MINTS: Minimalist Thompson Sampling

Kaizheng Wang
Department of IEOR and Data Science Institute
Columbia University
New York, NY 10027
kaizheng.wang@columbia.edu

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 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. We further analyze MINTS for multi-armed bandits and establish near-optimal regret guarantees.

1 Introduction

Effective sequential decision-making requires balancing the exploration-exploitation tradeoff: gathering new information versus exploiting current knowledge. Simple strategies like explore-then-commit divide the time horizon into distinct phases, but determining the split is difficult. A more adaptive paradigm, optimism in the face of uncertainty, uses upper confidence bounds (UCBs) to guide exploration [20, 2], though this often requires careful, problem-specific calibration. The Bayesian paradigm offers a principled alternative, automatically balancing the tradeoff by maintaining a posterior distribution over all unknown parameters [31, 17]. While elegant, this requirement to specify a full prior becomes a significant bottleneck when incorporating complex structural knowledge, such as shape constraints, as designing a tractable prior that is faithful to them can be prohibitively difficult.

We introduce a minimalist Bayesian framework that allows the user to place a prior only on the location of the optimum, and handles other parameters with constraints via profile likelihood [3]. The reduced-dimension prior and profile likelihood yield a generalized posterior distribution. Based on that, we develop a MINimalist Thompson Sampling (MINTS) algorithm and derive near-optimal regret bounds in multi-armed bandits.

Related work Our framework relates to Thompson sampling (TS) [31, 27]. While TS derives its posterior on the optimum indirectly from a full probabilistic model (e.g., priors on all arm means), our approach specifies a prior directly on the optimum itself. A similar idea exists in full-information online learning [22, 33, 7, 6, 4], but our work addresses the more challenging partial-feedback setting.

Existing methods reasoning about the optimum are either tailored to special structures [34], or still rely on full probabilistic models for belief updates [32, 14, 15, 26]. Closer in spirit, Souza et al. (2021) augment a standard model with a prior on the optimum [29], whereas our framework uses profile likelihood to replace the nuisance parameter model entirely.

Finally, our framework offers a unified approach to Bayesian optimization under structural constraints, a domain largely addressed with case-by-case solutions tailored to specific scenarios [30, 25].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2 Preliminaries

In the standard setup of stochastic optimization, an agent seeks to solve a maximization problem:

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

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 the cumulative regret $\sum_{t=1}^{T} [\max_{x \in \mathcal{X}} f(x) - f(x_t)]$, or the simple regret $\max_{x \in \mathcal{X}} f(x) - f(x_T)$.

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 [19])**.** *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 [8])**.** *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 $\mathcal{P}_{x_t}$.*

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. Consider a family of problem instances $\{P_\theta\}_{\theta \in \Theta}$, indexed by a parameter θ in a 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 θ is drawn from a prior distribution $\mathcal{Q}_0$ over Θ [31, 17, 28, 11]. 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:

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

2. (Decision-making) Choose x_{t+1} by optimizing a criterion based on $\mathcal{Q}_t$ [24, 17, 10, 31, 18, 26].

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. 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 $f(x) = x\theta_x$. The likelihood of data $\mathcal{D}_t$ is $\mathcal{L}(\theta; \mathcal{D}_t) = \prod_{i=1}^{t} \theta_{x_i}^{\phi_i} (1 - \theta_{x_i})^{1 - \phi_i}$.

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 [23, 9, 13]. On the other hand, nonparametric models offer greater flexibility but run into the obstacle of structural constraints. Below we show the challenges arising from these constraints.

Suppose that $\mathcal{X}$ only consists of K prices $p_1 < \cdots < p_K$. Then, the function θ is represented by a vector $\boldsymbol{\theta} = (\theta_{p_1}, \cdots, \theta_{p_K})^\top$. It is common to assume that θ_x is non-increasing and M-Lipschitz in x, where $M > 0$ is a constant. The structural constraints confine $\boldsymbol{\theta}$ to the following convex set:

$$\{\boldsymbol{v} \in \mathbb{R}^K : \ 0 \le v_K \le \cdots \le v_1 \le 1 \text{ and } v_j - v_{j+1} \le M(p_{j+1} - p_j), \ \forall j \in [K]\}.$$

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 inapplicable.

