hoeffding, 1994). For sub-Gaussian distributions with general variance proxies, we can reduce to this case by rescaling.

We present a regret bound with explicit dependence on the horizon T, number of arms K, and the sub-optimality gaps of arms. The proof is deferred to Appendix B.2.

Theorem 6.1 (Regret bound). *For the multi-armed bandit in Example 2.1, run MINTS (Algorithm 2) with a uniform prior over the K arms and the Gaussian likelihood (3.1) with $\sigma > 1$. Define $\Delta_j = \max_{k \in [K]} \mu_k - \mu_j$ for $j \in [K]$. Under Assumption 6.1, there exists a constant C determined by σ such that*

$$\mathbb{E}[\mathcal{R}(T)] \le C \left(\min \left\{ \sum_{j:\ \Delta_j > 0} \frac{\log T}{\Delta_j}, \sqrt{KT \log K} \right\} + \sum_{j=1}^K \Delta_j \right).$$

This result demonstrates the near-optimality of MINTS. The sum $\sum_{j=1}^K \Delta_j$ stems from the fact that each arm must be pulled at least once. When the problem instance is fixed and T is sufficiently large, the problem-dependent bound $\sum_{j:\ \Delta_j > 0} \Delta_j^{-1} \log T$ matches the lower bound for Gaussian bandits in Garivier et al. (2019) up to a constant factor. For any fixed T, the problem-independent bound $\sqrt{TK \log K}$ matches the regret bound for Thompson sampling using Gaussian priors and likelihood (Agrawal and Goyal, 2017), achieving the minimax lower bound in Bubeck and Cesa-Bianchi (2012) up to a $\sqrt{\log K}$ factor.

Theorem 6.1 is a corollary of the more refined result below. See Appendix B.1 for the proof.

Theorem 6.2 (Regret bound). *Under the setup in Theorem 6.1, there exists a constant C determined by σ such that*

$$\mathbb{E}[\mathcal{R}(T)] \le C \inf_{\delta \ge 0} \left\{ \sum_{j:\ \Delta_j > \delta} \left(\frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j} + \Delta_j \right) + T \max_{j:\ \Delta_j \le \delta} \Delta_j \right\}.$$

The regret bound has the same order as that in Auer and Ortner (2010), achieved by a carefully designed upper confidence bound algorithm with arm elimination. Our algorithm is simpler.

7 Discussion

We introduced a minimalist Bayesian framework for stochastic optimization that only requires a prior for the component of interest and handles nuisance parameters via profile likelihood. The lightweight modeling makes it easy to incorporate structural constraints on problem parameters, opening several promising avenues for future research. First, designing scalable algorithms for sampling from the generalized posterior is critical for handling continuous or high-dimensional spaces. Second, developing more sophisticated acquisition rules beyond simple posterior sampling could further improve performance. Beyond these refinements, extending the minimalist principle to contextual bandits and reinforcement learning presents an exciting frontier. Finally, a crucial theoretical task will be to accompany these new algorithms with rigorous guarantees.

Acknowledgement

A Proofs of Section 4

A.1 Proof of Lemma 4.1

Choose any x such that $\bar{\mathcal{L}}(x; \mathcal{D}_t) = 1$. By definition, there exists $h \in \Theta$ such that $h(x) = \max_{x' \in \mathcal{X}} h(x')$ and $h(x_i) = \phi_i$ for all $i \in [t]$. It is easily seen that $(h(x), h(x_1), \cdots, h(x_t)) \in S(x)$ and thus $S(x) \neq \varnothing$.

Now, suppose that x makes $S(x) \neq \varnothing$. We need to show that $\bar{\mathcal{L}}(x; \mathcal{D}_t) = 1$. Choose any $(v, v_1, \cdots, v_t) \in S(x)$. The Kirszbraun theorem (Kirszbraun, 1934) guarantees the existence of a function $g : \mathcal{X} \to \mathbb{R}$ that is M-Lipschitz and satisfies $g(x) = v$, $g(x_1) = v_1$, $\cdots$, $g(x_t) = v_t$. Define $h(\cdot) = \max\{g(\cdot), v\}$. Then, h remains M-Lipschitz. It is maximized at x and agrees with g on $x, x_1, \cdots, x_t$. Therefore, $\bar{\mathcal{L}}(x; \mathcal{D}_t) = 1$.

A.2 Proof of Lemma 4.2

Define $L(v_1, \cdots, v_n) = \frac{1}{2\sigma^2} \sum_{i=1}^{t} (v_i - \phi_i)^2$. It is easily seen that

$$- \log \mathcal{L}(h; \mathcal{D}_t) = L(h(x_1), \cdots, h(x_t)) + C, \qquad \forall h \in \Theta$$

holds with a constant C. For any $\varepsilon > 0$, there exists $g \in \Theta$ such that $g(x) = \max_{x' \in \mathcal{X}} g(x')$, $g(x_i) = \phi_i$ for all $i \in [t]$, and

$$- \log \mathcal{L}(g; \mathcal{D}_t) \leq - \log \bar{\mathcal{L}}(x; \mathcal{D}_t) + \varepsilon.$$

Note that $(g(x), g(x_1), \cdots, g(x_t))$ is feasible for the program (4.1). Hence,

$$V(x) \leq L(g(x_1), \cdots, g(x_t)) = - \log \mathcal{L}(g; \mathcal{D}_t) - C \leq - \log \bar{\mathcal{L}}(x; \mathcal{D}_t) - C + \varepsilon.$$

We get $- \log \bar{\mathcal{L}}(x; \mathcal{D}_t) \geq V(x) + C$.

It remains to prove the other direction, choose any optimal solution $(v, v_1, \cdots, v_t)$ to (4.1). Similar to the proof of Lemma 4.1, we can construct $h \in \Theta$ that is maximized at x and satisfies $h(x) = v$, $h(x_1) = v_1$, $\cdots$, $h(x_t) = v_t$. Then,

$$V(x) = L(h(x_1), \cdots, h(x_t)) = - \log \mathcal{L}(h; \mathcal{D}_t) - C \geq - \log \bar{\mathcal{L}}(x; \mathcal{D}_t) - C.$$

We get $- \log \bar{\mathcal{L}}(x; \mathcal{D}_t) \leq V(x) + C$.

