BEST ARM IDENTIFICATION IN GENERALIZED LINEAR BANDITS WITH 1-BIT FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

In many real-world applications such as IoT sensor networks, digital marketing, and clinical trials, decision-making relies on binary feedback or operates under strict 1-bit communication constraints. Generalized linear models (GLMs) provide a natural framework for modeling outcomes using high-dimensional contextual features. When rewards are observed only through thresholded binary feedback, a central question is how the resulting information loss affects best arm identification (BAI). In this paper, we study BAI in GLMs with 1-bit observations, where the latent reward follows an exponential-family distribution but the learner observes only thresholded binary feedback. Our algorithm constructs confidence sets directly for the pairwise reward gaps between arms, formulated in terms of the feature matrices V and an m-dimensional parameter vector θ. Under suitable regularity and identifiability assumptions, we prove that the proposed algorithm is (ϵ, δ)-PAC and establish upper and lower sample-complexity bounds governed by the same instance-dependent hardness, up to logarithmic factors and the stated regularity and design constants. The cost of quantization enters through the censored Fisher information along reward-gap directions, making explicit how the information retained by each binary observation affects the exploration required to identify an approximately optimal arm.

Keywords: generalized linear bandits, best-arm identification, 1-bit feedback, Fisher information, exponential family.

1 INTRODUCTION

In many sequential decision problems, a learner must identify the best arm while observing only whether its reward crossed a threshold. In our setting, each arm has a structured latent reward Y_k, whose distribution depends on a shared unknown parameter vector θ^*. Rather than observing the full reward, the learner receives a single bit of feedback $G_k = \mathbf{1}\{Y_k \leq \gamma_k\}$, where γ_k is a fixed arm-specific threshold. The objective remains to identify the best arm according to its latent reward. This observation bottleneck raises a fundamental question: *How much extra exploration is required to identify the best arm from only one bit of feedback per pull?*

This constraint arises naturally in several applications. In sensor fusion, limited power and communication bandwidth motivate transmitting thresholded sensor observations instead of full measurements (Kumar & Vatedka, 2023; Fang et al., 2013). In survival analysis with current-status observations, the available feedback records only whether an event has occurred by a specified inspection time, while the exact event time remains unobserved (Jewell & van der Laan, 2003). In both cases, the learner must make decisions using binary observations of an underlying quantity of interest.

We study this problem through generalized linear models (GLMs), where latent rewards follow an exponential-family distribution whose natural parameter is determined by the arm's features and the shared parameter θ^*. Although thresholding removes direct access to reward magnitudes, the probability $\Pr_\theta(Y_k \leq \gamma_k)$ retains a known parametric structure. The learner can therefore estimate θ^* from the binary likelihood and use the model to infer reward gaps between arms. Because the parameter is shared, observations from one arm can also provide information about other arms.

The central statistical object is the *censored Fisher information matrix*, which describes the local information about θ^* retained by the binary observations. This information depends on both the arm

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

features and the observation thresholds, so the information available under full reward observations does not directly describe uncertainty under one-bit feedback. For best arm identification, the relevant uncertainty concerns parameter directions that change the reward gaps between competing arms. This motivates our approach: construct confidence sets directly for pairwise reward gaps and guide exploration using the information retained after thresholding.

1.1 Related Work

1.1.1 Best Arm Identification in Generalized Linear Bandits

Fixed-confidence best arm identification in generalized linear bandits exploits a shared unknown parameter to identify an approximately optimal arm using as few observations as possible. Kazerouni & Wein (2021) propose a gap-based exploration algorithm that constructs confidence sets directly for pairwise reward gaps and selects arms to reduce uncertainty in competitive comparisons. Their setting assumes direct observations of the modeled reward, which may itself be binary. Our setting introduces an additional observation constraint: the learner observes only a thresholded version of a latent reward, while the identification objective remains defined through the latent reward functional. This requires relating uncertainty in reward gaps to the information retained after thresholding.

1.1.2 Communication-Constrained and Quantized Bandits

Distributed best arm identification studies how multiple agents can collaborate under limited communication, including tradeoffs between exploration time and communication rounds (Hillel et al., 2013; Tao et al., 2019). A complementary line of work considers the number of bits transmitted per observation. Hanna et al. (2022) develop reward-quantization schemes that preserve regret guarantees using few bits, while Vial et al. (2020) show that suitable encoding of reward histories allows one-bit feedback to asymptotically match the regret of full-feedback upper-confidence-bound policies. More directly related to pure exploration, Lau & Scarlett (2025) study best arm identification for quantile rewards under one-bit feedback and derive instance-dependent sample-complexity bounds. Our focus is identifying an arm according to a latent reward functional under a shared GLM and fixed arm-specific thresholds. The resulting question concerns how the information retained by this observation mechanism supports comparisons between structurally related arms.

1.1.3 One-Bit Parameter Estimation

One-bit compressed sensing studies signal recovery from quantized linear measurements (Boufounos & Baraniuk, 2008). Most closely related to our observation model, Shah et al. (2025) study multi-parameter maximum likelihood estimation for exponential-family GLMs with one-bit measurements. They characterize the Fisher information under censored and uncensored observations and establish regularity conditions for consistency and asymptotic normality of the estimator. These results describe parameter estimation under one-bit observations, but do not directly provide a sequential fixed-confidence best arm identification guarantee. Our work connects this censored-likelihood framework to gap-based exploration, where sampling must resolve the reward comparisons needed to certify an approximately optimal arm.

1.2 Our Contributions and Paper Outline

We study fixed-confidence best arm identification in generalized linear models with one-bit observations and fixed arm-specific thresholds. Our main contributions are as follows:

- **Gap-based exploration from one-bit observations.** We propose 1BIT-GLGAPE, which estimates the shared parameter θ^* using the Bernoulli likelihood of the thresholded observations and constructs confidence sets directly for pairwise latent reward gaps. At each round, the algorithm identifies the empirical best arm and its most competitive challenger, then uses an ℓ_1-based allocation and tracking rule to target uncertainty in their comparison. It stops when the gap confidence bounds certify that the recommended arm is ϵ-optimal.

- **Information-dependent sample complexity.** Our upper-bound analysis and local asymptotic lower-bound analysis identify the common instance-dependent hardness

$$\mathcal{H}_\epsilon(\theta^*, \gamma) = \sum_{k \in \mathcal{K}_\epsilon} \frac{1}{(\Delta_k - \epsilon)^2 \mathcal{C}_k(\theta^*, \gamma_k)}.$$

  The coefficient $\mathcal{C}_k(\theta^*, \gamma_k)$ captures the directional information retained after thresholding, making the effect of one-bit observations explicit in the sample-complexity bounds. Under the stated assumptions, we establish (ϵ, δ)-PAC correctness and an upper bound on the expected stopping time of 1BIT-GLGAPE. We complement this guarantee with a local asymptotic lower bound based on the change-of-measure framework of Kaufmann et al. (2016). The bounds share the displayed hardness quantity, with differences given by logarithmic factors and the stated regularity and design constants.

The remainder of the paper is organized as follows. Section 2 introduces the one-bit generalized linear model and the (ϵ, δ)-PAC objective. Section 3 presents 1BIT-GLGAPE and its gap-selection and arm-selection subroutines. Section 4 gives the correctness guarantee and sample-complexity analysis. Section 5 presents numerical experiments. Section 6 concludes and discusses future directions. Supporting lemmas and detailed proofs are provided in the appendix.

2 PROBLEM SETUP

Consider that we find the best arm in a set of K available arms. Let $[K] = 1, 2, 3, ..., K$ denote the set of possible arms. $V_k \in \mathbb{R}^{d \times m}$ is a known matrix representing the pre-determined or observable environmental features associated with arm k. This matrix is available to the decision maker and encapsulates all available information regarding the arm.

To model the reward generation process, we assume that the latent reward is modeled by a random variable Y taking values in a sample space $\mathcal{Y}$ that belongs to the exponential family. The probability density function of the reward is parameterized by a general parameter vector $\theta \in \Theta \subset \mathbb{R}^m$. Specifically, for an arm with feature matrix V_k, its natural parameter is defined as $\eta_{k,\theta} = V_k\theta \in \mathbb{R}^d$, the general density fuction is formulated as:

$$p_Y(y; \theta) = h(y) \exp\left(\langle V_k\theta, T_y\rangle - \phi(V_k\theta)\right), \tag{1}$$

with equation 1, we have that:

- $h(y) : \mathcal{Y} \to (0, +\infty)$ is the base measure.
- $T_y : \mathcal{Y} \to \mathbb{R}^d$ is the d-dimensional sufficient statistic vector.
- $\phi : \mathbb{R}^d \to \mathbb{R}$ is the log-partition function.

In our sequential setting, the environment operates based on a fixed, unknown true parameter vector $\theta^* \in \Theta$. At each time step $t = 1, 2, .., n$, denoted by $\mathcal{K} = [K]$, the agent selects an action $a_t \in \mathcal{K}$. The environment then generates a latent, unobservable reward Y_t drawn strictly from the true underlying distribution $p_Y(y; \theta^*)$.

Instead of observing the exact real-valued reward Y_t, the agent only receives a 1-bit censored measurement $G_t \in \{0, 1\}$ based in a specific thresholding mechanism:

$$G_t = \begin{cases} 1 & \text{if } Y_t \le \gamma_k, \\ 0 & \text{if } Y_t > \gamma_k, \end{cases} \tag{2}$$

where γ_k is a fixed threshold and an interior point of $\mathcal{Y}$. Let $\mathcal{Y}(1) = \{y \in \mathcal{Y} : y \le \gamma_k\}$ and $\mathcal{Y}(0) = \{y \in \mathcal{Y} : y > \gamma_k\}$. For any general parameter θ, the probability distribution of the 1-bit observation is mathematically defined as:

$$P_{G_t}(g; \theta) = \Pr(G_t = g \mid \theta) = \int_{\mathcal{Y}(g)} p_Y(y; \theta)\, dy. \tag{3}$$

Consequently, the actual observed sequence of feedback is generated according to $P_{G_t}(g, \theta^*)$. The agent uses these 1-bit observations to estimate the unknown parameter, typically yielding an estimator $\hat{\theta}_t$ at each time step.

To formulate the BAI objective, we evaluate the arms based on the true parameter θ^*. The true expected sufficient statistic vector of arm k is given by:

$$\mu_k = \mathbb{E}[T_{Y_k} \mid \theta^*] = \nabla_\eta \phi(V_k \theta^*) \in \mathbb{R}^d. \tag{4}$$

Equation 4 was proven through lemma 22 in Appendix. Since the reward of each arm is the expected sufficient statistic $\mu_k(\theta^*) = \mathbb{E}_{\theta^*}[T_{Y_k}] \in \mathbb{R}^d$, which may be multi-dimensional, we define a scalar reward functional $r : \mathbb{R}^d \to \mathbb{R}$ that captures the quantity of interest. Specifically, $e^\top \in \mathbb{R}^d$ denote the projection onto the first component, and $\|e\|_2 = 1$. The true optimal arm is defined as:

$$a^* = \arg\max_{k \in \mathcal{K}} r(\mu_k(\theta^*)) = \arg\max_{k \in \mathcal{K}} e^\top \mu_k(\theta^*) \tag{5}$$

which in the GLM parameterisation reduces to:

$$a^* = \arg\max_{k \in \mathcal{K}} e^\top \nabla_\eta \phi(V_k \theta^*) \tag{6}$$

An exploration strategy is defined as $\mathcal{S} = (\mathcal{A}, \tau)$, where $\mathcal{A}$ is a sequential sampling rule mapping historical observations $\mathcal{D}_{t-1} = \{(a_s, G_s)\}_{s=1}^{t-1}$ to the current action a_t, and τ is the stopping time. Upon stopping at time τ, the strategy outputs a declared optimal arm $\hat{a}_\tau$.

Assumption 1. (**Boundedness and Regularity.**) Let $\mathcal{V} = \{V_k\}_{k \in \mathcal{K}}$ denote the set of available feature matrices. We assume that the unknown true parameter vector is bounded, meaning there exists a constant $S > 0$ such that $\|\theta\|_2 \leq S$. Consequently, the true natural parameter satisfies $\|\eta_{k,\theta}\| \leq \|V_k\|_{\text{op}} S \leq L.S$ (with $\|V_k\|_{\text{op}} \leq L$ for all $k \in \mathcal{K}$), where $\|V_k\|_{\text{op}}$ denotes the largest singular value. Additionally, the mapping of the expected sufficient statistic $\mu(\eta) = \nabla_\eta \phi(\eta)$ is continuously differentiable and Lipchitz continuous with a constant $k_\mu \geq \|\nabla^2 \phi(\eta)\|_{\text{op}}$ for all η. Finally, the Hessian matrix satisfies:

$$c_\mu = \inf_{\|\theta\|_2 \leq S} \lambda_{\min}(\nabla^2_\eta \phi(V_k \theta)) > 0, \quad \text{for all} \quad k \in \mathcal{K}$$

Definition 2. Given $\epsilon > 0$ and $\delta \in (0, 1)$, an exploration strategy $\mathcal{S} = (\mathcal{A}, \tau)$ is said to be (ϵ, δ)-PAC if

$$\Pr\big(e^\top \mu_{a_*}(\theta^*) - e^\top \mu_{\hat{a}_\tau}(\theta^*) \geq \epsilon\big) \leq \delta \tag{7}$$

where $e \in \mathbb{R}^d$ is a fixed contrast vector specifying the scalar reward functional of interest, ϵ is the tolerance, and $1 - \delta$ is the confidence level. The objective is to design an (ϵ, δ)-PAC strategy that minimises the expected stopping time $\mathbb{E}[\tau]$.