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 $H_j : \boldsymbol{\theta} \in \Theta_j = \{\boldsymbol{v} \in \Theta : v_j \geq v_k, \forall k \in [K]\}$. The likelihood $\mathcal{L}$ 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 [3]. Define the profile likelihood of Arm j as the maximum likelihood achievable by any parameter vector consistent with H_j:

$$\bar{\mathcal{L}}(j; \mathcal{D}_t) = \max_{\boldsymbol{v} \in \Theta_j} \mathcal{L}(\boldsymbol{v}; \mathcal{D}_t). \tag{3.2}$$

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

$$\mathcal{Q}_t(j) = \frac{\bar{\mathcal{L}}(j; \mathcal{D}_t)\mathcal{Q}_0(j)}{\sum_{k=1}^{K} \bar{\mathcal{L}}(k; \mathcal{D}_t)\mathcal{Q}_0(k)}, \qquad j \in [K]. \tag{3.3}$$

It represents our updated belief about which arm is optimal.

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. We have $-\log \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}_t) = \frac{1}{2\sigma^2} \sum_{j=1}^{K} |I_j|(\theta_j - \hat{\mu}_j)^2 + C$, where C is a constant. Denote by $L(\boldsymbol{\theta})$ the first term on the right-hand side. Then, $\log \bar{\mathcal{L}}(j; \mathcal{D}_t) = -\min_{\boldsymbol{\theta} \in \Theta_j} L(\boldsymbol{\theta}) - C$. 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:

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

Remark 2 (Structured bandits)**.** Structural constraints on the parameter $\boldsymbol{\theta}$ can be seamlessly incorporated by restricting the parameter space Θ, e.g., adding the Lipschitz condition in Example 2.2. The rest of the inferential procedure remains exactly the same.

Remark 3 (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. Let f_θ denote the objective function in the problem instance P_θ. After t rounds of interaction, we obtain data $\mathcal{D}_t$ and construct the profile likelihood

$$\bar{\mathcal{L}}(x; \mathcal{D}_t) = \sup \left\{ \mathcal{L}(\theta; \mathcal{D}_t) : \theta \in \Theta \text{ and } f_\theta(x) = \max_{x' \in \mathcal{X}} f_\theta(x') \right\}, \qquad x \in \mathcal{X}, \tag{3.4}$$

and derive a generalized posterior distribution $\mathcal{Q}_t$ by reweighting the prior $\mathcal{Q}_0$:

$$\frac{\mathrm{d}\mathcal{Q}_t}{\mathrm{d}\mathcal{Q}_0}(x) = \frac{\bar{\mathcal{L}}(x; \mathcal{D}_t)}{\int_{\mathcal{X}} \bar{\mathcal{L}}(x'; \mathcal{D}_t)\mathcal{Q}_0(\mathrm{d}x')}, \qquad x \in \mathcal{X}. \tag{3.5}$$

Then, it is natural to draw the next decision x_{t+1} from $\mathcal{Q}_t$. We name this procedure as **MINimalist Thompson Sampling** (MINTS). It is conceptually simpler than standard Thompson Sampling.

4 Theoretical analysis for multi-armed bandits

We now provide theoretical guarantees for MINTS on the multi-armed bandit problem. Consider the multi-armed bandit problem in Example 2.1. To implement the MINTS algorithm, 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 rather than a theoretical assumption.

To state the formal results, we introduce the performance measure and a tail condition for the rewards.

Definition 4.1 (Regret)**.** *The performance of a decision sequence $\{x_t\}_{t=1}^T$ is measured by its regret:*

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

Assumption 4.1 (Sub-Gaussian reward)**.** *The reward distributions $\{\mathcal{P}_j\}_{j=1}^K$ are 1-sub-Gaussian:*

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

Assumption 4.1 is standard for bandit studies [21]. 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 [16]. For sub-Gaussian distributions with general variance proxies, we can reduce to this case by rescaling.

We present a general regret bound with explicit dependence on the sub-optimality gaps of arms. The proof can be found in Appendix A.1.

Theorem 4.1 (Regret bound)**.** *For the multi-armed bandit in Example 2.1, run MINTS 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 4.1, there exists a constant C determined by σ such that*

$$\mathbb{E}[\mathcal{R}(T)] \leq C \inf_{\delta \geq 0} \left\{ \sum_{j:\ \Delta_j > \delta} \left(\frac{\log T}{\Delta_j} + \Delta_j \right) + T \max_{j:\ \Delta_j \leq \delta} \Delta_j \right\}, \qquad \forall T \geq 2.$$

Next, we further derive more interpretable results under an additional assumption that the mean rewards $\{\mu_j\}_{j=1}^K$ belong to a constant-width interval, which is frequently used in the literature. See Appendix A.2 for the proof.