A.3 Proof of the claims in Example 4.1

Since the feedback is noiseless, the likelihood and profile likelihood are binary-valued:

$$\mathcal{L}(f; \mathcal{D}_t) = \mathbf{1}(\phi_i \in \partial f(x_i), \ \forall i \in [t]),$$
$$\bar{\mathcal{L}}(x; \mathcal{D}_t) = \mathbf{1}\Big(\exists f \in \mathcal{F}, \ \text{s.t.} \ f(x) = \max_{x' \in \mathcal{X}} f(x') \ \text{and} \ \phi_i \in \partial f(x_i), \ \forall i \in [t]\Big).$$

14

Hence, the generalized posterior $\mathcal{Q}_t$ is the uniform distribution over the set

$$\mathcal{X}_t = \left\{ x \in \mathcal{X} : \ \exists f \in \mathcal{F} \text{ s.t. } f(x) = \min_{x' \in \mathcal{X}} f(x') \text{ and } \phi_i \in \partial f(x_i), \ \forall i \in [t] \right\}.$$

On the other hand, let $\mathcal{S} = \{ x \in \mathcal{X} : \ \phi_i^\top (x - x_i) \leq 0, \ \forall i \in [t] \}$. We only need to prove that $\mathcal{X}_t = \mathcal{S}$.

The first step is to show $\mathcal{X}_t \subseteq \mathcal{S}$. For any $x \in \mathcal{X}_t$, there exists a convex function f on $\mathcal{X}$ that attains its minimum value at x and satisfies $\phi_i \in \partial f(x_i)$ for all $i \in [t]$. The optimality of x and convexity of f imply that

$$0 \geq f(x) - f(x_i) \geq \phi_i^\top (x - x_i), \qquad \forall i \in [t]$$

and thus $x \in \mathcal{S}$. Consequently, $\mathcal{X}_t \subseteq \mathcal{S}$.

It remains to prove $\mathcal{S} \subseteq \mathcal{X}_t$. Choose any $x \in \mathcal{S}$ and define

$$f(x') = \max_{i \in [t]} [\phi_i^\top (x' - x_i)]_+, \qquad x' \in \mathcal{X}.$$

Here, $(\cdot)_+ = \max\{\cdot, 0\}$ denotes the positive part of a real number. The function f is clearly convex and nonnegative. Meanwhile, the assumption $x \in \mathcal{S}$ forces $f(x) = 0$, which further implies that x is an optimum of f. Therefore, $x \in \mathcal{X}_t$. We get $\mathcal{S} \subseteq \mathcal{X}_t$.

A.4 Proof of the claim in Example 4.2

By applying the first result in Example 4.1 to the new procedure, we see that $\bar{\mathcal{Q}}_{t+1}$ is the uniform distribution over a half ellipsoid $\{ x \in E(c_t, A_t) : \ \phi_{t+1}^\top (x - x_{t+1}) \leq 0 \}$. Let $\bar{E}_{t+1}$ denote this region.

Choose any $\mathcal{Q} = \mathcal{Q}(c, A) \in \Phi$. To make $D_{\mathrm{KL}}(\bar{\mathcal{Q}}_{t+1} \| \mathcal{Q})$ finite, we must have $\bar{\mathcal{Q}}_{t+1} \ll \mathcal{Q}$ and thus $\bar{E}_{t+1} \subseteq E(c, A)$. Then, we have

$$\frac{\mathrm{d}\bar{\mathcal{Q}}_{t+1}}{\mathrm{d}\mathcal{Q}}(x) = \frac{\mathrm{Vol}[E(c, A)]}{\mathrm{Vol}(\bar{E}_{t+1})} \mathbf{1}(x \in \bar{E}_{t+1}),$$

where $\mathrm{Vol}(\cdot)$ denotes the volume of a region. Consequently,

$$D_{\mathrm{KL}}(\bar{\mathcal{Q}}_{t+1} \| \mathcal{Q}) = \mathbb{E}_{x \sim \bar{\mathcal{Q}}_{t+1}} \left[\log \left(\frac{\mathrm{d}\bar{\mathcal{Q}}_{t+1}}{\mathrm{d}\mathcal{Q}}(x) \right) \right] = \frac{1}{\mathrm{Vol}(\bar{E}_{t+1})} \cdot \log \left(\frac{\mathrm{Vol}[E(c, A)]}{\mathrm{Vol}(\bar{E}_{t+1})} \right).$$

This is monotonically increasing in $\mathrm{Vol}[E(c, A)]$. As $\mathcal{Q}_{t+1}$ minimizes the divergence, $E(c_{t+1}, A_{t+1})$ must have the smallest volume among all ellipsoids covering $\bar{E}_{t+1}$. This is precisely the updating rule for the ellipsoid method in convex optimization.

B   Proofs of Section 6

B.1   Proof of Theorem 6.2

We first decompose the regret into the contributions of individual arms:

$$\mathbb{E}[\mathcal{R}(T)] = \sum_{t=1}^{T} \sum_{j=1}^{K} \left(\max_{k \in [K]} \mu_k - \mu_j \right) \mathbb{P}(x_t = j) = \sum_{j=1}^{K} \Delta_j \left(\sum_{t=1}^{T} \mathbb{P}(x_t = j) \right). \tag{B.1}$$

Next, we invoke a lemma on the expected number of pulls of any sub-optimal arm. The proof borrows ideas from the analysis of Thompson sampling by Agrawal and Goyal (2017) and is deferred to Appendix B.3.

Lemma B.1. *There exists a universal constant $C_0 > 0$ such that if $\Delta_j > 0$, then*

$$\sum_{t=1}^{T} \mathbb{P}(x_t = j) \leq C_0 \left(\frac{\sigma^2}{1 - \sigma^{-2}} \cdot \frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j^2} + \frac{1}{\sqrt{1 - \sigma^{-2}}} \right).$$

Choose any $\delta \geq 0$. Then,

$$\sum_{j:\, \Delta_j \leq \delta} \Delta_j \left(\sum_{t=1}^{T} \mathbb{P}(x_t = j) \right) \leq \max_{j:\, \Delta_j \leq \delta} \Delta_j \cdot \sum_{t=1}^{T} \sum_{j=1}^{K} \mathbb{P}(x_t = j) = T \max_{j:\, \Delta_j \leq \delta} \Delta_j.$$

When $\Delta_j > \delta$, we use Lemma B.1 to obtain that

$$\Delta_j \sum_{t=1}^{T} \mathbb{P}(x_t = j) \lesssim \frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j} + \Delta_j,$$

where $\lesssim$ hides a constant factor determined by σ. Plugging these estimates into (B.1) finishes the proof.

B.2 Proof of Theorem 6.1

The result is trivial when $K = 1$ or $T = 1$. From now on, we assume $K \geq 2$, $T \geq 2$ and use $\lesssim$ to hide constant factors determined by σ.