Definition 3. Let $\mu_k = \nabla_\eta \phi(V_k \theta)$ be the true expected sufficient statistic vector of arm k. For any two arms $i, j \in \mathcal{K}$, we define their corresponding vector gap as $\Delta(i, j) = e^\top(\mu_i(\theta^*) - \mu_j(\theta^*))$. Furthermore, the optimal scalar gap Δ_i associated with an arm $i \in \mathcal{K}$ is mathematically defined as:

$$\Delta_i = \begin{cases} r(\mu_{a^*}(\theta^*)) - r(\mu_i(\theta^*)) = e^\top(\mu_{a^*}(\theta^*) - \mu_i(\theta^*)) & \text{if } i \neq a^*, \\ r(\mu_{a^*}(\theta^*)) - \max_{j \neq a^*} r(\mu_j(\theta^*)) = e^\top(\mu_{a^*}(\theta^*) - \max_{j \neq a^*} \mu_j(\theta^*)) & \text{if } i = a^*. \end{cases}$$

3 The Proposed Algorithm

In this section, we propose an exploration strategy designed to identify the best arm under censored data conditions, where observations are strictly limited to binary outcomes (0 or 1), it means 1-bit feedback. This exploration strategy should reduce uncertainty in the most competitive reward-gap direction. However, the information accumulates according to the censored Fisher matrix rather than the uncensored design matrix.

Specifically, Algorithm 1 is the central control loop. At each round, it invokes SELECTGAP (Algorithm 2) to identify the most uncertain arm pair and an exploratory direction, followed by SELECTARM (Algorithm 3) to select the physical arm whose observations best realize that direction.

The outer loop in Algorithm 1 begins with random warm-up pulls to avoid an undefined MLE and a singular Fisher information matrix. It then alternates between updating the parameter estimate from 1-bit observations, evaluating gap confidence bounds, checking the stopping rule, and choosing the next arm. If the maximal pessimistic gap is at most ϵ, it returns the empirical best arm; otherwise it continues with a Newton–Raphson update and a new exploratory action.

3.1 MLE ESTIMATOR UNDER 1-BIT DATA

From the sequentially observed data samples, we seek to estimate the m-dimensional vector θ using the MLE method. Let t denote the current time step, s denote a past time step, and a_s denote the specific arm played at time s. Let $G_s \in \{0, 1\}$ be the random binary variable observed at time step s. The MLE at time t is defined as follows:

$$\hat{\theta}_t = \arg\max_{\theta \in \Theta} \ell_t \left(\theta; \{G_s\}_{s=1}^t\right) \tag{8}$$

By substituting the natural parameter $\eta_{s,\theta} = V_{a_s}\theta$ directly into the formulation, the cumulative log-likehood function ℓ_t based on Bernoulli observations is given by:

$$\ell_t \left(\theta; \{G_s\}_{s=1}^t\right) = \sum_{s=1}^{t} \left[\phi(V_{a_s}\theta; G_s) - \phi(V_{a_s}\theta)\right], \tag{9}$$

where $\phi(V_{a_s}\theta)$ is the standard log-partition function of the uncensored exponential family, and $\phi(V_{a_s}\theta, G_s)$ represents the conditional log-partition function restricted by the 1-bit observation G_s.

We utilize the Fisher Information Matrix (FIM) results from previous studies to support further analyses. The FIM contribution of an arm k is defined using the covariance of the expected sufficient statistic:

$$J_k(\theta) = V_k^\top \operatorname{Cov}\big(\mathbb{E}[T_{Y_k} \mid G_k]\big)\, V_k. \tag{10}$$

Specifically, the covariance expands to:

$$\operatorname{Cov}(\mathbb{E}[T_{Y_k} \mid G_k]) = p_k(1 - p_k)\,(\mu_1^k - \mu_0^k)(\mu_1^k - \mu_0^k)^\top, \tag{11}$$

where $\mu_g^k = \mathbb{E}[T_{Y_k} \mid G_k = g] \in \mathbb{R}^d$.

Result from equation 11 was proven in lemma 20 in Appendix.

In our algorithm, we update the estimate $\hat{\theta}$ using the Newton-Raphson method. At iteration t, after selecting arm a_t and observing the binary feedback G_t, the Fisher Information Matrix J_t is updated by the information gain:

$$J_t = J_{t-1} + V_{a_t}^\top \operatorname{Cov}\left(\mathbb{E}[T_{Y_{a_t}} \mid G_{a_t}]\right) V_{a_t}, \tag{12}$$

where the censored covariance term is explicitly evaluated as:

$$\operatorname{Cov}\left(\mathbb{E}[T_{Y_{a_t}} \mid G_{a_t}]\right) = p_{a_t}(1 - p_{a_t})(\mu_1^{a_t} - \mu_0^{a_t})(\mu_1^{a_t} - \mu_0^{a_t})^\top. \tag{13}$$

After, using the updated inverse Fisher Information Matrix J_t^{-1}, the parameter estimate is updated via:

$$\hat{\theta}_t = \hat{\theta}_{t-1} + J_t^{-1}(\hat{\theta}_{t-1})\nabla \ell_t(\hat{\theta}_{t-1}). \tag{14}$$

3.2 Confidence Sets for Gaps

We build confidence sets directly for the reward gaps instead of individual arm rewards. Let $\lambda_0 > 0$ be the minimum eigenvalue of the initial matrix J_{E+1}, and L be the Lipschitz parameter. We define:

$$\kappa = \sqrt{3 + 2\log(1 + 2L^2/\lambda_0)}. \tag{15}$$

Let $\Delta_t(i,j)$ be an estimate for the gap between arms $i, j \in \mathcal{K}$ at period t, let $\Delta_t(i,j) = e^\top \left(\mu_i(\hat{\theta}_t) - \mu_j(\hat{\theta}_t)\right)$. Given a confidence level $\delta \in (0,1)$, we propose C_t is a time-varying quantity that scales the width of the confidence sets for all pairs of arms.

$$C_t = \alpha\sqrt{2m\log t \log(\frac{\pi^2 m t^2}{6\delta})} \tag{16}$$

where $\alpha = \frac{2\kappa R}{c_\mu}$, and m is the dimension of θ.

We consider the following confidence set for the gap between any two arms $i, j \in \mathcal{K}$ based on the observations prior to period t:

$$C_t(i,j) = \left\{\Delta \in \mathbb{R} : |\Delta - \Delta_t(i,j)| \le C_t \max_{c,c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_j)^\top e\|_{J_t^{-1}}\right\}. \tag{17}$$

The confidence set is centered around the estimated gap. Its total width, denoted as $\beta_t(i,j)$ is defined as:

$$\beta_t(i,j) = C_t \max_{c,c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_j)^\top e\|_{J_t^{-1}} \tag{18}$$

3.3 Algorithm

Algorithm 1 gives the complete outer loop; the two selection subroutines are specified in Appendix A.

Algorithm 1 1BIT-GLGAPE

Require: $\{V_k\}_{k=1}^K$, tolerance ϵ, confidence δ, initial pulls E, parameters α, S, L, thresholds $\{\gamma_k\}_{k=1}^K$
Ensure: approximately best arm $\hat{a}_\tau$

```
Pull E > d random arms, gather D_E = {(V_{a_s}, G_s)}_{s=1}^E, initialize pull counts T_k(E)
Compute initial MLE θ̂_E and FIM J_E = Σ_{s=1}^E V_{a_s}^⊤ Cov(E[T_{Y_{a_s}} | G_{a_s}]) V_{a_s}
for t = E+1, E+2, ... do
    (i_t, j_t, B(t), z_t) ← SELECTGAP({V_k}, D_{t-1})
    if B(t) ≤ ε then
        return i_t as the approximately best arm
    end if
    a_t ← SELECTARM({V_k}, T_k(t-1), z_t)
    Play arm a_t, observe G_t ∈ {0,1}
    Newton–Raphson update: θ̂_t = θ̂_{t-1} + J_t^{-1}(θ̂_{t-1}) ∇ℓ_t(θ̂_{t-1})
    D_t ← D_{t-1} ∪ {(V_{a_t}, G_t)},   J_t = J_{t-1} + V_{a_t}^⊤ p_{a_t}(1-p_{a_t})(μ_1^{a_t} − μ_0^{a_t})(μ_1^{a_t} − μ_0^{a_t})^⊤ V_{a_t}
    T_k(t) = T_k(t-1) + I{a_t = k}
end for
```

4 Theoretical Guarantees

Both upper and lower sample complexity bounds are governed by the same instance hardness $\mathcal{H}_\epsilon(\theta^*, \gamma)$, where the information loss from 1-bit quantization is captured directly by the directional coefficient $\mathcal{C}_k(\theta^*, \gamma_k)$.

Specifically, for each suboptimal arm k, let $\Delta_k = e^\top(\mu_{a^*}(\theta^*) - \mu_k(\theta^*))$ and define $\mathcal{K}_\epsilon = \{k \neq a^* : \Delta_k \geq \epsilon\}$. The instance-dependent hardness is

$$\mathcal{H}_\epsilon(\theta^*, \gamma) = \sum_{k \in \mathcal{K}_\epsilon} \frac{1}{(\Delta_k - \epsilon)^2 \mathcal{C}_k(\theta^*, \gamma_k)}, \tag{19}$$

where $\mathcal{C}_k(\theta^*, \gamma_k)$ is the directional information coefficient induced by the censored Fisher information (Appendix B.3).

Theorem 4 ((ϵ, δ)-PAC correctness)**.** *Under the boundedness, regularity, and information-matrix conditions stated above, let $\mathcal{E}$ be the uniform confidence event satisfying $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. On the event $\mathcal{E}$, the algorithm* 1BIT-GLGAPE *terminates at a finite stopping time $\tau < \infty$ and outputs an arm $\hat{a}_\tau$ that deterministically satisfies:*

$$e^\top \mu_{a^*}(\theta^*) - e^\top \mu_{\hat{a}_\tau}(\theta^*) < \epsilon. \tag{20}$$

Consequently, 1BIT-GLGAPE *is (ϵ, δ)-PAC, guaranteeing that with probability at least $1 - \delta$:*

$$\mathbb{P}\left(e^\top \mu_{a^*}(\theta^*) - e^\top \mu_{\hat{a}_\tau}(\theta^*) < \epsilon\right) \geq 1 - \delta. \tag{21}$$

The proof establishes that on the high-probability event $\mathcal{E}$, the pairwise gap estimator remains bounded within the MLE confidence sets, ensuring that the stopping condition $B(\tau) \leq \epsilon$ guarantees ϵ-optimality upon termination; see Appendix B.1.

Proposition 5 (Asymptotic Lower Bound)**.** *For $\delta \in (0, 0.15)$, any (ϵ, δ)-PAC algorithm satisfies:*

$$\mathbb{E}[\tau] \geq 2 \log\left(\frac{1}{2.4\delta}\right) \mathcal{H}_\epsilon(\theta^*, \gamma). \tag{22}$$

Theorem 6 (Tracking-based upper bound)**.** *Suppose the tracking allocation has positive design information and $\lambda_1 = \lambda_{\min}(\sum_k V_{\Delta,k}^\top V_{\Delta,k}) > 0$. Under the conditions in Appendix B.3, the manuscript derives*

$$\mathbb{E}[\tau] \leq \frac{2048\, \kappa^2 R^2 k_\mu^2 L^2 m d}{c_\mu^4 \lambda_1} \log^2\left(\frac{1}{\delta}\right) \mathcal{H}_\epsilon(\theta^*, \gamma). \tag{23}$$

Hence the displayed bounds share the same problem-dependent hardness, up to logarithmic factors and the stated regularity/design constants. Detailed spectral, allocation, and elimination-time calculations are in Appendix B.3.

5 NUMERICAL RESULTS

In this section, we test the proposed algorithm in various scenarios. All scenarios considered involve a logistic link function. We explore the algorithm with E random arms to ensure that J_{E+1} is nonsingular. The implementation details and techniques are detailed in the appendix D.