Corollary 4.1. *Consider the setup in Theorem 4.1. Assume that $\max_{j \in [K]} \mu_j - \min_{j \in [K]} \mu_j \leq c$ holds for some constant c. There exists another constant C determined by c and σ such that*

$$\mathbb{E}[\mathcal{R}(T)] \leq C \min \left\{ \sum_{j:\ \Delta_j > 0} \frac{\log T}{\Delta_j}, \sqrt{KT \log T} \right\}, \qquad \forall T \geq K. \tag{4.1}$$

Corollary 4.1 demonstrates the near-optimality of MINTS through a logarithmic, problem-dependent regret bound and a root-T, problem-independent one. When the problem instance is fixed and T is sufficiently large, the first result matches the lower bound for Gaussian bandit in [12] up to a constant factor. On the other hand, for a fixed T, the second result achieves the minimax lower bound in [5] up to a $\sqrt{\log T}$ factor. We believe that $\log T$ can be further sharpened to $\log K$, as in the regret bound for Gaussian Thompson sampling [1].

5 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.

Acknowledgments and Disclosure of Funding

We thank Yeon-Koo Che, Yaqi Duan and Chengpiao Huang for helpful discussions. This research is supported by NSF grants DMS-2210907 and DMS-2515679.

References

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

[2] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

[3] O. E. Barndorff-Nielsen and D. R. Cox. *Inference and asymptotics*. Monographs on Statistics and Applied Probability. Chapman & Hall, 1994.

[4] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

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

[6] Olivier Catoni. *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour XXXI-2001*. Springer, 2004.

[7] Nicoló Cesa-Bianchi, Yoav Freund, David Haussler, David P. Helmbold, Robert E. Schapire, and Manfred K. Warmuth. How to use expert advice. *Journal of the ACM (JACM)*, 44(3):427–485, May 1997.

[8] Arnoud V Den Boer. Dynamic pricing and learning: Historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

[9] Vivek F Farias and Benjamin Van Roy. Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29, 2010.

[10] Peter Frazier, Warren Powell, and Savas Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS Journal on Computing*, 21(4):599–613, 2009.

[11] Peter I Frazier. Bayesian optimization. In *Recent advances in optimization and modeling of contemporary problems*, pages 255–278. Informs, 2018.

[12] Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

[13] J Michael Harrison, N Bora Keskin, and Assaf Zeevi. Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Science*, 58(3):570–586, 2012.

[14] Philipp Hennig and Christian J Schuler. Entropy search for information-efficient global optimization. *The Journal of Machine Learning Research*, 13(1):1809–1837, 2012.

[15] José M Hernández-Lobato, Matthew W Hoffman, and Zoubin Ghahramani. Predictive entropy search for efficient global optimization of black-box functions. *Advances in neural information processing systems*, 27, 2014.

[16] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[17] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[18] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On Bayesian upper confidence bounds for bandit problems. In *Artificial Intelligence and Statistics*, pages 592–600. PMLR, 2012.

[19] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, page 681–690, New York, NY, USA, 2008. Association for Computing Machinery.

[20] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[21] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[22] N Littlestone and MK Warmuth. The weighted majority algorithm. In *30th Annual Symposium on Foundations of Computer Science*, pages 256–261. IEEE Computer Society, 1989.

[23] Andrew McLennan. Price dispersion and incomplete learning in the long run. *Journal of Economic Dynamics and Control*, 7(3):331–347, 1984.

[24] Jonas Močkus. On Bayesian methods for seeking the extremum. In *IFIP Technical Conference on Optimization Techniques*, pages 400–404. Springer, 1974.

[25] Stefano Paladino, Francesco Trovo, Marcello Restelli, and Nicola Gatti. Unimodal Thompson sampling for graph-structured arms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[26] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

[27] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

[28] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[29] Artur Souza, Luigi Nardi, Leonardo B Oliveira, Kunle Olukotun, Marius Lindauer, and Frank Hutter. Bayesian optimization with a prior for the optimum. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 265–296. Springer, 2021.

[30] Laura P Swiler, Mamikon Gulian, Ari L Frankel, Cosmin Safta, and John D Jakeman. A survey of constrained Gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2), 2020.