By taking $\delta = 0$ in the regret bound in Theorem 6.2, we obtain that

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \sum_{j:\, \Delta_j > 0} \frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j} + \sum_{j=1}^{K} \Delta_j.$$

Note that

$$\frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j} = \frac{\max\{\log T + 2\log \Delta_j, 1\}}{\Delta_j} \leq \frac{1 + \log T}{\Delta_j} + \frac{2\log(1 + \Delta_j)}{\Delta_j}$$

$$\overset{\text{(i)}}{\leq} \frac{1 + \log T}{\Delta_j} + 2 \overset{\text{(ii)}}{\leq} \frac{1 + \log T}{\Delta_j} + \left(\Delta_j + \frac{1}{\Delta_j} \right) \lesssim \frac{\log T}{\Delta_j} + \Delta_j,$$

where (i) and (ii) follow from elementary inequalities $\log(1 + z) \leq z$ and $z + 1/z \geq 2$ for all $z > 0$. As a result,

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \sum_{j:\, \Delta_j > 0} \frac{\log T}{\Delta_j} + \sum_{j=1}^{K} \Delta_j. \tag{B.2}$$

On the other hand, Theorem 6.2 implies that

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \inf_{\delta \geq 0} \left\{ \sum_{j:\, \Delta_j > \delta} \frac{\log(\max\{T\Delta_j^2, e\})}{\delta} + T\delta \right\} + \sum_{j=1}^{K} \Delta_j.$$

Choose any $\delta \geq e/\sqrt{T}$. When $\Delta_j > \delta$, we have $T\Delta_j^2 \geq e^2 > e$ and

$$\frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j} = \frac{2\log(\sqrt{T}\Delta_j)}{\Delta_j} = 2\sqrt{T} \cdot \frac{\log(\sqrt{T}\Delta_j)}{\sqrt{T}\Delta_j} \leq 2\sqrt{T} \cdot \frac{\log(\sqrt{T}\delta)}{\sqrt{T}\delta} = \frac{2\log(\sqrt{T}\delta)}{\delta},$$

where the inequality follows from the facts that $\sqrt{T}\Delta_j \geq \sqrt{T}\delta \geq e$ and $z \mapsto z^{-1}\log z$ is decreasing on $[e, +\infty)$. Consequently,

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \inf_{\delta \geq e/\sqrt{T}} \left\{ \frac{K\log(\sqrt{T}\delta)}{\delta} + T\delta \right\} + \sum_{j=1}^{K} \Delta_j.$$

Taking $\delta = e\sqrt{T^{-1}K\log K}$, we get

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \sqrt{TK\log K} + \sum_{j=1}^{K} \Delta_j. \tag{B.3}$$

The proof is completed by combining (B.2) and (B.3).

B.3 Proof of Lemma B.1

B.3.1 Preparations

Following the convention, we represent the reward by y_t rather than ϕ_t. We now introduce some key quantities for tracking the iterates.

Definition B.1. *For any $j \in [K]$ and $t \in \mathbb{Z}_+$, denote by $S_j(t) = \{i \in [t-1] : x_i = j\}$ the set of pulls for Arm j in the first $(t-1)$ rounds, and $N_j(t) = |S_j(t)|$ the number of pulls. When $N_j(t) \geq 1$, let*

$$\hat{\mu}_j(t) = \frac{1}{N_j(t)} \sum_{i \in S_j(t)} y_i$$

be the empirical mean reward of Arm j. When $N_j(t) = 0$, let $\hat{\mu}_j(t) = 0$.

Definition B.2. *Denote by $\tau_{j,k}$ the time of the k-th pull of Arm j. Let $\xi_{j,k} = \frac{1}{k}\sum_{i=1}^{k} y_{\tau_{j,i}}$ be the average reward over the first k pulls of Arm j. Let $\mathcal{H}_t$ be the σ-field generated by the data $\mathcal{D}_t$.*

Choose any $M > 0$ and $j \in [K]$ such that $\Delta_j > 0$. Define $u_j = \mu_j + \Delta_j/3$, $v_j = \mu_j + 2\Delta_j/3$, and

$$\mathcal{J}_1 = \sum_{t=1}^{T} \mathbb{P}[x_t = j, N_j(t) \leq M],$$

$$\mathcal{J}_2 = \sum_{t=1}^{T} \mathbb{P}[x_t = j, \hat{\mu}_j(t) \geq u_j],$$

$$\mathcal{J}_3 = \sum_{t=1}^{T} \mathbb{P}[x_t = j, N_j(t) > M, \hat{\mu}_j(t) < u_j].$$

We have a decomposition

$$\sum_{t=1}^{T} \mathbb{P}(x_t = j) \leq \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3. \tag{B.4}$$

By definition,

$$\mathcal{J}_1 = \sum_{t=1}^{T} \mathbb{E}\left(\mathbf{1}[x_t = j, N_j(t) \leq M] \right) = \mathbb{E}\left(\sum_{t=1}^{T} \mathbf{1}[x_t = j, N_j(t) \leq M] \right) \leq M + 1, \tag{B.5}$$

17

$$\mathcal{J}_2 \leq \sum_{k=1}^{\infty} \mathbb{P}(\xi_{j,k} \geq u_j) = \sum_{k=1}^{\infty} \mathbb{P}(\xi_{j,k} - \mu_j \geq \Delta_j/3). \tag{B.6}$$

We invoke useful concentration bounds on the difference between the empirical average reward $\xi_{j,k}$ and the expectation μ_j.

Lemma B.2. *Under Assumption 6.1, we have*

$$\mathbb{P}(\xi_{j,k} - \mu_j \geq t) \leq e^{-kt^2/2}, \qquad \forall t \geq 0,$$
$$\mathbb{P}(\xi_{j,k} - \mu_j \leq -t) \leq e^{-kt^2/2}, \qquad \forall t \geq 0,$$
$$\mathbb{E} e^{\lambda k (\xi_{j,k} - \mu_j)^2/2} \leq \frac{1}{\sqrt{1-\lambda}}, \qquad \forall \lambda \in [0, 1).$$

Proof of Lemma B.2. Note that $\{\xi_{j,k} - \mu_j\}_{k=1}^{\infty}$ is a martingale difference sequence with respect to the filtration $\{\mathcal{H}_{\tau_{j,k}}\}_{k=1}^{\infty}$. Theorem 2.19 in Wainwright (2019) yields the desired tail bounds on $\xi_{j,k} - \mu_j$, together with the fact that $\xi_{j,k}$ is k^{-1}-sub-Gaussian. The proof is then completed by applying Theorem 2.6 in Wainwright (2019). $\square$

By (B.6) and Lemma B.2,

$$\mathcal{J}_2 \leq \sum_{k=0}^{\infty} e^{-k\Delta_j^2/18} = \frac{1}{1 - e^{-\Delta_j^2/18}}.$$

For any $z > 0$, we have $e^z \geq 1 + z$ and thus $e^{-z} \leq (1+z)^{-1}$. Then,

$$\mathcal{J}_2 \leq \frac{1}{1 - (1 + \Delta_j^2/18)^{-1}} = \frac{18}{\Delta_j^2} + 1. \tag{B.7}$$

It remains to bound $\mathcal{J}_3$.