Since our algorithm is significantly influenced by the feature dimension of V_k, we present two experimental scenarios in this evaluation across various values of ϵ and feature dimensions d_{feature}. In both cases, we fix $\delta = 0.05$.

Measuring the stopping time for the number of $d-$features. We check with the number of features being $[10, 15, 20, 25, 30]$. To value of ϵ and δ, we fix similar to the test scenario for the number of arms. Besides, we fix with $K = 100$ arms. The result is presented in Figure 1.

Fixing the feature dimension $d = 10$ and the number of arms $K = 100$, we plot the stopping time of various ϵ. As expected, a smaller tolerance requires a larger amount of exploration. The result is presented in Figure 2.

To the best of our knowledge, there are currently no directly comparable baselines for our proposed algorithm. While prior literature has addressed the 1-bit best-arm identification problem, existing methods operate outside linear or generalized linear settings and fail to apply d-dimensional features. So, we present the standalone performance of our algorithm without direct baseline comparisons.

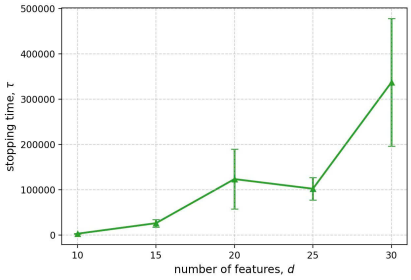

Figure 1: The amount of exploration required by the proposed algorithm with tolerance $\epsilon = 0.1$, $\delta = 0.05$, and K = 100 arms.

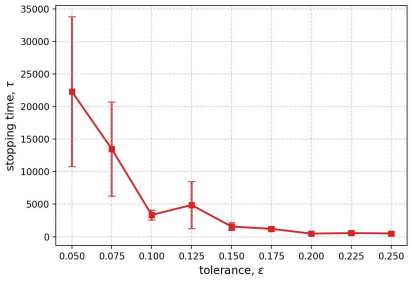

Figure 2: The amount of exploration required by the proposed algorithm with $\delta = 0.05$, $K = 100$ arms, and $d = 10$ features.

6 Conclusion

Our study examines best arm identification in generalized linear bandits under 1-bit feedback, demonstrating that transitioning from full to binary observations imposes a precise, directional sample complexity cost in parameter space. To model this, we establish the censored Fisher matrix as the core curvature object that unifiedly governs maximum likelihood estimation, confidence widths, and information coefficients. When observing binary feedback instead of full rewards, the information available about θ^* per pull of arm k is no longer captured by the uncensored design contribution, but by the censored Fisher matrix revealed by the binary outcome.

Furthermore, the censored information coefficients $\mathcal{C}_k(\theta^*, \gamma_k)$ in the denominator pinpoint precisely where this sample complexity price is paid: they measure how much a single 1-bit observation of arm k reduces uncertainty along the reward-gap direction. Consequently, an arm that is difficult to distinguish from the optimal arm under full observation becomes strictly harder to identify under 1-bit feedback. Finally, because the thresholds τ_k in this work remain fixed after a short initialization phase and are never updated, this framework naturally raises a concrete follow-up question: whether threshold selection itself can be optimized jointly with arm allocation.

AI use statement

In preparing this revision, generative AI tools were used to reorganize and condense existing mathematical statements for the main-paper page limit, edit prose for readability, format references, and convert the manuscript to the official ICLR 2027 LaTeX template. They were not used in this formatting process to generate synthetic data, develop the underlying theoretical model, supply ingredients for proofs, write or alter proofs, propose hypotheses, design experiments, implement the research method, translate content, prepare or analyze datasets, or interpret results; dataset- and experiment-related tasks were not applicable to this theoretical manuscript. The AI-assisted formatting and restructuring changes were checked against the supplied source manuscript. Before submission, the authors will review the complete revision and update this statement to cover any

other use of generative AI. The authors take responsibility for the final content, including text, claims, proofs, citations, and artifacts produced with AI assistance.

Reproducibility statement

This is a theoretical study. The problem specification, algorithmic steps, assumptions, and main guarantees are stated in the paper, and supporting derivations and proofs are provided in the appendix.

References

Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pp. 16–21. IEEE, 2008.

Jun Fang, Yumeng Liu, Hongbin Li, and Shaoqian Li. One-Bit quantizer design for multisensor GLRT fusion. *IEEE Signal Processing Letters*, 20(3):257–260, 2013.

Osama A Hanna, Lin Yang, and Christina Fragouli. Solving Multi-Arm Bandit using a few bits of communication. In *International Conference on Artificial Intelligence and Statistics*, pp. 11215–11236. PMLR, 2022.

Eshcar Hillel, Zohar S Karnin, Tomer Koren, Ronny Lempel, and Oren Somekh. Distributed exploration in Multi-Armed Bandits. *Advances in Neural Information Processing Systems*, 26, 2013.

Nicholas P Jewell and Mark van der Laan. Current status data: Review, recent developments and open problems. *Handbook of statistics*, 23:625–642, 2003.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in Multi-Armed Bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Abbas Kazerouni and Lawrence M Wein. Best arm identification in Generalized Linear Bandits. *Operations Research Letters*, 49(3):365–371, 2021.

Ritesh Kumar and Shashank Vatedka. Communication-constrained distributed mean estimation of log-concave distributions. In *2023 National Conference on Communications (NCC)*, pp. 1–6. IEEE, 2023.

Ivan Lau and Jonathan Scarlett. Quantile multi-armed bandits with 1-bit feedback. In Gautam Kamath and Po-Ling Loh (eds.), *Proceedings of The 36th International Conference on Algorithmic Learning Theory*, volume 272 of *Proceedings of Machine Learning Research*, pp. 664–699. PMLR, 24–27 Feb 2025.

Jaimin Shah, Martina Cardone, Cynthia Rush, and Alex Dytso. Generalized Linear Models with 1-Bit measurements: Asymptotics of the maximum likelihood estimator. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Chao Tao, Qin Zhang, and Yuan Zhou. Collaborative learning with limited interaction: Tight bounds for distributed exploration in Multi-Armed Bandits. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 126–146. IEEE, 2019.

Daniel Vial, Sanjay Shakkottai, and R Srikant. One-Bit feedback is sufficient for upper confidence bound policies. *arXiv preprint arXiv:2012.02876*, 2020.

A Detailed Algorithmic Subroutines

This section records the two subroutines used by Algorithm 1 and the detailed implementation discussion.

Algorithm 2 SelectGap

Require: $\{V_k\}_{k\in\mathcal{K}}$, History $\mathcal{D}_{t-1}$ with 1-bit observations.
Ensure: Gap to be explores $(i_t, j_t, B(t), z_t)$.

1: Compute the Maximum Likehood Estimation (MLE) $\hat{\theta}_t$ from 1-bit observation history $\mathcal{D}_{t-1}$ by maximizing the Bernoulli log-likehood, and compute the cumulative Fisher Information Matrix (FIM) J_t.

$$\hat{\theta}_t = \arg\max_{\theta} \sum_{s\in D_{t-1}} [G_{a_s} \log p_{a_s}(\gamma_{a_s}; V_{a_s}, \theta) + (1 - G_{a_s}) \log (1 - p_{a_s}(\gamma_{a_s}; V_{a_s}, \theta))]$$

$$J_t = \sum_{s\in D_{t-1}} V_{a_s}^\top \text{Cov}(\mathbb{E}[T_{Y_{a_t}}|G_{a_t}])V_{a_s}$$

2: Find candidate best arm:

$$i_t \leftarrow \arg\max_{k\in\mathcal{K}} r(\mu_k(\hat{\theta}_t) = \arg\max_{k\in\mathcal{K}} e^\top(\nabla_\eta \phi(V_k\hat{\theta}_t)).$$

3: Find most challenging arm:

$$\Delta_t(j, i_t) \leftarrow e^\top(\mu_{i_t}(\hat{\theta}_t) - \mu_j(\hat{\theta}_t))$$
$$j_t \leftarrow \arg\max_{j\in\mathcal{K},\ j\neq i_t} [\Delta_t(j, i_t) + \beta_t(i_t, j)]$$

4: Find optimized contrast vector:

$$c_1, c_2 = \arg\max_{c,c'\in[c_\mu,k_\mu]} \left\|(cV_{i_t} - c'V_{j_t})^\top e\right\|_{J_t^{-1}}$$

$$z_t \leftarrow [(c_1V_{i_t} - c_2V_{j_t})^\top e] \in \mathbb{R}^m$$

5: Compute bounding gap:

$$B(t) \leftarrow \Delta_t(j_t, i_t) + \beta_t(i_t, j_t)$$

6: **return** $(i_t, j_t, B(t), z_t)$

Algorithm 3 SelectArm

Require: $\{V_k\}_{k\in\mathcal{K}}$, pull counts $T_k(t-1)$, target contrast vector z_t.
Ensure: Arm to be played a_t.

1: Solve the linear program:

$$w^* \leftarrow \arg\min_{w\in\mathbb{R}^K} \|w\|_1 \quad \text{s.t.} \quad \sum_{k=1}^K w_k V_k^\top = z_t$$

2: For each $k \in \mathcal{K}$, compute sampling proportions:

$$p_k(z_k) \leftarrow \frac{|w_k^*|}{\sum_{a=1}^K |w_a^*|}$$

3: Select the action:

$$a_t \leftarrow \arg\min_{k:p_k>0} \frac{T_k(t)}{p_k(z_k)}$$

4: **return** a_t

A.1 SELECT GAP

The algorithm starts its main loop after initial phase of playing $E > m$ random arms to ensure the empirical covariance matrix is nonsingular. At any subsequent period $t > E$, utilizing the paramters $\hat{\theta}_t$ received from the main loop, the algorithm first finds the empirically best arm:

$$i_t = \arg\max_{k \in \mathcal{K}} e^\top \mu_k(\hat{\theta}_t)$$

Then, it verifies whether arm i_t is within an ϵ distance of the true optimal arm using a pessimistic approach. Specially, it finds another arm j_t that is the most advantageous over arm i_t within the gap confidence bounds:

$$j_t = \arg\max_{j \in [K], j \neq i_t} [\Delta_t(j, i_t) + \beta_t(i_t, j)]$$

This pessimistic gap consists of the estimated gap and the uncertainty width. If the maximum pessimistic gap is less than ϵ, the algorithm stops and declares i_t as the optimal arm. Otherwise, it constructs the target directional vector z_t to guide the arm selection process to maximally reduce the uncertainty component:

$$c_1, c_2 = \arg\max_{c, c' \in [c_\mu, k_\mu]} \|(cV_{i_t} - c'V_{j_t})^\top e\|_{J_t^{-1}}$$

$$z_t = [(c_1 V_{i_t} - c_2 V_{j_t})^\top e]$$

A.2 SELECT ARM

With select arm, we use an idea of previous studies. Specific, in this subroutine, we utilize the available arm features V_k, the historical pull counts $T_k(t-1)$, and the target directional vector z_t. The objective is to find a specific physical arm a_t to play at time step t that optimally reduces the uncertainty $\|z_t\|_{J_t^{-1}}$.

Let $p_k(z_t)$ be the ideal frequency of pulling arm k. We define it mathematically as:

$$p_k(\boldsymbol{z}_t) = \frac{|w_k^*|}{\sum_{a=1}^K |w_a^*|}$$

where the weight vector $w^* \in \mathbb{R}^K$ is the solution to the sparsity-including L_1 linear program:

$$\min_{w \in \mathbb{R}^K} \|w\|_1 \quad \text{s.t.} \quad \sum_{k=1}^K w_k V_k^\top = z_t$$

To physically minimize the uncertainty in practice the arm a_t chosen to be played is determined by a tracking formula that penalizes over-sampled arms and prioritizes under sampled ones:

$$a_t = \arg\min_{k \in \mathcal{K}: p_k(\boldsymbol{y}_t) > 0} \frac{T_k(t-1)}{p_k(\boldsymbol{z}_t)}$$

B DETAILED THEORETICAL ANALYSIS

B.1 CORRECTNESS

Definition 7. Some definitions for proof correctness theorem:

- **Confidence Set Definition.** The confidence set at time step t, denoted by $\mathcal{E}_t$, is defined as the set of parameters θ such that the squared norm of the estimation error, weighted by the information matrix J_t, is bounded by β_t. Mathematically, this is expressed as $\mathcal{E} = \{\theta : \|\hat{\theta}_t - \theta\|_{J_t}^2 \leq \beta_t\}$.