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

[32] Julien Villemonteix, Emmanuel Vazquez, and Eric Walter. An informational approach to the global optimization of expensive-to-evaluate functions. *Journal of Global Optimization*, 44(4):509–534, 2009.

[33] Volodimir G. Vovk. Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, page 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.

[34] Rolf Waeber, Peter I Frazier, and Shane G Henderson. Bisection search with noisy responses. *SIAM Journal on Control and Optimization*, 51(3):2261–2279, 2013.

[35] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

A Proofs of Section 4

We present the proofs of Theorem 4.1 and Corollary 4.1.

A.1 Proof of Theorem 4.1

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

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

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 [1] and is deferred to Appendix A.3.

Lemma A.1. *If $\Delta_j > 0$ and $T \geq 2$, then*

$$\sum_{t=1}^{T}\mathbb{P}(x_t = j) \leq \frac{234\sigma^2}{1-\sigma^{-2}}\cdot\frac{\log T}{\Delta_j^2} + \frac{7}{\sqrt{1-\sigma^{-2}}}.$$

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

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

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

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

where $\lesssim$ only hides a constant factor determined by σ. Hence,

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \sum_{j:\ \Delta_j>\delta}\left(\frac{\log T}{\Delta_j} + \Delta_j\right) + T\max_{j:\ \Delta_j\leq\delta}\Delta_j, \qquad \forall\delta \geq 0.$$

A.2 Proof of Corollary 4.1

By taking $\delta = 0$ in the regret bound in Theorem 4.1 and using the assumption $\max_{j\in[K]}\mu_j - \min_{j\in[K]}\mu_j \lesssim 1$, we obtain that

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \sum_{j:\ \Delta_j>0}\left(\frac{\log T}{\Delta_j} + \Delta_j\right) \lesssim \sum_{j:\ \Delta_j>0}\frac{\log T}{\Delta_j}.$$

It remains to prove $\mathbb{E}[\mathcal{R}(T)] \lesssim \sqrt{KT\log T}$. The result is trivial when $K = 1$. Suppose that $K \geq 2$. By Theorem 4.1,

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \sum_{j:\ \Delta_j>\delta}\left(\frac{\log T}{\Delta_j} + \Delta_j\right) + T\delta.$$

Let $\delta = \sqrt{T^{-1}K\log T}$. Then,

$$\mathbb{E}[\mathcal{R}(T)] \lesssim \sum_{j:\ \Delta_j>\delta}\left(\frac{\log T}{\sqrt{T^{-1}K\log T}} + c\right) + T\sqrt{\frac{K\log T}{T}} \lesssim \sqrt{KT\log T}.$$

A.3 Proof of Lemma A.1

A.3.1 Preparations

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

Definition A.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 A.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

$$\begin{aligned}
\mathcal{J}_1 &= \sum_{t=1}^{T} \mathbb{P}[x_t = j, \hat{\mu}_j(t) < u_j, N_j(t) > M], \\
\mathcal{J}_2 &= \sum_{t=1}^{T} \mathbb{P}[x_t = j, \hat{\mu}_j(t) \geq u_j, N_j(t) > M], \\
\mathcal{J}_3 &= \sum_{t=1}^{T} \mathbb{P}[x_t = j, N_j(t) \leq M].
\end{aligned}$$

We have a decomposition

$$\sum_{t=1}^{T} \mathbb{P}(x_t = j) = \mathcal{J}_1 + \mathcal{J}_2 + \mathcal{J}_3. \tag{A.2}$$

By definition,

$$\mathcal{J}_3 = \sum_{t=1}^{T} \mathbb{E}\Big(\mathbf{1}[x_t = j, N_j(t) \leq M]\Big) = \mathbb{E}\bigg(\sum_{t=1}^{T} \mathbf{1}[x_t = j, N_j(t) \leq M]\bigg) \leq M + 1. \tag{A.3}$$

To bound $\mathcal{J}_2$, note that

$$\sum_{t=1}^{T} \mathbf{1}[x_t = j, \hat{\mu}_j(t) \geq u_j, N_j(t) > M] \leq \sum_{k=M+1}^{T} \mathbf{1}(\xi_{j,k} \geq u_j).$$

Hence,

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

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

Lemma A.2. *Under Assumption 4.1, we have*

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

Proof of Lemma A.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 [35] 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 [35]. □

By Lemma A.2, for any $k > M$, we have

$$\mathbb{P}(\xi_{j,k} - \mu_j \geq \Delta_j/3) \leq e^{-k(\Delta_j/3)^2/2} = e^{-k\Delta_j^2/18}.$$

As a result,

$$\mathcal{J}_2 \leq \sum_{k=M+1}^{\infty} e^{-k\Delta_j^2/18} = \frac{e^{-(M+1)\Delta_j^2/18}}{1 - e^{-\Delta_j^2/18}}. \tag{A.4}$$

It remains to bound $\mathcal{J}_1$.