B.3.2 Bounding $\mathcal{J}_3$

Without loss of generality, we assume $\mu_1 = \max_{k \in [K]} \mu_k$ throughout the proof. As a result, $\Delta_j = \mu_1 - \mu_j$. Let $c \in (0, 1)$ be a constant to be determined, and $M' = (1-c)M$. We have

$$\begin{aligned}
&\mathbb{P}[x_t = j, N_j(t) > M, \hat{\mu}_j(t) < u_j] \\
&= \mathbb{P}[x_t = j, N_j(t) > M, \hat{\mu}_j(t) < u_j, N_1(t) > M', \hat{\mu}_1(t) > v_j] \\
&\quad + \mathbb{P}[x_t = j, N_j(t) > M, \hat{\mu}_j(t) < u_j, N_1(t) > M', \hat{\mu}_1(t) \leq v_j] \\
&\quad + \mathbb{P}[x_t = j, N_j(t) > M, \hat{\mu}_j(t) < u_j, 1 \leq N_1(t) < M', \hat{\mu}_1(t) > \hat{\mu}_j(t)] \\
&\quad + \mathbb{P}[x_t = j, N_j(t) > M, \hat{\mu}_j(t) < u_j, 1 \leq N_1(t) < M', \hat{\mu}_1(t) \leq \hat{\mu}_j(t)] \\
&\quad + \mathbb{P}[x_t = j, N_j(t) > M, \hat{\mu}_j(t) < u_j, N_1(t) = 0].
\end{aligned}$$

Denote by $\mathcal{E}_{1,t}$, $\mathcal{E}_{2,t}$, $\mathcal{E}_{3,t}$, $\mathcal{E}_{4,t}$ and $\mathcal{E}_{5,t}$ the five summands on the right-hand side. We have

$$\mathcal{J}_3 \leq \sum_{t=1}^{T} (\mathcal{E}_{1,t} + \mathcal{E}_{2,t} + \mathcal{E}_{3,t} + \mathcal{E}_{4,t} + \mathcal{E}_{5,t}) \tag{B.8}$$

We will control the $\mathcal{E}_{j,t}$'s individually. The following fact will come in handy: for any $\mathcal{H}_{t-1}$-measurable event $\mathcal{A}$,

$$\mathbb{P}(\{x_t = j\} \cap \mathcal{A}) = \mathbb{E}[\mathbb{P}(\{x_t = j\} \cap \mathcal{A}|\mathcal{H}_{t-1})] = \mathbb{E}[\mathbb{P}(x_t = j|\mathcal{H}_{t-1}) \cdot \mathbf{1}(\mathcal{A})] = \mathbb{E}[\mathcal{Q}_{t-1}(j)\mathbf{1}(\mathcal{A})]. \tag{B.9}$$

We also need to characterize the generalized posterior $\mathcal{Q}_t$. Remark 1 implies that

$$\mathcal{Q}_t(j) = \frac{e^{-\Lambda(j,\mathcal{D}_t)}\mathcal{Q}_0(j)}{\sum_{k=1}^K e^{-\Lambda(k,\mathcal{D}_t)}\mathcal{Q}_0(k)},$$

where

$$\Lambda(j,\mathcal{D}_t) = \min_{\boldsymbol{\theta}\in\Theta_j}\left\{\frac{1}{2\sigma^2}\sum_{k=1}^K N_k(t+1)[\hat{\mu}_k(t+1) - \theta_k]^2\right\}. \tag{B.10}$$

Then, we have

$$\frac{\mathcal{Q}_t(j)}{\mathcal{Q}_t(i)} = e^{\Lambda(i,\mathcal{D}_t)-\Lambda(j,\mathcal{D}_t)}\frac{\mathcal{Q}_0(j)}{\mathcal{Q}_0(i)}. \tag{B.11}$$

We invoke some useful estimates for Λ, whose proof is deferred to Appendix B.4.

Lemma B.3. *Suppose that $i,j \in [K]$ and $\hat{\mu}_i(t) \geq \hat{\mu}_j(t)$.*

1. *We have*

$$\Lambda(j,\mathcal{D}_{t-1}) \geq \frac{1}{2\sigma^2}\cdot\frac{[\hat{\mu}_i(t) - \hat{\mu}_j(t)]^2}{1/N_i(t) + 1/N_j(t)}.$$

2. *If $N_j(t) \geq N_i(t)$, then $\Lambda(i,\mathcal{D}_{t-1}) - \Lambda(j,\mathcal{D}_{t-1}) \leq 0$.*

3. *If $N_j(t) < N_i(t)$, then*

$$\Lambda(i,\mathcal{D}_{t-1}) - \Lambda(j,\mathcal{D}_{t-1}) \geq -\frac{1}{2\sigma^2}\cdot\frac{(\hat{\mu}_i - \hat{\mu}_j)^2}{1/N_j - 1/N_i}.$$

We are now in a position to tackle the summands $\{\mathcal{E}_{j,t}\}_{j=1}^5$ in (B.8).

Bounding $\mathcal{E}_{1,t}$. Let $\mathcal{A}$ be the event $\{N_j(t) > M, \hat{\mu}_j(t) < u_j, N_1(t) > M', \hat{\mu}_1(t) > v_j\}$. Choose $\hat{x}_t \in \mathrm{argmax}_{j\in[K]}\hat{\mu}_j(t)$. We have $\Lambda(\hat{x}_t,\mathcal{D}_{t-1}) = 0$. Then, the relation (B.11) and the uniformity of $\mathcal{Q}_0$ yield

$$\mathcal{Q}_{t-1}(j) = e^{\Lambda(\hat{x}_t,\mathcal{D}_{t-1})-\Lambda(j,\mathcal{D}_{t-1})}\mathcal{Q}_{t-1}(\hat{x}_t) \leq e^{-\Lambda(j,\mathcal{D}_{t-1})}.$$

Under $\mathcal{A}$, Part 1 of Lemma B.3 implies that

$$\mathcal{Q}_{t-1}(j) \leq \exp\left(-\frac{1}{2\sigma^2}\cdot\frac{(v_j - u_j)^2}{1/M + 1/M'}\right) = \exp\left(-\frac{M'\Delta_j^2}{36\sigma^2}\right).$$

By (B.9),

$$\mathcal{E}_{1,t} = \mathbb{E}[\mathcal{Q}_{t-1}(j)\cdot\mathbf{1}(\mathcal{A})] \leq \exp\left(-\frac{M'\Delta_j^2}{36\sigma^2}\right). \tag{B.12}$$