- **Threshold Parameter.** The threshold parameter β_t is defined using the quantile function $\chi^2_{m,1-\delta_t}$.
- **Error Probability.** The time varying error probability δ_t, is set to $\frac{6\delta}{\pi^2 t^2}$.
- **Confidence Scale.** The scaling factor for the confidence intervals, denoted as C_t, is formulated as $C_t = \frac{2\kappa R}{c_\mu}\sqrt{2m\log t \cdot \log\left(\frac{\pi^2 m t^2}{6\delta}\right)}$.
- **Confidence Width.** The width of the confidence interval between two arms i and k, denoted as $\beta_t(i,k)$, is defined as

$$\beta_t(i,k) = C_t \max_{c,c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}}$$

.

Lemma 8 (Validity of the Confidence Set)**.** *Let $\mathcal{G} = \{\theta^* \in \mathcal{E}_t, \forall t \geq E\}$ denote the good event that the true parameter vector θ^* remains within the confidence set $\mathcal{E}_t$ for all time steps $t \geq E$. Then, the probability of event $\mathcal{G}$ satisfies:*

$$\mathbb{P}(\mathcal{G}) \geq 1 - \delta. \tag{24}$$

Proof. By the asymptotic normality of the Maximum Likelihood Estimator (MLE) specifically referencing Theorem 1 of Shah et al (Shah et al., 2025), the failure probability of the confidence set at any specific round t is bounded by:

$$\mathbb{P}(\theta^* \notin \mathcal{E}_t) \leq \delta_t = \frac{6\delta}{\pi^2 t^2}.$$

Applying a union bound over all $t \geq E$ and using the Euler-Basel series identity $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$, the probability of the complementary event (failure of the confidence set) is bounded as:

$$\mathbb{P}(\mathcal{G}^c) = \mathbb{P}\left(\exists t \geq E : \theta^* \notin \mathcal{E}_t\right) \leq \sum_{t=E}^{\infty} \delta_t \leq \sum_{t=1}^{\infty} \frac{6\delta}{\pi^2 t^2} = \frac{6\delta}{\pi^2} \cdot \frac{\pi^2}{6} = \delta.$$

Applying the complement rule of probability yields $\mathbb{P}(\mathcal{G}) = 1 - \mathbb{P}(\mathcal{G}^c) \geq 1 - \delta$. □

Lemma 9 (Gap Concentration Bound)**.** *On the good event $\mathcal{G}$, for all arms $k \in \mathcal{K}$ and all time steps $t \geq E$, the absolute difference between the empirical gap $\Delta_t(k, i_t) = e^\top[\mu_{i_t}(\hat{\theta}_t) - \mu_k(\hat{\theta}_t)]$ and the true gap $\Delta_t^*(k, i_t) = e^\top[\mu_{i_t}(\theta^*) - \mu_k(\theta^*)]$ is bounded by the confidence width $\beta_t(i_t, k)$:*

$$|\Delta_t(k, i_t) - \Delta_t^*(k, i_t)| \leq \beta_t(i_t, k), \quad \forall k \in \mathcal{K}, \forall t \geq E. \tag{25}$$

Proof. Expanding the difference between the empirical gap and the true gap yields:

$$\begin{aligned}\Delta_t(k, i_t) - \Delta_t^*(k, i_t) &= e^\top[\mu_{i_t}(\hat{\theta}_t) - \mu_k(\hat{\theta}_t)] - e^\top[\mu_{i_t}(\theta^*) - \mu_k(\theta^*)] \\ &= e^\top[\mu_{i_t}(\hat{\theta}_t) - \mu_{i_t}(\theta^*)] - e^\top[\mu_k(\hat{\theta}_t) - \mu_k(\theta^*)].\end{aligned} \tag{26}$$

By the Mean Value Theorem, there exist intermediate parameter vectors $\bar{\theta}_{i_t}$ and $\bar{\theta}_k$ on the line segment connecting $\hat{\theta}_t$ and θ^* such that:

$$\Delta_t(k, i_t) - \Delta_t^*(k, i_t) = e^\top \left(c_{i_t} V_{i_t} - c_k V_k\right)(\hat{\theta}_t - \theta^*),$$

where $c_{i_t} = \nabla^2_\eta \phi(V_{i_t}\bar{\theta}_{i_t}) \in [c_\mu, k_\mu]$ and $c_k = \nabla^2_\eta \phi(V_k \bar{\theta}_k) \in [c_\mu, k_\mu]$.

Applying the Cauchy-Schwarz inequality with respect to the design matrix J_t, we obtain:

$$|\Delta_t(k, i_t) - \Delta_t^*(k, i_t)| \leq \left\|(c_{i_t} V_{i_t} - c_k V_k)^\top e\right\|_{J_t^{-1}} \|\hat{\theta}_t - \theta^*\|_{J_t}.$$

Under event $\mathcal{G}$, the estimation error satisfies $\|\hat{\theta}_t - \theta^*\|_{J_t} \leq \sqrt{\beta_t}$. Taking the maximum over $c, c' \in [c_\mu, k_\mu]$ and recalling from Definition 7 that $C_t \geq \sqrt{\beta_t}$, we directly arrive at:

$$|\Delta_t(k, i_t) - \Delta_t^*(k, i_t)| \leq C_t \max_{c,c' \in [c_\mu, k_\mu]} \left\|(cV_{i_t} - c'V_k)^\top e\right\|_{J_t^{-1}} = \beta_t(i_t, k). \tag{27}$$

□

Theorem 10 ((ϵ, δ)-PAC Correctness)**.** *The algorithm* 1BIT-GLGAPE *satisfies the* (ϵ, δ)*-Probably Approximately Correct (PAC) property. Formally, on the good event* $\mathcal{G}$ *(which occurs with probability at least* $1 - \delta$*), the algorithm terminates at a finite stopping time* $\tau < \infty$ *and outputs an arm* $\hat{a}_\tau$ *satisfying:*

$$e^\top \mu_{a^*}(\theta^*) - e^\top \mu_{\hat{a}_\tau}(\theta^*) \leq \epsilon. \tag{28}$$

Proof. We evaluate the error probability by conditioning on the good event $\mathcal{G}$. By the law of total probability, the overall error probability is bounded as:

$$\mathbb{P}(\text{error}) = \mathbb{P}(\text{error} \cap \mathcal{G}) + \mathbb{P}(\text{error} \cap \mathcal{G}^c) \leq \mathbb{P}(\text{error} \mid \mathcal{G}) + \mathbb{P}(\mathcal{G}^c).$$

From Lemma 8, the failure probability is bounded by $\mathbb{P}(\mathcal{G}^c) \leq \delta$. To establish PAC correctness, it suffices to demonstrate that $\mathbb{P}(\text{error} \mid \mathcal{G}) = 0$.

When the algorithm terminates at stopping time τ, the stopping condition $B(\tau) \leq \epsilon$ is satisfied, where $B(\tau) = \max_{k \neq i_\tau} \{\Delta_\tau(k, i_\tau) + \beta_\tau(i_\tau, k)\}$. Consequently, for an arbitrary arm $k \neq i_\tau$, the inequality holds:

$$\Delta_\tau(k, i_\tau) + \beta_\tau(i_\tau, k) \leq B(\tau) \leq \epsilon.$$

Assuming the empirically chosen arm $i_\tau \neq a^*$, we apply this stopping condition specifically to $k = a^*$:

$$\Delta_\tau(a^*, i_\tau) + \beta_\tau(i_\tau, a^*) \leq \epsilon. \tag{29}$$

Under event $\mathcal{G}$, Lemma 9 guarantees that the true gap $e^\top \mu_{a^*}(\theta^*) - e^\top \mu_{i_\tau}(\theta^*)$ is upper-bounded by the empirical gap plus the confidence width:

$$e^\top \mu_{a^*}(\theta^*) - e^\top \mu_{i_\tau}(\theta^*) \leq \Delta_\tau(a^*, i_\tau) + \beta_\tau(i_\tau, a^*). \tag{30}$$

Combining (29) and (30) yields:

$$e^\top \mu_{a^*}(\theta^*) - e^\top \mu_{\hat{a}_\tau}(\theta^*) \leq \epsilon.$$

Thus, conditioned on event $\mathcal{G}$, the selected arm $\hat{a}_\tau = i_\tau$ is deterministically ϵ-optimal, implying $\mathbb{P}(\text{error} \mid \mathcal{G}) = 0$. Consequently, $\mathbb{P}(\text{error}) \leq \delta$. □

B.2 Information-Theoretic Lower Bound

To begin analyzing lower bound, we use theorem Kaufmann et al. (2016). With $\delta < 1.5$, we have fomula:

$$\sum_k \mathbb{E}[T_k(\tau)] \, \mathrm{KL}_k(\theta^* \| \theta'_j) \geq \log \frac{1}{2.4\delta}. \tag{31}$$

Definition 11 (Hard arms and gap directions)**.** 1. $\mathcal{K}_\epsilon = \{\, k \neq a^* : \Delta_k \geq \epsilon \,\}$ The set of arms whose gap is *at least* ϵ compared to the *best arm* — that is, the arms that can be distinguished from a^*.

2. Let

$$(V_{a^*} - V_j)^\top e = V_{\Delta,j}^\top e = v_j \tag{32}$$

Lemma 12. *For each arm* $j \in \mathcal{K}_\epsilon$*, construct* θ'_j *closest to* θ^* *such that* j *becomes* ϵ*-optimal under* θ'_j *and* $\theta'_j = \theta^* + s_j V_{\Delta,j}^\top e$*, then*

$$s_j = \frac{\Delta_j(\theta^*) - \epsilon}{e^\top H_j V_{\Delta,j} V_{\Delta,j}^\top e} \tag{33}$$

with H_j *is Hessian of* $\nabla_\eta^2 \phi(V_j \theta^*)$

Proof. Let the gap of arm j under θ is

$$\Delta_j(\theta) = e^\top(\mu_{a^*}(\theta) - \mu_j(\theta)) = e^\top(\nabla_\eta\phi(V_{a^*}\theta) - \nabla_\eta\phi(V_j\theta)).$$

Let derivative with respect to θ at θ^*:

$$\nabla_\theta\Delta_j(\theta^*) = (V_{a^*} - V_j)^\top\nabla_\eta^2\phi(V_j\theta^*)e.$$

To θ'_j closer to θ^*, we require

$$\Delta_j(\theta'_j) = \varepsilon.$$

First-order expansion:

$$\begin{aligned}\Delta_j(\theta'_j) &\approx \Delta_j(\theta^*) + \nabla_\theta\Delta_j(\theta^*)^\top(\theta'_j - \theta^*) &(34)\\ &= \Delta_j(\theta^*) + \left(\nabla_{\Delta,j}^\top H_j e\right)^\top \left(s_j V_{\Delta,j}^\top e\right) &(35)\\ &= \Delta_j(\theta^*) - s_j e^\top H_j \nabla_{\Delta,j}\nabla_{\Delta,j}^\top e. &(36)\end{aligned}$$

(Eq 29: a negative sign is introduced to minimize the gap between θ^* and θ'_j) Hence:

$$s_j = \frac{\Delta_j(\theta^*) - \varepsilon}{c^\top H_j V_{\Delta,j} V_{\Delta,j}^\top c}.$$

□

Lemma 13. *With the value received is* $\{0, 1\}$,

$$\begin{aligned}\mathrm{KL}_k(\theta \,\|\, \theta') &= \sum_{g\in\{0,1\}} P_{G_k}(g;\theta)\log\frac{P_{G_k}(g;\theta)}{P_{G_k}(g;\theta')}\\ &= p_k(\theta)\log\frac{p_k(\theta)}{p_k(\theta')} + (1-p_k(\theta))\log\frac{1-p_k(\theta)}{1-p_k(\theta')} &(37)\end{aligned}$$

If $\|\theta' - \theta\|$ *small, we have second-order expansion of KL*

$$\mathrm{KL}_k(\theta \,\|\, \theta') = \frac{1}{2}(\theta' - \theta)^\top J_k(\theta)(\theta' - \theta) + O(\|\theta' - \theta\|^3). \tag{38}$$

Proof. Let $\delta p_k = p_k(\theta') - p_k(\theta)$. We have

$$\log\frac{p_k(\theta)}{p_k(\theta')} = -\log\left(1 + \frac{\delta p_k}{p_k(\theta)}\right) = -\frac{\delta p_k}{p_k(\theta)} + \frac{(\delta p_k)^2}{2p_k(\theta)^2} - \frac{(\delta p_k)^3}{3p_k(\theta)^3} + \cdots,$$

and

$$\log\frac{1-p_k(\theta)}{1-p_k(\theta')} = \frac{\delta p_k}{1-p_k(\theta)} + \frac{(\delta p_k)^2}{2(1-p_k(\theta))^2} + O((\delta p_k)^3).$$