A.3.2 Bounding $\mathcal{J}_1$

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

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

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

$$\mathcal{J}_1 \leq \sum_{t=1}^{T} (\mathcal{E}_{1,t} + \mathcal{E}_{2,t} + \mathcal{E}_{3,t} + \mathcal{E}_{4,t} + \mathcal{E}_{5,t}) \tag{A.5}$$

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 $\mathcal{A}$,

$$\mathbb{P}(\{x_t = j\} \cap \mathcal{A}) = \mathbb{E}[\mathbb{P}(\{x_t = j\} \cap \mathcal{A} | \mathcal{H}_{t-1})] = \mathbb{E}[\mathbb{P}(x_t = j | \mathcal{H}_{t-1}) \cdot \mathbf{1}(\mathcal{A})] = \mathbb{E}[\mathcal{Q}_{t-1}(j)\mathbf{1}(\mathcal{A})]. \tag{A.6}$$

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

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

holds for some constant C. For any $j \in [K]$, define

$$\Lambda(j, \mathcal{D}_t) = \min_{\boldsymbol{\theta} \in \Theta_j} \left\{ \frac{1}{2\sigma^2} \sum_{k=1}^{K} N_k(t+1)[\hat{\mu}_k(t+1) - \theta_k]^2 \right\}. \tag{A.7}$$

Then, we have

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

It is easily seen that

$$\frac{\mathcal{Q}_t(j)}{\mathcal{Q}_t(i)} = e^{\Lambda(i,\mathcal{D}_t) - \Lambda(j,\mathcal{D}_t)} \frac{\mathcal{Q}_0(j)}{\mathcal{Q}_0(i)}. \tag{A.8}$$

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

Lemma A.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 (A.5).

Bounding $\mathcal{E}_{1,t}$. Let $\mathcal{A}$ be the event $\{\hat{\mu}_j(t) < u_j, N_j(t) > M, \hat{\mu}_1(t) > v_j, N_1(t) > M'\}$. Choose $\hat{x}_t \in \operatorname{argmax}_{j\in[K]} \hat{\mu}_j(t)$. We have $\Lambda(\hat{x}_t, \mathcal{D}_{t-1}) = 0$. Then, the relation (A.8) 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 A.3 implies that

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

By (A.6),

$$\mathcal{E}_{1,t} = \mathbb{E}[\mathcal{Q}_{t-1}(j) \cdot \mathbf{1}(\mathcal{A})] \leq \exp\bigg(-\frac{M'\Delta_j^2}{36\sigma^2} \bigg). \tag{A.9}$$

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

$$\mathcal{E}_{2,t} \leq \mathbb{P}[\hat{\mu}_1(t) \leq v_j, N_1(t) > M'] \leq \sum_{k=M'+1}^{\infty} \mathbb{P}[\hat{\mu}_1(t) \leq v_j, N_1(t) = k] \leq \sum_{k=M'+1}^{\infty} \mathbb{P}(\xi_{1,k} \leq v_j).$$

For any $k > M'$, Lemma A.2 yields

$$\mathbb{P}(\xi_{1,k} \leq v_j) = \mathbb{P}(\xi_{1,k} - \mu_1 \leq -\Delta_j/3) \leq e^{-k(\Delta_j/3)^2/2} = e^{-k\Delta_j^2/18}.$$

Hence,

$$\mathcal{E}_{2,t} \leq \sum_{k=M'+1}^{\infty} e^{-k\Delta_j^2/18} = \frac{e^{-(M'+1)\Delta_j^2/18}}{1 - e^{-\Delta_j^2/18}}. \tag{A.10}$$