Bounding $\mathcal{E}_{2,t}$. We have

$$\mathcal{E}_{2,t} \leq \mathbb{P}[\hat{\mu}_1(t) \leq v_j, N_1(t) > M']. \tag{B.13}$$

For large M', the probability should be small because if Arm 1 is pulled many times, then its empirical average reward should be close to the population mean. We will study this via concentration of self-normalized martingale (Abbasi-Yadkori et al., 2011). For any $t \in \mathbb{Z}_+$, denote by $y_{t,1}$ the potential reward if Arm 1 is pulled in the t-th round, and $\bar{\mathcal{H}}_{t-1}$ the σ-field generated by $\mathcal{D}_{t-1}$ and x_t. Define $\eta_0 = 0$ and $\eta_t = y_{t,1} - \mu_1$ for $t \geq 1$. Then, η_t is $\bar{\mathcal{H}}_t$-measurable, $\mathbb{E}(\eta_t|\bar{\mathcal{H}}_{t-1}) = 0$, and Assumption 6.1 implies that

$$\mathbb{E}(e^{\lambda \eta_t}|\bar{\mathcal{H}}_{t-1}) \leq e^{\lambda^2/2}, \qquad \forall \lambda \in \mathbb{R}.$$

Define

$$V_t = M' + \sum_{i=0}^{t-1} \mathbf{1}(x_i = 1) = M' + N_1(t),$$

$$S_t = \sum_{i=0}^{t-1} \eta_i \mathbf{1}(x_i = 1) = N_1(t)[\hat{\mu}_1(t) - \mu_1].$$

Choose any $\delta > 0$. Theorem 1 in Abbasi-Yadkori et al. (2011) implies that

$$\mathbb{P}\left(|S_t|^2/V_t \leq 2\log(\delta^{-1}\sqrt{V_t/M'}), \ \ \forall t \geq 1 \right) \geq 1 - \delta.$$

When the above event happens, for all $t \geq 1$ we have

$$\frac{\{N_1(t)[\hat{\mu}_1(t) - \mu_1]\}^2}{M' + N_1(t)} \leq 2\log(1/\delta) + \log\left(\frac{M' + N_1(t)}{M'}\right) \leq 2\log(1/\delta) + \frac{N_1(t)}{M'},$$

where the last inequality follows from the elementary fact $\log(1 + z) \leq z, \forall z \geq 0$. Therefore, we have

$$\mathbb{P}\left(\frac{N_1(t)}{M' + N_1(t)}[\hat{\mu}_1(t) - \mu_1]^2 \leq \frac{2\log(1/\delta)}{N_1(t)} + \frac{1}{M'}, \ \ \forall t \geq 1 \right) \geq 1 - \delta.$$

Denote by $\mathcal{A}_\delta$ the above event.

We now come back to (B.13). On the event $\{\hat{\mu}_1(t) \leq v_j, N_1(t) > M'\}$, we have

$$\frac{N_1(t)}{M' + N_1(t)}[\hat{\mu}_1(t) - \mu_1]^2 > \frac{\Delta_j^2}{18} \qquad \text{and} \qquad \frac{2\log(1/\delta)}{N_1(t)} + \frac{1}{M'} < \frac{2\log(1/\delta) + 1}{M'}.$$

Take $\delta = e^{1/2 - M'\Delta_j^2/36}$. We have $[2\log(1/\delta) + 1]/M' = \Delta_j^2/18$ and thus

$$\mathbb{P}[\hat{\mu}_1(t) \leq v_j, N_1(t) > M'] \leq \mathbb{P}\left(\frac{N_1(t)}{M' + N_1(t)}[\hat{\mu}_1(t) - \mu_1]^2 > \frac{2\log(1/\delta)}{N_1(t)} + \frac{1}{M'} \right) \leq \mathbb{P}(\mathcal{A}_\delta^c) \leq \delta.$$

Then, (B.13) leads to

$$\mathcal{E}_{2,t} \leq e^{1/2 - M'\Delta_j^2/36}. \tag{B.14}$$

20

Bounding $\mathcal{E}_{3,t}$. Let $\mathcal{A}$ be the event $\{N_j(t) > M, \hat{\mu}_j(t) < u_j, 1 \le N_1(t) < M', \hat{\mu}_1(t) > \hat{\mu}_j(t)\}$. Under $\mathcal{A}$, we apply the relation (B.11) and Part 2 of Lemma B.3 to Arms 1 and j (as the i and j therein) and then obtain $\mathcal{Q}_{t-1}(j) \le \mathcal{Q}_{t-1}(1)$. Therefore, by (B.9),

$$\mathcal{E}_{3,t} \le \mathbb{E}[\mathcal{Q}_{t-1}(1) \cdot \mathbf{1}(\mathcal{A})] \le \mathbb{E}\Big(\mathbf{1}[x_t = 1, 1 \le N_1(t) < M']\Big),$$

$$\sum_{t=1}^{T} \mathcal{E}_{3,t} \le \mathbb{E}\Big(\sum_{t=1}^{T} \mathbf{1}[x_t = 1, 1 \le N_1(t) < M']\Big) \le \lceil M' \rceil - 1. \tag{B.15}$$

Bounding $\mathcal{E}_{4,t}$. Let $\mathcal{A}$ be the event $\{N_j(t) > M, \hat{\mu}_j(t) < u_j, 1 \le N_1(t) < M', \hat{\mu}_1(t) \le \hat{\mu}_j(t)\}$. By (B.9),

$$\mathcal{E}_{4,t} = \mathbb{E}\Big(\frac{\mathcal{Q}_{t-1}(j)}{\mathcal{Q}_{t-1}(1)} \mathcal{Q}_{t-1}(1) \cdot \mathbf{1}(\mathcal{A}) \Big) = \mathbb{E}\Big(\frac{\mathcal{Q}_{t-1}(j)}{\mathcal{Q}_{t-1}(1)} \cdot \mathbf{1}(\{x_t = 1\} \cap \mathcal{A}) \Big).$$

Under $\mathcal{A}$, we can apply the relation (B.11) and Part 3 of Lemma B.3 to Arms j and 1 (as the i and j therein). This yields

$$\frac{\mathcal{Q}_{t-1}(j)}{\mathcal{Q}_{t-1}(1)} \le \exp\Big(\frac{1}{2\sigma^2} \cdot \frac{[\hat{\mu}_j(t) - \hat{\mu}_1(t)]^2}{1/N_1(t) - 1/N_j(t)} \Big) \le \exp\Big(\frac{1}{2\sigma^2} \cdot \frac{[\mu_1 - \hat{\mu}_1(t)]^2}{1/N_1(t) - 1/N_j(t)} \Big).$$