So,

$$\begin{aligned}\mathrm{KL}_k(\theta||\theta') &= p_k(\theta)\left(-\frac{\delta p_k}{p_k(\theta)} + \frac{(\delta p_k)^2}{2p_k(\theta)^2}\right)\\ &\quad + (1-p_k(\theta))\left(\frac{\delta p_k}{1-p_k(\theta)} + \frac{(\delta p_k)^2}{2(1-p_k(\theta))^2}\right) + O((\delta p_k)^3)\\ &= -\delta p_k + \delta p_k + \frac{(\delta p_k)^2}{2p_k(\theta)} + \frac{(\delta p_k)^2}{2(1-p_k(\theta))} + O((\delta p_k)^3)\\ &= \frac{(\delta p_k)^2}{2p_k(\theta)(1-p_k(\theta))} + O((\delta p_k)^3). &\text{(KL1)}\end{aligned}$$

We also have

$$\delta p_k = \nabla_\theta p_k(\theta)^\top (\theta' - \theta) + O(\|\theta' - \theta\|^2).$$

Substitute into (KL1):

$$\begin{aligned}\mathrm{KL}_k(\theta||\theta') &= \frac{\left[\nabla_\theta p_k^\top (\theta' - \theta)\right]^2}{2p_k(\theta)(1 - p_k(\theta))} + O(\|\theta' - \theta\|^3) \\ &= \frac{1}{2}(\theta' - \theta)^\top \frac{\nabla_\theta p_k \nabla_\theta p_k^\top}{p_k(\theta)(1 - p_k(\theta))}(\theta' - \theta) + O(\|\theta' - \theta\|^3) \\ &= \frac{1}{2}(\theta' - \theta)^\top J_k(\theta)(\theta' - \theta) + O(\|\theta' - \theta\|^3).\end{aligned}$$

□

Consider arm $j \in \mathcal{K}_\epsilon$ and applying lemma 2 for $\mathrm{KL}_j(\theta^*||\theta'_j)$, we have

$$\mathrm{KL}_j(\theta^*\|\theta'_j) = \frac{1}{2}(\theta'_j - \theta^*)^\top J_j(\theta^*)(\theta'_j - \theta^*) + O(\|\theta'_j - \theta^*\|^3). \tag{39}$$

To calculate $\theta'_j - \theta^*$, we use lemma 1:

$$\theta'_j - \theta^* = s_j v_j = s_j V_{\Delta,j}^\top e. \tag{40}$$

So,

$$\begin{aligned}\mathrm{KL}_j(\theta^*\|\theta'_j) &= \frac{s_j^2}{2}(V_{\Delta,j}^\top e)^\top J_j(\theta^*)(V_{\Delta,j}^\top e) + O(\|\theta'_j - \theta^*\|^3) \\ &= \frac{s_j^2}{2} e^\top V_{\Delta,j} J_j(\theta^*) V_{\Delta,j}^\top e.\end{aligned}$$

Replace s_j from equal above:

$$\mathrm{KL}_j(\theta^*\|\theta'_j) = \frac{(\Delta_j - \varepsilon)^2}{2\left[e^\top H_j V_{\Delta,j} V_{\Delta,j}^\top e\right]^2} e^\top V_{\Delta,j} J_k(\theta^*) V_{\Delta,j}^\top e + O(\|\theta'_j - \theta^*\|^3).$$

Let

$$\mathcal{C}_j(\theta^*, \gamma_j) = \frac{e^\top V_{\Delta,j} J_j(\theta^*) V_{\Delta,j}^\top e}{[e^\top H_j V_{\Delta,j} V_{\Delta,j}^\top e]^2}.$$

Then

$$\mathrm{KL}_j(\theta^*\|\theta'_j) = \mathcal{C}_j(\theta^*, \gamma_j)\,\frac{(\Delta_j - \varepsilon)^2}{2} + O(\|\theta'_j - \theta^*\|^3). \tag{41}$$

We also have,

$$\sum_{k=1}^{K} \mathbb{E}[T_k(\tau)] \cdot \mathrm{KL}_k\big(\theta^* \,\|\, \theta'_j\big) \geq \mathbb{E}[T_j(\tau)] \cdot \mathrm{KL}_j\big(\theta^* \,\|\, \theta'_j\big)\,. \tag{42}$$

Let w_j as representing the proportion of times that arm j is pulled by the algorithm as the total number of steps n approaches infinity. $T_j(n)$ is the pull count of arm j after n rounds. So $\mathbb{E}[T_j(\tau)] = w_j \mathbb{E}[\tau]$ equation 42 becomes:

$$\sum_{k=1}^{K} \mathbb{E}[T_k(\tau)] \cdot \mathrm{KL}_k\big(\theta^* \,\|\, \theta'_j\big) \geq \mathbb{E}[\tau] w_j \mathrm{KL}_j(\theta^*||\theta'_j) = \mathbb{E}[\tau] w_j \mathcal{C}_j(\theta^*, \gamma_j) \frac{(\Delta_j - \epsilon)^2}{2}. \tag{43}$$

and

$$\mathbb{E}[\tau] w_j \mathcal{C}_j(\theta^*, \gamma_j) \frac{(\Delta_j - \epsilon)^2}{2} \geq \log \frac{1}{2.4\delta}. \tag{44}$$

So,

$$w_j \geq \frac{2\log(1/2.4\delta)}{\mathbb{E}[\tau](\Delta_j - \epsilon)^2 \mathcal{C}_j(\theta^*, \gamma_j)}.$$

Sum both $j \in K_\epsilon$:

$$\sum_{j \in K_\epsilon} w_j \geq \frac{2\log(1/2.4\delta)}{\mathbb{E}[\tau]} \cdot \sum_{j \in K_\epsilon} \frac{1}{(\Delta_j - \epsilon)^2 \mathcal{C}_j(\theta^*, \gamma_j)}.$$

We also have,

$$\sum_{j \in K_\epsilon} w_j = 1,$$

so ,

$$1 \geq \frac{2\log(1/2.4\delta)}{\mathbb{E}[\tau]} \cdot \sum_{j \in K_\epsilon} \frac{1}{(\Delta_j - \epsilon)^2 \mathcal{C}_j(\theta^*, \gamma_j)}.$$

Finally,

$$\mathbb{E}[\tau] \geq 2\log\frac{1}{2.4\delta} \cdot \sum_{k \in K_\epsilon} \frac{1}{(\Delta_k - \epsilon)^2 \mathcal{C}_k(\theta^*, \gamma_k)}. \tag{45}$$

B.3 Stopping-Time Upper Bound

The goal is to determine how many iterations (n_k) are required for an optimal algorithm to eliminate suboptimal arms k. It means, we need what is the maximum number of pulls we need to eliminate with confidence gap small enough.

From lemma **??**, in good event $\mathcal{G}$, we have:

$$\Delta_k - \beta_t(i_t, k) \leq \Delta_t(k, i_t) \leq \Delta_k + \beta_t(i_t, k). \tag{46}$$

Lower Confidence Bound is defined as the estimated difference gap between a specific arm and the currently chosen arm, minus the uncertainty margin of that estimation. In BAI (ϵ, δ)-PAC problem, to find best arm with high probability $1 - \delta$, we need $\text{LCB}_k(\tau) \geq \epsilon$ and we have consequence:

$$\text{LCB}_k(t) = \Delta_t(k, i_t) - \beta_t(i_t, k) \geq \Delta_k - 2\beta_t.$$

We only want to remove the arm if even the lower bound of the gap (LCB) is greater than ϵ. So,

$$\Delta_k - 2\beta_t \geq \epsilon.$$

From that,

$$2\beta_t \leq \Delta_k - \epsilon. \tag{47}$$

Based on study's Kazerouni et al (Kazerouni & Wein, 2021) to build gap, we choose parameter set up like their research. We have,

$$\beta_t(i, k) = C_t. \max_{c, c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}}.$$

(with $C_t = \frac{2\kappa}{c_\mu}\sqrt{2m\log(t)\ \log\left(\frac{\pi^2 m^2 t^2}{6\delta}\right)}$, m is dimension of θ.)

So,

$$2C_t. \max_{c, c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}} \leq \Delta_k - \epsilon, \tag{48}$$

Our next goal is to find $\max_{c, c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}}$. To solve this problem, we will have some definitions and lemmas.

Definition 14. We unify some of the definitions that have appeared in lower bound

1. Let
$$(V_i - V_k)^\top e = V_{\Delta,k}^\top e = v_k. \tag{49}$$

2. Let
$$\mathcal{C}_k(\theta^*, \gamma_k) = \frac{(V_{\Delta,k}^\top e)^\top J_k(\theta^*)(V_{\Delta,k}^\top e)}{[e^\top H_k V_{\Delta,k} V_{\Delta,k}^\top e]^2}, \tag{50}$$
with H_k is Hessian $\nabla_\eta^2 \phi(V_k \theta^*)$.

3.
$$\mathrm{supp}(p) = \{k : p_k(z_t) > 0\} = \{k : w_k^* > 0\}.$$
That is, the set of arms receiving positive allocation, i.e., the arms that are sampled under the water-filling allocation.

4. Let
$$c_0 = \frac{1}{2}\lambda_{\min}\left(\sum_{j \in \mathrm{supp}(p)} w_j^* V_j^\top C_j(\theta^*, \gamma_j) V_j\right) > 0, \tag{51}$$
where $C_j(\theta^*, \gamma_j) = \mathrm{Cov}(\mathbb{E}[T_{Y_j}|G_j])$.

Lemma 15. *With c_0, $\mathcal{C}_k(\theta^*, \gamma_k)$ is defied in equation 50 and equation 51, respecitvely.*

1. $\max_{c,c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}} \le 2k_\mu \max_j \|V_j^\top e\|_{J_t^{-1}}$.

2. $\|V_j^\top e\|_{J_t^{-1}}^2 \le \frac{dL^2\|e\|_2^2}{\lambda_{\min}(J_t)}$, *with* $L = \max_j \|V_j\|_{op}$, *d is dimension of sufficient statistic.*

Proof. (1) We have
$$\max_{c,c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}} = \max_{c,c' \in [c_\mu, k_\mu]} \sqrt{[(cV_i - c'V_k)^\top e] J_t^{-1} [(cV_i - c'V_k)^\top e]^\top}.$$

Let $f(c, c')$ = $(cV_i - c'V_k)^\top e J_t^{-1} e^\top (cV_i - c'V_k)^\top$.

We have J_t^{-1} is a positive definite matrix, so $f(c, c')$ is quadratic matrix and a convex function. We also have c, c' in $[c_\mu, k_\mu]$, so constraint set is compact, closed volume. Therefore, $f(c, c')$ is maximized at boundary point of the constraint set. It means,

- $c = k_\mu$ and $c' = k_\mu$.
- $c = k_\mu$ and $c' = c_\mu$.

So,
$$\|(cV_i - c'V_k)^\top e\|_{J_t^{-1}} = \|cV_i^\top e - c'V_k^\top e\|_{J_t^{-1}}.$$

Since $c, c' \le k_\mu$, we substitute this maximum value to obtain the absolute upper bound:
$$\le |c| \cdot \|V_i^\top e\|_{J_t^{-1}} + |c'| \cdot \|V_k^\top e\|_{J_t^{-1}} \le k_\mu(\|V_i^\top e\|_{J_t^{-1}} + \|V_k^\top e\|_{J_t^{-1}}).$$

We also have $\|V_i^\top e\|_{J_t^{-1}} \le \max_j \|V_j^\top e\|_{J_t^{-1}}$, and $\|V_k^\top e\|_{J_t^{-1}} \le \max_j \|V_j^\top e\|_{J_t^{-1}}$.