Bounding $\mathcal{E}_{3,t}$. Let $\mathcal{A}$ be the event $\{\hat{\mu}_j(t) < u_j, N_j(t) > M, \hat{\mu}_1(t) > \hat{\mu}_j(t), 1 \leq N_1(t) < M'\}$. Under $\mathcal{A}$, we apply the relation (A.8) and Part 2 of Lemma A.3 to the indices 1 and j (as the i and j therein) and then obtain $\mathcal{Q}_{t-1}(j) \leq \mathcal{Q}_{t-1}(1)$. Therefore, by (A.6),

$$\begin{aligned}
&\mathcal{E}_{3,t} \leq \mathbb{E}[\mathcal{Q}_{t-1}(1) \cdot \mathbf{1}(\mathcal{A})] \leq \mathbb{E}\Big(\mathbf{1}[x_t = 1, 1 \leq N_1(t) < M']\Big), \\
&\sum_{t=1}^{T} \mathcal{E}_{3,t} \leq \mathbb{E}\bigg(\sum_{t=1}^{T} \mathbf{1}[x_t = 1, 1 \leq N_1(t) < M'] \bigg) \leq \lceil M' \rceil - 1.
\end{aligned} \tag{A.11}$$

Bounding $\mathcal{E}_{4,t}$. Let $\mathcal{A}$ be the event $\{\hat{\mu}_j(t) < u_j, N_j(t) > M, \hat{\mu}_1(t) \leq \hat{\mu}_j(t), 1 \leq N_1(t) < M'\}$. By (A.6),

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

Under $\mathcal{A}$, we can apply the relation (A.8) and Part 3 of Lemma A.3 to the indices j and 1 (as the i and j therein). This yields

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

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

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

Hence,

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

We have

$$\begin{aligned}\sum_{t=1}^{T}\mathcal{E}_{4,t} &\le \mathbb{E}\Big[\sum_{t=1}^{T}\exp\left(\frac{N_1(t)[\hat{\mu}_1(t)-\mu_1]^2}{2c\sigma^2}\right)\mathbf{1}[x_t=1, 1\le N_1(t)<M']\Big]\\ &\le \mathbb{E}\Big[\sum_{k=2}^{\lceil M'\rceil-1}\exp\left(\frac{(k-1)[\hat{\mu}_1(\tau_{1,k})-\mu_1]^2}{2c\sigma^2}\right)\Big] = \sum_{s=1}^{\lceil M'\rceil-2}\mathbb{E}\Big[\exp\left(\frac{s(\xi_{1,s}-\mu_1)^2}{2c\sigma^2}\right)\Big],\end{aligned}$$

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

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

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

Therefore,

$$\sum_{t=1}^{T}\mathcal{E}_{4,t} \le \frac{\lceil M'\rceil - 2}{\sqrt{1-1/(c\sigma^2)}}. \tag{A.12}$$

Bounding $\mathcal{E}_{5,t}$. If $N_1(t)=0$, then $\Lambda(1,\mathcal{D}_{t-1})=0$. The relation (A.8) 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 (A.6),

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

We have

$$\sum_{t=1}^{T}\mathcal{E}_{5,t} \le \mathbb{E}\Big(\sum_{t=1}^{T}\mathbf{1}[x_t=1, N_1(t)=0]\Big) \le 1. \tag{A.13}$$

Bounding $\mathcal{J}_1$. Summarizing (A.9), (A.10), (A.11), (A.12) and (A.13), we get

$$\begin{aligned}\mathcal{J}_1 &\le T\exp\Big(-\frac{M'\Delta_j^2}{36\sigma^2}\Big) + T\frac{e^{-(M'+1)\Delta_j^2/18}}{1-e^{-\Delta_j^2/18}} + (\lceil M'\rceil - 1) + \frac{\lceil M'\rceil - 2}{\sqrt{1-1/(c\sigma^2)}} + 1\\ &\le \frac{2T}{1-e^{-\Delta_j^2/18}}\exp\Big(-\frac{M'\Delta_j^2}{36\sigma^2}\Big) + \frac{2(M'+1)}{\sqrt{1-1/(c\sigma^2)}}\\ &\le \frac{2T}{1-e^{-\Delta_j^2/18}}\exp\Big(-\frac{(1-c)M\Delta_j^2}{36\sigma^2}\Big) + \frac{2[(1-c)M+1]}{\sqrt{1-1/(c\sigma^2)}},\end{aligned}$$

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,

$$\begin{aligned}\mathcal{J}_1 &\le \frac{2T}{1-e^{-\Delta_j^2/18}}\exp\Big(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\Big) + \frac{2\sqrt{2}}{\sqrt{1-\sigma^{-2}}}\Big(\frac{1-\sigma^{-2}}{2}M+1\Big)\\ &\le \frac{2T}{1-e^{-\Delta_j^2/18}}\exp\Big(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\Big) + \sqrt{2}M + \frac{2\sqrt{2}}{\sqrt{1-\sigma^{-2}}}.\end{aligned} \tag{A.14}$$