Since $1 \le N_1(t) < (1-c)M < M < N_j(t)$, we have $N_j(t) > N_1(t)/(1-c)$ and

$$\frac{1}{N_1(t)} - \frac{1}{N_j(t)} \ge \frac{1}{N_1(t)} - \frac{1}{N_1(t)/(1-c)} = \frac{c}{N_1(t)}.$$

Hence,

$$\frac{\mathcal{Q}_{t-1}(j)}{\mathcal{Q}_{t-1}(1)} \le \exp\Big(\frac{N_1(t)[\mu_1 - \hat{\mu}_1(t)]^2}{2c\sigma^2} \Big).$$

We have

$$\sum_{t=1}^{T} \mathcal{E}_{4,t} \le \mathbb{E}\Big[\sum_{t=1}^{T} \exp\Big(\frac{N_1(t)[\hat{\mu}_1(t) - \mu_1]^2}{2c\sigma^2} \Big) \mathbf{1}[x_t = 1, 1 \le N_1(t) < M'] \Big]$$

$$\le \mathbb{E}\Big[\sum_{k=2}^{\lceil M' \rceil - 1} \exp\Big(\frac{(k-1)[\hat{\mu}_1(\tau_{1,k}) - \mu_1]^2}{2c\sigma^2} \Big) \Big] = \sum_{s=1}^{\lceil M' \rceil - 2} \mathbb{E}\Big[\exp\Big(\frac{s(\xi_{1,s} - \mu_1)^2}{2c\sigma^2} \Big) \Big],$$

where $\tau_{1,k}$ is the time of the k-th pull of Arm 1, see Definition B.2.

When $c\sigma^2 > 1$, we obtain from Lemma B.2 that

$$\mathbb{E}\Big[\exp\Big(\frac{s(\xi_{1,s} - \mu_1)^2}{2c\sigma^2} \Big) \Big] \le \frac{1}{\sqrt{1 - 1/(c\sigma^2)}}, \qquad \forall s \in \mathbb{Z}_+.$$

Therefore,

$$\sum_{t=1}^{T} \mathcal{E}_{4,t} \le \frac{\lceil M' \rceil - 2}{\sqrt{1 - 1/(c\sigma^2)}}. \tag{B.16}$$

Bounding $\mathcal{E}_{5,t}$. If $N_1(t) = 0$, then $\Lambda(1, \mathcal{D}_{t-1}) = 0$. The relation (B.11) yields

$$\frac{\mathcal{Q}_{t-1}(j)}{\mathcal{Q}_{t-1}(1)} = e^{\Lambda(1,\mathcal{D}_{t-1})-\Lambda(j,\mathcal{D}_{t-1})} \le 1.$$

Then, by (B.9),

$$\mathcal{E}_{5,t} \le \mathbb{P}[x_t = j, N_1(t) = 0] = \mathbb{E}[\mathcal{Q}_{t-1}(j) \cdot \mathbf{1}(N_1(t) = 0)]$$
$$\le \mathbb{E}[\mathcal{Q}_{t-1}(1) \cdot \mathbf{1}(N_1(t) = 0)] = \mathbb{E}\Big(\mathbf{1}[x_t = 1, N_1(t) = 0]\Big).$$

We have

$$\sum_{t=1}^{T} \mathcal{E}_{5,t} \le \mathbb{E}\Big(\sum_{t=1}^{T} \mathbf{1}[x_t = 1, N_1(t) = 0]\Big) \le 1. \tag{B.17}$$

Bounding $\mathcal{J}_3$. Below we use $\lesssim$ to hide universal constant factors. Summarizing (B.12), (B.14), (B.15), (B.16) and (B.17), we get

$$\mathcal{J}_3 \le T \exp\left(-\frac{M'\Delta_j^2}{36\sigma^2}\right) + Te^{1/2-M'\Delta_j^2/36} + (\lceil M'\rceil - 1) + \frac{\lceil M'\rceil - 2}{\sqrt{1-1/(c\sigma^2)}} + 1$$
$$\lesssim T \exp\left(-\frac{M'\Delta_j^2}{36\sigma^2}\right) + \frac{(M'+1)}{\sqrt{1-1/(c\sigma^2)}} = T \exp\left(-\frac{(1-c)M\Delta_j^2}{36\sigma^2}\right) + \frac{[(1-c)M+1]}{\sqrt{1-1/(c\sigma^2)}},$$

so long as $1/\sigma^2 < c < 1$. Let $c = (1 + \sigma^{-2})/2$. We have $1 - c = (1 - \sigma^{-2})/2$ and $1 - 1/c\sigma^2 = (\sigma^2 - 1)/(\sigma^2 + 1) \ge (1 - \sigma^{-2})/2$. Hence,

$$\mathcal{J}_3 \lesssim T \exp\left(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\right) + \frac{1}{\sqrt{1-\sigma^{-2}}}\left(\frac{1-\sigma^{-2}}{2}M + 1\right)$$
$$\lesssim T \exp\left(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\right) + M + \frac{1}{\sqrt{1-\sigma^{-2}}}. \tag{B.18}$$

B.3.3 Final steps

Combining (B.4), (B.5), (B.7) and (B.18), we get

$$\sum_{t=1}^{T} \mathbb{P}(x_t = j) \lesssim \left[T \exp\left(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\right) + M + \frac{1}{\sqrt{1-\sigma^{-2}}}\right] + \left(\frac{1}{\Delta_j^2} + 1\right) + (M+1)$$
$$\lesssim T \exp\left(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\right) + \frac{1}{\Delta_j^2} + M + \frac{1}{\sqrt{1-\sigma^{-2}}}, \qquad \forall M > 0.$$

By taking

$$M = \frac{72\sigma^2}{1-\sigma^{-2}} \cdot \frac{\log(\max\{T\Delta_j^2, e\}}{\Delta_j^2},$$

we get

$$\sum_{t=1}^{T} \mathbb{P}(x_t = j) \lesssim Te^{-\log(\max\{T\Delta_j^2, e\})} + \frac{1}{\Delta_j^2} + \frac{\sigma^2}{1-\sigma^{-2}} \cdot \frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j^2} + \frac{1}{\sqrt{1-\sigma^{-2}}}$$
$$= \frac{\sigma^2}{1-\sigma^{-2}} \cdot \frac{\log(\max\{T\Delta_j^2, e\})}{\Delta_j^2} + \frac{1}{\sqrt{1-\sigma^{-2}}}.$$