So,
$$\max_{c,c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}} \le k_\mu(\|(V_i)^\top e\|_{J_t^{-1}} + \|(V_k)^\top e\|_{J_t^{-1}}) \le 2k_\mu \max_j \|V_j^\top e\|_{J_t^{-1}}. \tag{52}$$

(2) Apply weighted Frobenius norm, we have
$$\|V_j^\top e\|_{J_t^{-1}}^2 = \mathrm{tr}((V_j^\top e)^\top J_t^{-1}(V_j^\top e)) = \mathrm{tr}(e^\top V_j J_t^{-1} V_j^\top e).$$

We also have,

$$\operatorname{tr}(e^\top V_j J_t^{-1} V_j^\top e) \le \frac{\operatorname{tr}(e^\top V_j V_j^\top e)}{\lambda_{\min}(J_t)} = \frac{\|V_j^\top e\|_2^2}{\lambda_{\min}(J_t)} \le \frac{dL^2\|e\|_2^2}{\lambda_{\min}(J_t)}. \tag{53}$$

From equation 52 and 53, we have

$$\max_{c,c' \in [c_\mu, k_\mu]} \|(cV_i - c'V_k)^\top e\|_{J_t^{-1}} \le 2k_\mu \|e\|_2 \frac{\sqrt{dL^2}}{\sqrt{\lambda_{\min}(J_t)}} = 2k_\mu \frac{\sqrt{dL^2}}{\sqrt{\lambda_{\min}(J_t)}}. \tag{54}$$

□

Looking at equation 54, our next step is to find the lower bound for $\lambda_{\min}(J_t)$. From the tracking rule, with sufficiently large t, we have $T_j(t) \ge \frac{w_j^* t}{2}$:

$$J_t = \sum_{s=1}^{t} V_{a_s}^\top C_{a_s}(\hat{\theta}_{s-1}, \gamma_{a_s}) V_{a_s}$$

$$= \sum_{j=1}^{K} \sum_{a_s = j} V_j^\top C_j(\hat{\theta}_{s-1}, \gamma_j) V_j \ge \sum_{j=1}^{K} \sum_{a_s = j} V_j^\top C_j(\theta^*, \gamma_j) V_j.$$

Then, $V_j^\top C_j(\theta^*, \gamma_j) V_j$ no longer depends on the time step s. The number of terms in this sum is precisely the number of times arm j is pulled, denoted $T_j(t)$. So,

$$J_t \ge \sum_{j=1}^{K} T_j(t) V_j^\top C_j(\theta^*, \gamma_j) V_j,$$

and according SELECTARM algorithm we have $T_j(t) \ge \frac{w_j^* t}{2}$:

$$J_t \ge \sum_{j=1}^{K} \frac{w_j^* t}{2} V_j^\top C_j(\theta^*, \gamma_j) V_j \ge \sum_{j \in \text{supp}(p)} \frac{w_j^* t}{2} \cdot V_j^\top C_j(\theta^*, \gamma_j) V_j$$

$$= \frac{t}{2} \sum_{j \in \text{supp}(p)} w_j^* V_j^\top C_j(\theta^*, \gamma_j) V_j.$$

So,

$$\lambda_{\min}(J_t) \ge \frac{t}{2} \lambda_{\min}\left(\sum_{j \in \text{supp}(p)} w_j^* V_j^\top C_j(\theta^*, \gamma_j) V_j \right) = c_0 t.$$

With $\lambda_{\min}(J_t) \ge c_0 t$, substitute into equation 54, we have

$$\max_{c,c' \in [c_\mu, k_\mu]} \|(cV_{i_t} - c'V_k)^\top e\|_{J_t^{-1}} \le 2k_\mu \frac{\sqrt{dL^2}}{\sqrt{\lambda_{\min}(J_t)}} \le \frac{2k_\mu \sqrt{dL^2}}{\sqrt{c_o t}}. \tag{55}$$

Our next goal is to find lower bound for c_0. To find upper value c_0, we use result from lemma 16 below.

Lemma 16. *With c_0, $\mathcal{C}_k(\theta^*, \gamma)$ is defied in 50, 51 respecitvely.*

1. $c_0 \ge \frac{c_\mu^2}{2} \min_j(w_j^* \mathcal{C}_j(\theta^*, \gamma_j))\lambda_1$, *with* $\lambda_1 = \lambda_{\min}\left(\sum_j V_{\Delta,j}^\top V_{\Delta,j}\right)$.

Proof. From definitation of $\mathcal{C}_k(\theta^*, \gamma)$ as 50 Let $v_j = V_{\Delta,j}^\top e$, we have

$$\begin{aligned} v_j^\top J_j(\theta^*) v_j &= \mathcal{C}_j(\theta^*, \gamma_j)[e^\top H_k V_{\Delta,k} V_{\Delta,k}^\top e]^2 \\ &\geq \mathcal{C}_j(\theta^*, \gamma_j)(c_\mu \|v_j\|^2)^2. \end{aligned} \tag{56}$$

By equation 56, we also have

$$\lambda_{\min}(J_j) \geq \frac{v_j^\top J_j v_j}{\|v_j\|^2} \geq c_\mu^2 \mathcal{C}_j(\theta^*, \gamma_j)\|v_j\|^2 \geq c_\mu^2 \mathcal{C}_j(\theta^*, \gamma_j)\lambda_{\min}(V_{\Delta,j}^\top V_{\Delta,j}). \tag{57}$$

With c_0 defined as in 51:

$$c_0 = \frac{1}{2}\lambda_{\min}\left(\sum_{j \in \text{supp}(p)} w_j^* V_j^\top C_j(\theta^*, \gamma_j) V_j\right) = \frac{1}{2}\lambda_{\min}(\sum_{j \in \text{supp}(p)} w_j^* J_j).$$

Apply equation 57, we have

$$c_0 = \frac{1}{2}\lambda_{\min}(\sum_{j \in \text{supp}(p)} w_j^* J_j) \geq \frac{1}{2} c_\mu^2 \sum_{j \in \text{supp}(p)} w_j^* . \mathcal{C}_j(\theta^*, \gamma_j)(V_{\Delta,j}^\top V_{\Delta,j}). \tag{58}$$

Relaxing inequality 58, we have

$$c_0 \geq \frac{1}{2} . c_\mu^2 \min_j (w_j^* \mathcal{C}_j(\theta^*, \gamma_j))\lambda_{\min}(\sum_j V_{\Delta,j}^\top V_{\Delta,j}) = \frac{1}{2} . c_\mu^2 \min_j (w_j^* \mathcal{C}_j(\theta^*, \gamma_j))\lambda_1. \tag{59}$$

□

Using lower bound of c_0 from lemma 16 substitute into equation 55, we have

$$\max_{c,c' \in [c_\mu, k_\mu]} \|(cV_{i_t} - c' V_k)^\top e\|_{J_t^{-1}} \leq \frac{2k_\mu \sqrt{dL^2}}{\sqrt{c_0 t}} \leq \frac{2k_\mu \sqrt{dL^2}}{\sqrt{\frac{1}{2} t . c_\mu^2 . \lambda_1 \min_j (w_j^* \mathcal{C}_j(\theta^*, \gamma_j))}}. \tag{60}$$

Our next goal is to find upper bound value for $\min_j (w_j^* \mathcal{C}_j(\theta^*, \gamma_j))$.

Lemma 17. *Suppose that*

$$\max_{\substack{w_j \geq 0 \\ \sum_j w_j = 1}} \min_{j \in \mathcal{K}_\epsilon} w_j (\Delta_j - \epsilon)^2 \mathcal{C}_j(\theta^*, \gamma_j),$$

having w_j^* is optimal solution then $w_j^*(\Delta_j - \epsilon)^2 \mathcal{C}_j(\theta, \gamma_j) = b$ with all j.

Proof. Suppose for contradiction that at the optimal solution w^*, there exist j_1 and j_2 such that:

$$w_{j_1}^* (\Delta_{j_1} - \epsilon)^2 \mathcal{C}_{j_1}(\theta^*, \gamma_{j_1}) < w_{j_2}^* (\Delta_{j_2} - \epsilon)^2 \mathcal{C}_{j_2}(\theta^*, \gamma_{j_2}).$$

This implies that j_1 is bottleneck and j_2 is slack element. There exists $\epsilon' > 0$ and let

$$w'_{j_1} = w_{j_1}^* + \epsilon' \quad w'_{j_2} = w_{j_2}^* - \epsilon', \quad w'_j = w_j^* \quad \forall j \neq j_1, j_2.$$

So, we have

$$(w_{j_1}^* + \epsilon')(\Delta_{j_1} - \epsilon)\mathcal{C}_{j_1}(\theta^*, \gamma_{j_1}) > w_{j_1}^* (\Delta_{j_1} - \epsilon)^2 \mathcal{C}_{j_1}(\theta^*, \gamma_{j_1}), \tag{61}$$

and

$$(w_{j_2}^* - \epsilon')(\Delta_{j_2} - \epsilon)\mathcal{C}_{j_2}(\theta^*, \gamma_{j_2}) < w_{j_2}^* (\Delta_{j_2} - \epsilon)\mathcal{C}_{j_2}(\theta^*, \gamma_{j_2}). \tag{62}$$

Let $\mathcal{C}_{j_1} = \mathcal{C}_{j_1}(\theta^*, \gamma_{j_1})$, $\mathcal{C}_{j_2} = \mathcal{C}_{j_2}(\theta^*, \gamma_{j_2})$, and $\mathcal{C}_j = \mathcal{C}_j(\theta^*, \gamma_j)$.

Assume that existing ϵ' sufficiently small such that:

$$(w_{j_2}^* - \epsilon')(\Delta_{j_2} - \epsilon)\mathcal{C}_{j_2} \geq (w_{j_1}^* + \epsilon')(\Delta_{j_1} - \epsilon)\mathcal{C}_{j_1}.$$

So,

$$\max_w \min_j (w_{j_2}^*(\Delta_{j_2} - \epsilon)\mathcal{C}_{j_2}) = (w_{j_1}^* + \epsilon')(\Delta_{j_2} - \epsilon)\mathcal{C}_{j_1}.$$

This contradicts the assumption that w^* is the optimal solution. □

We have $w_j^*(\Delta_j - \epsilon)^2\mathcal{C}_j = \text{const} = b$, so:

$$w_j^* = \frac{b}{(\Delta_j - \epsilon)^2\,\mathcal{C}_j}. \tag{63}$$

We also have,

$$w_j^*\mathcal{C}_j = \frac{b}{(\Delta_j - \epsilon)^2}.$$

$$\sum_j w_j^* = 1 = \sum_j \frac{b}{(\Delta_j - \epsilon)^2\mathcal{C}_j} = b\sum_j \frac{1}{(\Delta_j - \epsilon)^2\mathcal{C}_j}.$$

So,

$$b = \frac{1}{\sum_j \frac{1}{(\Delta_j - \epsilon)^2\mathcal{C}_j}}.$$

Subtile b in equation 63,

$$w_j^* = \frac{1/[(\Delta_j - \epsilon)^2\mathcal{C}_j]}{\sum_l 1/[(\Delta_l - \epsilon)^2\mathcal{C}_l]}.$$

$$w_j^*\mathcal{C}_j = \frac{1/(\Delta_j - \epsilon)^2}{\sum_l 1/[(\Delta_l - \epsilon)^2\mathcal{C}_l]}.$$

So,

$$\frac{1}{w_j^*(\Delta_j - \epsilon)^2\mathcal{C}_j} = \frac{1}{b} = \sum_{l \in K_\epsilon} \frac{1}{(\Delta_l - \epsilon)^2\mathcal{C}_l}. \tag{64}$$

From that,

$$\min_j (w_j^*\mathcal{C}_j) = \frac{b}{\max_j (\Delta_j - \epsilon)^2}. \tag{65}$$

Consider the stopping time; t is number pulls, let $t = n_k$.

By the elimination condition, we have:

$$2\beta_{n_k}(i,k) \leq \Delta_k - \epsilon$$

$$\Rightarrow 2\frac{4\kappa R k_\mu L\sqrt{2m}\,\sqrt{d}}{c_\mu\sqrt{c_0}}\sqrt{\frac{\log n_k \cdot \log\left(\frac{\pi^2 m n_k^2}{6\delta}\right)}{n_k}} \leq \Delta_k - \epsilon. \tag{66}$$

Let $A = \frac{4\kappa R k_\mu L\sqrt{2m}\,\sqrt{d}}{c_\mu\sqrt{c_0}}$. Squaring both sides, we have:

$$\frac{4A^2 \log n_k\ \log\left(\frac{\pi^2 m n_k^2}{6\delta}\right)}{n_k} \leq (\Delta_k - \epsilon)^2. \tag{67}$$

Then the next step, we need to find the minimum n_k that satisfies the equation 67.

Let $u = (\Delta_k - \epsilon)^2$.

Rewriting equation 67, we have:

$$n_k \geq \frac{4A^2}{(\Delta_k - \epsilon)^2} \log n_k \log\left(\frac{\pi^2 m n_k^2}{6\delta}\right). \tag{68}$$

We hypothesize that n_k take the form solution $n_k = \frac{C_2}{(\Delta_k-\epsilon)^2} \log^2\left(\frac{1}{(\Delta_k-\epsilon)^2\delta}\right)$, with C_2 is a constant to be determined.

Plug into $\log n_k$:

$$\log n_k = \log \frac{C_2}{(\Delta_k - \epsilon)^2} + \log\log \frac{1}{(\Delta_k - \epsilon)^2\delta} = O\left(\log \frac{1}{(\Delta_k - \epsilon)^2\delta}\right).$$

And

$$\log\left(\frac{\pi^2 m n_k^2}{6\delta}\right) = O\left(\log \frac{1}{(\Delta_k - \epsilon)^2\delta}\right),$$

(since n_k^2 only multiplies by a polynomial factor inside $1/((\Delta_k - \epsilon)^2\delta)$, which only increases the logarithm within the same order).