A.3.3 Final steps

Combining (A.2), (A.3), (A.4) and (A.14), we get

$$\sum_{t=1}^{T}\mathbb{P}(x_t=j) \le \left[\frac{2T}{1-e^{-\Delta_j^2/18}}\exp\left(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\right)+\sqrt{2}M+\frac{2\sqrt{2}}{\sqrt{1-\sigma^{-2}}}\right]+\frac{e^{-M\Delta_j^2/18}}{1-e^{-\Delta_j^2/18}}+(M+1)$$
$$\le \frac{3T}{1-e^{-\Delta_j^2/18}}\exp\left(-\frac{(1-\sigma^{-2})M\Delta_j^2}{72\sigma^2}\right)+\frac{5M}{2}+\frac{4}{\sqrt{1-\sigma^{-2}}}, \qquad \forall M>0.$$

By taking

$$M=\frac{72\sigma^2}{1-\sigma^{-2}}\cdot\frac{\log T}{\Delta_j^2}>0,$$

we get

$$\sum_{t=1}^{T}\mathbb{P}(x_t=j) \le \frac{3T}{1-e^{-\Delta_j^2/18}}e^{-\log T}+\frac{5}{2}\cdot\frac{72\sigma^2}{1-\sigma^{-2}}\cdot\frac{\log T}{\Delta_j^2}+\frac{4}{\sqrt{1-\sigma^{-2}}}$$
$$=\frac{3}{1-e^{-\Delta_j^2/18}}+\frac{180\sigma^2}{1-\sigma^{-2}}\cdot\frac{\log T}{\Delta_j^2}+\frac{4}{\sqrt{1-\sigma^{-2}}}.$$

The fact $1/(1-e^{-z})\le 1/z+1, \forall z>0$ yields

$$\sum_{t=1}^{T}\mathbb{P}(x_t=j) \le \frac{234\sigma^2}{1-\sigma^{-2}}\cdot\frac{\log T}{\Delta_j^2}+\frac{7}{\sqrt{1-\sigma^{-2}}}.$$

A.4 Proof of Lemma A.3

A.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 (A.7), we have

$$2\sigma^2\Lambda(j,\mathcal{D}_{t-1}) = \min_{\boldsymbol{\theta}\in\Theta_j}\left\{\sum_{k=1}^{K}N_k(\hat{\mu}_k-\theta_k)^2\right\}$$
$$\ge \min_{\boldsymbol{\theta}\in\Theta_j}\left\{N_j(\hat{\mu}_j-\theta_j)^2+N_i(\hat{\mu}_i-\theta_i)^2\right\} = \min_{\theta_j\ge\theta_i}\left\{N_j(\hat{\mu}_j-\theta_j)^2+N_i(\hat{\mu}_i-\theta_i)^2\right\}. \quad \text{(A.15)}$$

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,

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

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 (A.15) lead to the desired inequality.

A.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: $\bar{\theta}_j\ge\hat{\mu}_i$. 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

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

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

$$\Lambda(i, \mathcal{D}_{t-1}) - \Lambda(j, \mathcal{D}_{t-1}) = \min_{\boldsymbol{\theta} \in \Theta_i} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) - \min_{\boldsymbol{\theta} \in \Theta_j} \ell(\boldsymbol{\theta}, \mathcal{D}_{t-1}) \leq \ell(\boldsymbol{\eta}, \mathcal{D}_{t-1}) - \ell(\bar{\boldsymbol{\theta}}, \mathcal{D}_{t-1})$$
$$= \frac{1}{2\sigma^2}\left(N_i(\hat{\mu}_i - \eta_i)^2 + N_j(\hat{\mu}_j - \eta_j)^2\right) - \frac{1}{2\sigma^2}\left(N_i(\hat{\mu}_i - \bar{\theta}_i)^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2\right)$$
$$= 0 - \frac{1}{2\sigma^2}\left(N_i(\hat{\mu}_i - \bar{\theta}_j)^2 + N_j(\hat{\mu}_j - \bar{\theta}_j)^2\right) \leq 0.$$

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

Proof of Claim A.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 $\bar{\theta}_j = \hat{\mu}_j$. □

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

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