B.4 Proof of Lemma B.3

B.4.1 Part 1

For notational simplicity, we will suppress the time index t in $N_k(t)$'s and $\hat{\mu}_k(t)$'s. The result is trivial when $\hat{\mu}_i = \hat{\mu}_j$. Below we assume that $\hat{\mu}_i > \hat{\mu}_j$. By (B.10), we have

$$2\sigma^2 \Lambda(j, \mathcal{D}_{t-1}) = \min_{\boldsymbol{\theta} \in \Theta_j} \left\{ \sum_{k=1}^K N_k(\hat{\mu}_k - \theta_k)^2 \right\}$$

$$\geq \min_{\boldsymbol{\theta} \in \Theta_j} \left\{ N_j(\hat{\mu}_j - \theta_j)^2 + N_i(\hat{\mu}_i - \theta_i)^2 \right\} = \min_{\theta_j \geq \theta_i} \left\{ N_j(\hat{\mu}_j - \theta_j)^2 + N_i(\hat{\mu}_i - \theta_i)^2 \right\}. \tag{B.19}$$

Denote by $h(\theta_j, \theta_i)$ the function in the brackets. The assumption $\hat{\mu}_i > \hat{\mu}_j$ implies that for any θ_j,

$$\min_{\theta_i \leq \theta_j} h(\theta_j, \theta_i) = h\left(\theta_j, \min\{\hat{\mu}_i, \theta_j\}\right) = N_j(\hat{\mu}_j - \theta_j)^2 + N_i(\hat{\mu}_i - \theta_j)_+^2.$$

View the above as a function of θ_j. It is strictly increasing on $(\hat{\mu}_i, +\infty)$. On the complement set $(-\infty, \hat{\mu}_i]$, the expression simplies to $N_j(\hat{\mu}_j - \theta_j)^2 + N_i(\hat{\mu}_i - \theta_j)^2$. This function's minimizer and minimum value are

$$\frac{N_j \hat{\mu}_j + N_i \hat{\mu}_i}{N_j + N_i} \qquad \text{and} \qquad \frac{(\hat{\mu}_i - \hat{\mu}_j)^2}{1/N_i + 1/N_j}.$$

This fact and (B.19) lead to the desired inequality.

B.4.2 Part 2

Choose any $\bar{\boldsymbol{\theta}} \in \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta_j} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1})$.

Case 1: $\bar{\theta}_j \geq \hat{\mu}_i$. It is easily seen that $\bar{\theta}_i = \hat{\mu}_i$. Define $\boldsymbol{\eta} \in \mathbb{R}^K$ by

$$\eta_k = \begin{cases} \hat{\mu}_j & , \text{ if } k = j \\ \bar{\theta}_j & , \text{ if } k = i \\ \bar{\theta}_k & , \text{ otherwise} \end{cases}.$$

We have $\boldsymbol{\eta} \in \Theta_i$ and

$$\Lambda(i, \mathcal{D}_{t-1}) - \Lambda(j, \mathcal{D}_{t-1}) = \min_{\boldsymbol{\theta} \in \Theta_i} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) - \min_{\boldsymbol{\theta} \in \Theta_j} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) \leq \ell(\boldsymbol{\eta}, \mathcal{D}_{t-1}) - \ell(\bar{\boldsymbol{\theta}}, \mathcal{D}_{t-1})$$

$$= \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \eta_i)^2 + N_j(\hat{\mu}_j - \eta_j)^2 \right) - \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \bar{\theta}_i)^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2 \right)$$

$$= \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \bar{\theta}_j)^2 - N_j(\hat{\mu}_j - \bar{\theta}_j)^2 \right) \leq 0.$$

The last inequality follows from $\bar{\theta}_j \geq \hat{\mu}_i \geq \hat{\mu}_j$ and $N_j \geq N_i$.

Case 2: $\bar{\theta}_j < \hat{\mu}_i$. It is easily seen that $\hat{\mu}_j \leq \bar{\theta}_j = \bar{\theta}_i$. Define $\boldsymbol{\eta} \in \mathbb{R}^K$ by

$$\eta_k = \begin{cases} \hat{\mu}_j & , \text{ if } k = j \\ \hat{\mu}_i & , \text{ if } k = i \\ \bar{\theta}_k & , \text{ otherwise} \end{cases}.$$

23

We have $\boldsymbol{\eta} \in \Theta_i$ and

$$\Lambda(i, \mathcal{D}_{t-1}) - \Lambda(j, \mathcal{D}_{t-1}) = \min_{\boldsymbol{\theta} \in \Theta_i} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) - \min_{\boldsymbol{\theta} \in \Theta_j} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) \leq \ell(\boldsymbol{\eta}, \mathcal{D}_{t-1}) - \ell(\bar{\boldsymbol{\theta}}, \mathcal{D}_{t-1})$$

$$= \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \eta_i)^2 + N_j(\hat{\mu}_j - \eta_j)^2 \right) - \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \bar{\theta}_i)^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2 \right)$$

$$= 0 - \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \bar{\theta}_j)^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2 \right) \leq 0.$$

B.4.3   Part 3

Choose any $\bar{\boldsymbol{\theta}} \in \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta_i} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1})$. We invoke a useful result.

Claim B.1. $\bar{\theta}_i \geq \hat{\mu}_i \geq \hat{\mu}_j = \bar{\theta}_j$

Proof of Claim B.1. Define $\boldsymbol{\eta} \in \mathbb{R}^K$ by

$$\eta_k = \begin{cases} \max\{\bar{\theta}_i, \hat{\mu}_i\} & , \text{ if } k = i \\ \hat{\mu}_j & , \text{ if } k = j \\ \bar{\theta}_k & , \text{ otherwise} \end{cases}.$$

We have $\boldsymbol{\eta} \in \Theta_i$ and

$$0 \geq 2\sigma^2 [\ell(\bar{\boldsymbol{\theta}}, \mathcal{D}_{t-1}) - \ell(\boldsymbol{\eta}, \mathcal{D}_{t-1})] = [N_i(\hat{\mu}_i - \bar{\theta}_i)^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2] - [N_i(\hat{\mu}_i - \eta_i)^2 + N_j(\hat{\mu}_j - \eta_j)^2]$$

$$= N_i(\hat{\mu}_i - \bar{\theta}_i)_-^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2$$

The inequality forces $\bar{\theta}_i \geq \hat{\mu}_i$ and $\bar{\theta}_j = \hat{\mu}_j$. $\qquad \square$

We now come back to the main proof. Define $\boldsymbol{\eta} \in \mathbb{R}^K$ by

$$\eta_k = \begin{cases} \bar{\theta}_i & , \text{ if } k = j \\ \hat{\mu}_i & , \text{ if } k = i \\ \bar{\theta}_k & , \text{ otherwise} \end{cases}.$$