So,

$$\log n_k \log\left(\frac{\pi^2 m n_k^2}{6\delta}\right) = O\left(\log^2 \frac{1}{(\Delta_k - \epsilon)^2\delta}\right) = O\left(\log^2 \frac{1}{(\Delta_k - \epsilon)^2\delta}\right).$$

It follows that there exists a sufficiently large constant such that:

$$n_k \leq \frac{16A_1^2}{(\Delta_k - \epsilon)^2} \log^2\left(\frac{1}{(\Delta_k - \epsilon)^2\delta}\right), \text{with} \quad A_1 = \frac{4\kappa R k_\mu L\sqrt{2m}}{c_\mu \sqrt{c_0}}. \tag{69}$$

Combine equation 48 and equation 60, we obtain:

$$n_k \leq \frac{2048\,\kappa^2 R^2 k_\mu^2 L^2 m d\, \log^2\left(\frac{1}{(\Delta_k-\epsilon)^2\delta}\right)}{c_\mu^4 \lambda_1\, w_k^*\, (\Delta_k - \epsilon)^2\, \mathcal{C}_k(\theta^*, \gamma_k)}. \tag{70}$$

And,

$$\mathbb{E}[\tau] \leq \sum_{k\in\mathcal{K}_\epsilon} n_k \leq \frac{2048\kappa^2 R^2 k_\mu^2 L^2 m d}{c_\mu^4 \lambda_1} \cdot \log^2\left(\frac{1}{\delta}\right) \cdot \sum_{k\in\mathcal{K}_\epsilon} \frac{1}{w_k^*(\Delta_k - \epsilon)^2 \mathcal{C}_k(\theta^*, \gamma_k)}. \tag{71}$$

By combining equation 64 and equation 71, we have:

$$\mathbb{E}[\tau] \leq \frac{2048\,\kappa^2 R^2 k_\mu^2 L^2 m d\, \log^2\left(\frac{1}{\delta}\right)}{c_\mu^4 \lambda_1} \cdot \sum_{k\in\mathcal{K}_\epsilon} \frac{1}{(\Delta_k - \epsilon)^2 \mathcal{C}_k(\theta^*, \gamma_k)}. \tag{72}$$

C Auxiliary Results

Lemma 18. *Let $\hat{\theta}_t$ denotes the exact Maxium Likehood Estimator (MLE) at round t. Suppose the estimator is updated via the Newton-Raphson method with the previous estimator $\hat{\theta}_{t-1}$ as the initialization $\theta^{(0)}$:*

$$\theta^{(s+1)} = \theta^{(s)} + J_t^{-1}(\theta^{(s)}) \nabla \ell_t(\theta^{(s)}), \tag{73}$$

where $J_t(\theta) = -\nabla^2 \ell_t(\theta)$ is the Fisher Information Matrix (FIM). We operate under the following assumptions:

- *Initialization: The starting point satisfies* $\|\theta^{(0)} - \hat{\theta}_t\|_2 \leq r$ *for a sufficiently small radius* $r > 0$.
- *Strong Concavity: The log-likelihood function* $\ell_t(\theta)$ *is strongly concave within the Euclidean ball of radius* r *centered at* $\hat{\theta}_t$*, with a parameter* $c_\mu > 0$.
- *Smoothness (Hessian Lipschitz): The Hessian* $\nabla^2 \ell_t(\theta)$ *is* L_H*-Lipschitz continuous.*

Then, after M *iterations, the error is bounded by:*

$$\|\theta^{(M)} - \hat{\theta}_t\|_2 \leq \left(\frac{L_H r}{2c_\mu}\right)^{2^M - 1} \|\theta^{(0)} - \hat{\theta}_t\|_2 \tag{74}$$

Specifically, choosing $M = \mathcal{O}(\log\log(1/\epsilon_0))$ *guarantees an optimization error of* $\|\theta^{(M)} - \hat{\theta}_t\|_2 \leq \epsilon_0$.

Proof. We begin by analyzing the error term at iteration $s + 1$. By substituting the Newton-Raphson update rule, we have:

$$\theta^{(s+1)} - \hat{\theta}_t = \theta^{(s)} - \hat{\theta}_t + J_t^{-1}(\theta^{(s)}) \nabla \ell_t(\theta^{(s)})$$

Since $\hat{\theta}_t$ is the exact minimizer, the gradient vanishes, i.e. $\nabla \ell_t(\hat{\theta}_t) = 0$. Applying the Mean Value Theorem (or exact Taylor expansion) around $\hat{\theta}_t$, we obtain:

$$\nabla \ell_t(\theta^{(s)}) = \nabla \ell_t(\theta^{(s)}) - \nabla \ell_t(\hat{\theta}_t) = \nabla^2 \ell_t(\bar{\theta})(\theta^{(s)} - \hat{\theta}_t)$$

where $\bar{\theta}$ is a convex combination of $\theta^{(s)}$ and $\hat{\theta}_t$, we obtain:

$$\theta^{(s+1)} - \hat{\theta}_t = J_t^{-1}(\theta^{(s)}) \left[\nabla^2 \ell_t(\bar{\theta}) - \nabla^2 \ell_t(\theta^{(s)})\right] (\theta^{(s)} - \hat{\theta}_t)$$

Taking the L2-norm on both sides and applying the sub-multiplicative property of the operator norm, we get:

$$\|\theta^{(s+1)} - \hat{\theta}_t\|_2 \leq \|J_t^{-1}(\theta^{(s)})\|_2 \cdot \|\nabla^2 \ell_t(\bar{\theta}) - \nabla^2 \ell_t(\theta^{(s)})\|_2 \cdot \|\theta^{(s)} - \hat{\theta}_t\|_2$$

By the strong concavity assumption, the spectral norm of the inverse Hessian is bounded: $\|J_t^{-1}(\theta^{(s)})\|_2 \leq 1/c_\mu$. Furthermore, by the L_H-Lipschitz continuity of the Hessian and the fact that $\|\bar{\theta} - \theta^{(s)}\|_2 \leq \|\hat{\theta}_t - \theta^{(s)}\|_2$, we have:

$$\|\theta^{(s+1)} - \hat{\theta}_t\|_2 \leq \frac{1}{c_\mu} \cdot L_H \|\theta^{(s)} - \hat{\theta}_t\|_2 \cdot \|\theta^{(s)} - \hat{\theta}_t\|_2 = \frac{L_H}{c_\mu} \|\theta^{(s)} - \hat{\theta}_t\|_2^2$$

Let $e_s = \|\theta^{(s)} - \hat{\theta}_t\|_2$ and $C = L_H/c_\mu$. The recurrence relation becomes $e_{s+1} \leq C e_s^2$, which impies $C e_{s+1} \leq (C e_s)^2$. Unrolling this inequality over M iterations yields $C e_M \leq (C e_0)^{2^M}$, and $e_0 \leq r$ (assumption). Rearranging the terms, we arrive at the final bound:

$$e_M \leq \frac{1}{C}(C e_0)^{2^M} = \left(\frac{L_H r}{c_\mu}\right)^{2^M - 1} e_0.$$

□

Lemma 19 (Finite-Time Bound and Consistency of Adaptive Tracking)**.** *Let* $\mathcal{K} = \{1, \ldots, K\}$ *denote the set of arms. Let* $p^* = w^*(\theta^*) \in \Delta^K$ *be the optimal target allocation vector derived from the water-filling optimization problem at the true parameter* θ^**, with active support* $\mathrm{supp}(p^*) = \{k \in \mathcal{K} : p_k^* > 0\}$.

At each round $t \geq 1$*, the algorithm updates the parameter estimate* $\hat{\theta}_{t-1}$*, computes the adaptive target allocation vector* $p(t) = w^*(\hat{\theta}_{t-1})$*, and selects arm* a_t *according to the adaptive tracking rule:*

$$a_t = \arg\min_{i : p_i(t) > 0} \frac{T_i(t-1)}{p_i(t)}.$$

Then, the following properties hold:

1. ***Asymptotic Consistency:***

$$\lim_{t\to\infty}\hat{\theta}_t = \theta^* \quad \textit{and} \quad \lim_{t\to\infty}\frac{T_k(t)}{t} = p_k^* \quad \textit{a.s} \quad \forall k \in \mathcal{K}.$$

2. ***Finite-Time Lower Bound:*** *Under the concentration event $\mathcal{E}_\delta$ (where $\mathbb{P}(\mathcal{E}_\delta) \geq 1-\delta$), there exists a finite stabilization time $t_0(\delta) = \mathcal{O}\left(\frac{d\ln(1/\delta)}{r^2}\right)$ such that for all $t \geq t_0(\delta)$ and all $k \in \mathrm{supp}(p^*)$:*

$$T_k(t) \geq \frac{p_k^* t}{2}$$

Proof. Consider any sequence of target vectors $\{p(s)\}_{s=1}^t \subset \Delta^K$. For any arm $k \in \mathcal{K}$ and any round $t \geq 1$:

- **Case 1:** If $T_k(t) = 0$, then $T_k(t) - \sum_{s=1}^t p_k(s) = -\sum_{s=1}^t p_k(s) \leq 0$.
- **Case 2:** If $T_k(t) \geq 1$, let $t_{\text{last}} \leq t$ be the last round up to t in which arm k was selected ($a_{t_{\text{last}}} = k$). By definition of the tracking rule at round t_{last}:

$$\frac{T_k(t_{\text{last}}-1)}{p_k(t_{\text{last}})} \leq \sum_{i=1}^K p_i(t_{\text{last}})\frac{T_i(t_{\text{last}}-1)}{p_i(t_{\text{last}})} = \sum_{i=1}^K T_i(t_{\text{last}}-1) = t_{\text{last}} - 1.$$

  Thus, $T_k(t) = T_k(t_{\text{last}}) \leq p_k(t_{\text{last}})(t_{\text{last}}-1) + 1$.

Subtracting $\sum_{s=1}^t p_k(s)$ from both sides yields an upper bound on the allocation surplus:

$$T_k(t) - \sum_{s=1}^t p_k(s) \leq 1 + \sum_{s=1}^{t_{\text{last}}-1}\big(p_k(t_{\text{last}}) - p_k(s)\big). \tag{75}$$

Since the total allocation error across all K arms sums to zero ($\sum_{i=1}^K\big(T_i(t) - \sum_{s=1}^t p_i(s)\big) = 0$), the allocation deficit $D_k(t) \triangleq \sum_{s=1}^t p_k(s) - T_k(t)$ of arm k is upper bounded by:

$$D_k(t) \leq K - 1 + \sum_{j\neq k}\sum_{s=1}^{t_{\text{last}}(j)-1}\big|p_j(t_{\text{last}}(j)) - p_j(s)\big|, \tag{76}$$

where $t_{\text{last}}(j)$ denotes the last round up to t in which arm j was selected.

Firstly, we prove that every arm $k \in \mathrm{supp}(p^*)$ is sampled infinitely often ($T_k(t) \to \infty$). Suppose for contradiction that there exists an arm $k \in \mathrm{supp}(p^*)$ such that $\lim_{t\to\infty} T_k(t) = M < \infty$. As $t = \sum_{i=1}^K T_i(t) \to \infty$, there must exist at least one arm j such that $T_j(t) \to \infty$.

Consequently, the tracking ratio of arm j approaches infinity:

$$\lim_{t\to\infty}\frac{T_j(t-1)}{p_j(t)} = \infty.$$

In contrast, the tracking ratio of arm k remains bounded above by $\frac{M}{\inf_t p_k(t)} < \infty$ (since $p_k^* > 0$ and the parameter space is compact). Under the selection rule $a_t = \arg\min_i \frac{T_i(t-1)}{p_i(t)}$, the algorithm is compelled to select arm k in subsequent iterations, contradicting the assumption that $T_k(t)$ is bounded. Hence, $T_k(t) \to \infty$ almost surely for all $k \in \mathrm{supp}(p^*)$.

We have $J_t = \sum_{s=1}^t V_{a_s}^\top \mathrm{Cov}(\mathbb{E}[T_{Y_{a_s}}|G_{a_s}])V_{a_s}$. Let $C_{a_s}(\hat{\theta}_{s-1}, \gamma_{a_s}) = \mathrm{Cov}(\mathbb{E}[T_{Y_{a_s}}|G_{a_s}])$.