We have $\boldsymbol{\eta} \in \Theta_j$ and

$$\Lambda(i, \mathcal{D}_{t-1}) - \Lambda(j, \mathcal{D}_{t-1}) = \min_{\boldsymbol{\theta} \in \Theta_i} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) - \min_{\boldsymbol{\theta} \in \Theta_j} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) \geq \ell(\bar{\boldsymbol{\theta}}, \mathcal{D}_{t-1}) - \ell(\boldsymbol{\eta}, \mathcal{D}_{t-1})$$

$$= \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \bar{\theta}_i)^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2 \right) - \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \eta_i)^2 + N_j(\hat{\mu}_j - \eta_j)^2 \right)$$

$$= \frac{1}{2\sigma^2} \left(N_i(\hat{\mu}_i - \bar{\theta}_i)^2 - N_j(\hat{\mu}_j - \bar{\theta}_i)^2 \right) \geq \frac{1}{2\sigma^2} \inf_{z \geq \hat{\mu}_i} \left\{ N_i(\hat{\mu}_i - z)^2 - N_j(\hat{\mu}_j - z)^2 \right\}$$

$$= \frac{1}{2\sigma^2} \inf_{z \geq 0} \left\{ N_i z^2 - N_j[z + (\hat{\mu}_i - \hat{\mu}_j)]^2 \right\}.$$

Denote by $g(z)$ the function in the bracket. From

$$g'(z)/2 = N_i z - N_j[z + (\hat{\mu}_i - \hat{\mu}_j)] = (N_i - N_j)z - N_j(\hat{\mu}_i - \hat{\mu}_j).$$

and $N_i > N_j$, we derive that

$$\inf_{z \geq 0} g(z) = g\left(\frac{N_j(\hat{\mu}_i - \hat{\mu}_j)}{N_i - N_j} \right) = -\frac{N_i N_j}{N_i - N_j}(\hat{\mu}_i - \hat{\mu}_j)^2 = -\frac{(\hat{\mu}_i - \hat{\mu}_j)^2}{1/N_j - 1/N_i}.$$

References

ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* **24**.

AGRAWAL, S. and GOYAL, N. (2017). Near-optimal regret bounds for Thompson sampling. *Journal of the ACM (JACM)* **64** 1–24.

AUER, P., CESA-BIANCHI, N. and FISCHER, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning* **47** 235–256.

AUER, P. and ORTNER, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica* **61** 55–65.

BARNDORFF-NIELSEN, O. and COX, D. (1994). *Inference and asymptotics.* Monographs on Statistics and Applied Probability, Chapman & Hall.

BERTSIMAS, D. and VEMPALA, S. (2004). Solving convex programs by random walks. *Journal of the ACM (JACM)* **51** 540–556.

BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **78** 1103–1130.

BUBECK, S. and CESA-BIANCHI, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* **5** 1–122.

CATONI, O. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001.* Springer.

CESA-BIANCHI, N., FREUND, Y., HAUSSLER, D., HELMBOLD, D. P., SCHAPIRE, R. E. and WARMUTH, M. K. (1997). How to use expert advice. *Journal of the ACM (JACM)* **44** 427–485. URL https://doi.org/10.1145/258128.258179

COPE, E. (2007). Bayesian strategies for dynamic pricing in e-commerce. *Naval Research Logistics (NRL)* **54** 265–281.

DEN BOER, A. V. (2015). Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in operations research and management science* **20** 1–18.

FARIAS, V. F. and VAN ROY, B. (2010). Dynamic pricing with a prior on market response. *Operations Research* **58** 16–29.

FRAZIER, P., POWELL, W. and DAYANIK, S. (2009). The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing* **21** 599–613.

FRAZIER, P. I. (2018). Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems.* Informs, 255–278.

GARIVIER, A., MÉNARD, P. and STOLTZ, G. (2019). Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research* **44** 377–399.

HARRISON, J. M., KESKIN, N. B. and ZEEVI, A. (2012). Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science* **58** 570–586.

HENNIG, P. and SCHULER, C. J. (2012). Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research* **13** 1809–1837.

HERNÁNDEZ-LOBATO, J. M., HOFFMAN, M. W. and GHAHRAMANI, Z. (2014). Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems* **27**.

HOEFFDING, W. (1994). Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding* 409–426.

JONES, D. R., SCHONLAU, M. and WELCH, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13** 455–492.

KAUFMANN, E., CAPPÉ, O. and GARIVIER, A. (2012). On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*. PMLR.

KIRSZBRAUN, M. (1934). Über die zusammenziehende und Lipschitzsche Transformationen. *Fundamenta Mathematicae* **22** 77–108.

KLEINBERG, R., SLIVKINS, A. and UPFAL, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*. STOC '08, Association for Computing Machinery, New York, NY, USA.
URL https://doi.org/10.1145/1374376.1374475

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* **6** 4–22.

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.

LITTLESTONE, N. and WARMUTH, M. (1989). The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society.

MCLENNAN, A. (1984). Price dispersion and incomplete learning in the long run. *Journal of Economic Dynamics and Control* **7** 331–347.

MINKA, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. UAI '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

MOČKUS, J. (1974). On Bayesian methods for seeking the extremum. In *IFIP Technical Conference on Optimization Techniques*. Springer.

NESTEROV, Y. (2018). *Lectures on convex optimization*, vol. 137. Springer.

PALADINO, S., TROVO, F., RESTELLI, M. and GATTI, N. (2017). Unimodal Thompson sampling for graph-structured arms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31.

RUSSO, D. and VAN ROY, B. (2018). Learning to optimize via information-directed sampling. *Operations Research* **66** 230–252.

RUSSO, D. J., VAN ROY, B., KAZEROUNI, A., OSBAND, I. and WEN, Z. (2018). A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning* **11** 1–96.

SCOTT, S. L. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* **26** 639–658.

SHAHRIARI, B., SWERSKY, K., WANG, Z., ADAMS, R. P. and DE FREITAS, N. (2015). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE* **104** 148–175.

SOUZA, A., NARDI, L., OLIVEIRA, L. B., OLUKOTUN, K., LINDAUER, M. and HUTTER, F. (2021). Bayesian optimization with a prior for the optimum. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.

SWILER, L. P., GULIAN, M., FRANKEL, A. L., SAFTA, C. and JAKEMAN, J. D. (2020). A survey of constrained Gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing* **1**.

THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.

VAN PARYS, B. and GOLREZAEI, N. (2024). Optimal learning for structured bandits. *Management Science* **70** 3951–3998.

VILLEMONTEIX, J., VAZQUEZ, E. and WALTER, E. (2009). An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization* **44** 509–534.

VOVK, V. G. (1990). Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*. COLT '90, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

WAEBER, R., FRAZIER, P. I. and HENDERSON, S. G. (2013). Bisection search with noisy responses. *SIAM Journal on Control and Optimization* **51** 2261–2279.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge university press.