Since the parameter space Θ ($\theta \in \Theta$) is a compact set, the matrix function $C_k(\theta, \gamma_k)$ always has a minimum value greater than 0 over the entire domain Θ. Suppose:

$$c_{\min} = \min_k \inf_{\theta\in\Theta} \lambda_{\min}\left(C_k(\theta, \gamma_k)\right) > 0.$$

Then, at any round s, we always have:

$$C_k(\hat{\theta}_{s-1}, \gamma_k) \geq c_{\min} I.$$

From there, we have:

$$J_t = \sum_{k=1}^{K} V_k^\top \left[\sum_{s:a_s=k} C_k(\hat{\theta}_{s-1}, \gamma_k) \right] V_k \geq c_{\min} \sum_{k=1}^{K} T_k(t) V_k^\top V_k.$$

At this point:

$$\lambda_{\min}(J_t) \geq c_{\min} \cdot \min_{k \in \text{supp}(w^*)} T_k(t) \cdot \lambda_{\min} \left(\sum_{k \in \text{supp}(w^*)} V_k^\top V_k \right).$$

Since $T_k(t) \to \infty$, $\lambda_{\min}(J_t) \to \infty$. From that, $\hat{\theta}_t \xrightarrow{a.s.} \theta^*$. By local Lipschitz continuity of the water-filling optimization mapping $w^*(\cdot)$, we obtain:

$$p(t) = w^*(\hat{\theta}_{t-1}) \xrightarrow{a.s.} p^* = w^*(\theta^*).$$

Applying the Cesàro Mean Theorem to the convergent sequence $p(s) \to p^*$:

$$\lim_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} p(s) = p^* \quad \text{a.s.}$$

Dividing both sides of equations 75 and 76 by t and taking the limit $t \to \infty$, the term $\frac{1}{t} \sum_{s=1}^{t} |p(t) - p(s)| \to 0$ vanishes. Therefore:

$$\lim_{t \to \infty} \frac{T_k(t)}{t} = p_k^* \quad \text{a.s.} \quad \forall k \in \mathcal{K}.$$

(b) We also have established that $\frac{T_k(t)}{t} \to p_k^* > 0$. Thus, there exists an initial stabilization time $T_{\text{stab}}(\delta)$ such that $T_k(s) \geq c \cdot s$ for all $s \geq T_{\text{stab}}(\delta)$.

Under event $\mathcal{E}_\delta$, for all $s \geq T_{\text{stab}}(\delta)$:

$$\|p(s) - p^*\|_\infty \leq L_{\text{opt}} \|\hat{\theta}_{s-1} - \theta^*\|_2 \leq C_1 \sqrt{\frac{\ln(s/\delta)}{s}}.$$

Substituting this rate into equation 76, we integrate the tracking deficit:

$$\sum_{s=1}^{t} \left| p_j(t_{\text{last}}) - p_j(s) \right| \leq \sum_{s=1}^{t} \left(C_1 \sqrt{\frac{\ln(t/\delta)}{t}} + C_1 \sqrt{\frac{\ln(s/\delta)}{s}} \right) \leq C_2 \sqrt{t \ln(t/\delta)} + H(T_{\text{stab}}).$$

Therefore, the allocation deficit is bounded by:

$$D_k(t) = \sum_{s=1}^{t} p_k(s) - T_k(t) \leq C_3 \sqrt{t \ln(t/\delta)} + C_4.$$

So:

$$T_k(t) \geq \sum_{s=1}^{t} p_k(s) - C_3 \sqrt{t \ln(t/\delta)} - C_4.$$

Furthermore:

$$\sum_{s=1}^{t} p_k(s) \geq p_k^* t - p_k^* T_{\text{stab}}(\delta) - 2C_1 \sqrt{t \ln(t/\delta)}.$$

Combining these bounds yields:

$$T_k(t) \geq p_k^* t - C_5 \sqrt{t \ln(t/\delta)} - \left(p_k^* T_{\text{stab}}(\delta) + C_4 \right). \tag{77}$$

Choosing $t_0(\delta) \geq T_{\text{stab}}(\delta)$ sufficiently large such that for all $t \geq t_0(\delta)$:

$$C_5\sqrt{t\ln(t/\delta)} + p_k^* T_{\text{stab}}(\delta) + C_4 \leq \frac{p_k^* t}{2}.$$

Substituting this threshold into equation 77 yields the desired result:

$$T_k(t) \geq \frac{p_k^* t}{2} \quad \forall t \geq t_0(\delta),\, \forall k \in \text{supp}(p^*).$$

□

Lemma 20. *The covariance of the conditional expectation $\mathbb{E}[T_{Y_k}|G_k]$ is given by the following closed-from expression:*

$$\textit{Cov}(\mathbb{E}[T_{Y_k}|G_k]) = p_k(1-p_k)(\mu_1^k - \mu_0^k)(\mu_1^k - \mu_0^k)^\top, \tag{78}$$

where $\mu_g^k = \mathbb{E}[T_{Y_k} \mid G_k = g] \in \mathbb{R}^d$.

Proof.

$$\text{Cov}(\mathbb{E}[T_{Y_k}|G_k]) = \mathbb{E}[\mathbb{E}[T_{Y_k}|G_k] \cdot \mathbb{E}[T_{Y_k}|G_k]^\top] - \mu_k\mu_k^\top. \tag{79}$$

We also have the expectation of the outer product term is evaluated based on the underlying binary distribution:

$$\mathbb{E}[\mathbb{E}[T_{Y_k}|G_k]\mathbb{E}[T_{Y_k}|G_k]^\top] = p_k\mu_1^k(\mu_1^k)^\top + (1-p_k)\mu_0^k(\mu_0^k)^\top. \tag{80}$$

By substituting the definition $\mu_k = p_k\mu_1^k + (1-p_k)\mu_0^k$, the expression is fully expanded as:

$$\begin{aligned}\mu_k\mu_k^\top &= [p_k\mu_1^k + (1-p_k)\mu_0^k][p_k{\mu_1^k}^\top + (1-p_k){\mu_0^k}^\top] \\ &= p_k^2\mu_1^k{\mu_1^k}^\top + p_k(1-p_k)\mu_1^k{\mu_0^k}^\top + p_k(1-p_k)\mu_0^k{\mu_1^k}^\top + (1-p_k)^2\mu_0^k{\mu_0^k}^\top.\end{aligned} \tag{81}$$

From equation 79, equation 80, and equation 81, we have

$$\text{Cov}(\mathbb{E}[T_{Y_k}|G_k]) = p_k(1-p_k)\left[\mu_1^k{\mu_1^k}^\top - \mu_1^k{\mu_0^k}^\top - \mu_0^k{\mu_1^k}^\top + \mu_0^k{\mu_0^k}^\top\right]. \tag{82}$$

Simplifying the expression, we have:

$$\text{Cov}(\mathbb{E}[T_{Y_k}|B_k]) = p_k(1-p_k)[(\mu_1^k - \mu_0^k)(\mu_1^k - \mu_0^k)^\top]. \tag{83}$$

□

Lemma 21. *The covariance of the conditional expected sufficient statistic equals:*

$$\text{Cov}(\mathbb{E}[T_{Y_k} \mid G_k]) = p_k(1-p_k)(\mu_1^k - \mu_0^k)(\mu_1^k - \mu_0^k)^\top.$$

Futhermore, this equals the Fisher Information Matrix of the observed 1-bit log-likehood:

$$J_k(\theta) = -\nabla_\theta^2\ell_k(\theta) = V_k^\top \cdot \text{Cov}(\mathbb{E}[T_{Y_k} \mid G_k])\, V_k.$$

Proof. 1-bit log-likehood for $G_k \in \{0,1\}$:

$$\ell_k(\theta) = G_k \log p_k(\theta) + (1-G_k)\log\big(1-p_k(\theta)\big).$$

And gradient according θ:

$$\nabla_\theta\ell_k(\theta) = \frac{G_k}{p_k}\nabla_\theta p_k(\theta) - \frac{1-G_k}{1-p_k}\nabla_\theta p_k(\theta) = \frac{G_k - p_k(\theta)}{p_k(\theta)(1-p_k(\theta))} \cdot V_k^\top\nabla_\theta p_k(\theta).$$

So Hessian of FIM:

$$-\nabla_\theta^2\ell_k(\theta) = \mathbb{E}[\nabla_\theta\ell_k.(\nabla_\theta\ell_k)^\top] = \mathbb{E}[\frac{(G_k-p_k)^2}{p_k^2(1-p_k)^2}](\nabla_\theta p_k)(\nabla_\theta p_k)^\top.$$

Since $\mathbb{E}[(G_k - p_k)^2] = \text{Var}(G_k) = p_k(1-p_k)$:

$$J_k(\theta) = \frac{1}{p_k(1-p_k)} (\nabla_\theta p_k)(\nabla_\theta p_k)^\top. \tag{84}$$

Since $\mathbb{E}[(G_k - p_k)^2] = \text{Var}(G_k) = p_k(1-p_k)$:

$$J_k(\theta) = \frac{1}{p_k(1-p_k)} (\nabla_\theta p_k)(\nabla_\theta p_k)^\top. \tag{85}$$

In generalized linear/exponential models, the score direction of the expectation parameters links to parameter θ through the transformation matrix V_k:

$$\nabla_\theta p_k = p_k(1-p_k) V_k^\top (\mu_1^k - \mu_0^k).$$

Substitution into equation 85:

$$J_k(\theta) = \frac{1}{p_k(1-p_k)} \left[p_k(1-p_k) V_k^\top (\mu_1^k - \mu_0^k)\right] \left[p_k(1-p_k)(\mu_1^k - \mu_0^k)^\top V_k\right]. \tag{86}$$

So,

$$J_k(\theta) = V_k^\top \left[p_k(1-p_k)(\mu_1^k - \mu_0^k)(\mu_1^k - \mu_0^k)^\top\right] V_k.$$

We also have $\text{Cov}(\mathbb{E}[T_{Y_k}|G_k]) = p_k(1-p_k)(\mu_1^k - \mu_0^k)(\mu_1^k - \mu_0^k)^\top$.

So,

$$J_k(\theta) = V_k^\top \text{Cov}(\mathbb{E}[T_{Y_k}|G_k]) V_k.$$

□

Lemma 22. *Show that*

$$\nabla_\eta \phi(\eta) = \mathbb{E}_\eta[T_Y] \tag{87}$$

Proof.

$$p_Y(y;\eta) = h(y) \exp\big(\langle \eta, T_y \rangle - \phi(\eta)\big)$$

$$\int h(y) \exp\big(\langle \eta, T_y \rangle - \phi(\eta)\big)\, dy = 1$$

$$\Rightarrow \quad \phi(\eta) = \log \int h(y) \exp\big(\langle \eta, T_y \rangle\big)\, dy$$

$$\nabla_\eta \phi(\eta) = \frac{\nabla_\eta \int h(y) \exp\big(\langle \eta, T_y \rangle\big)\, dy}{\int h(y) \exp\big(\langle \eta, T_x \rangle\big)\, dy}$$

$$= \frac{\int h(y)\, T_y\, \exp\big(\langle \eta, T_y \rangle\big)\, dy}{\int h(y) \exp\big(\langle \eta, T_y \rangle\big)\, dy}$$

$$= \int h(y)\, T_y\, \exp\big(\langle \eta, T_y \rangle - \phi(\eta)\big)\, dy$$

$$= \int T_y\, p_Y(y;\eta)\, dy = \mathbb{E}_\eta[T_Y].$$

So, $\nabla_\eta \phi(\eta) = \mathbb{E}_\eta[T_Y]$. □

D EXPERIMENTAL DETAILS

This section provides additional detail about the conducted experiment, including parameters, numerical modification in algorithm and hardware description.

D.1 EXPERIMENT SETUP

We conduct our algorithm 1 in a toy environment, specifically, the reward distribution is Bernoulli distribution, and the inverse link function is a logistic function.
Additional to the configuration, which is described in the main text, we state here all the fixed parameters value used in the experiment. That is, the number of seeds ran in each configuration is 20, with confidence $\delta = 0.05$, k_μ is specifically set at $1/4$, while c_μ follows the formula $\min_k \sigma(\eta_k)(1 - \sigma(\eta_k))$, and η_k follows $V_k\theta^*$.

D.2 ADDITIONAL TECHNIQUE IN IMPLEMENTATION

Ridge regularization. We add a small ridge term $\lambda_{\mathrm{r}} = 10^{-4}$ in two places, both for numerical stability rather than as part of the method. First, we use $J_t + \lambda_{\mathrm{r}} I$ in place of J_t. In exact arithmetic the warm-up already makes J_t nonsingular: every weight $\sigma(\hat{\eta}_k)(1 - \sigma(\hat{\eta}_k))$ is strictly positive and the warm-up arms span $\mathbb{R}^d$. However, J_t is updated incrementally using past estimates; arms with large $|V_k\hat{\theta}_t|$ receive weights close to zero, so J_t can be severely ill-conditioned in floating point. The ridge bounds its condition number, which guarantees that the Cholesky factorisation used to compute $\|\cdot\|_{J_t^{-1}}$ succeeds and that all confidence widths are finite. Second, we add $\frac{\lambda_{\mathrm{r}}}{2}\|\theta\|_2^2$ to the negative log-likelihood. This makes the objective strongly convex, so every Newton system is well posed and the minimiser is unique. It also keeps the Hessian of the objective equal to the regularised J_t, so that the same matrix is used for estimation and for the confidence widths. The penalty is not what keeps the estimate bounded when the binary data are linearly separable; that role is played by the constraint $\|\theta\|_2 \le S$. In both places the ridge is several orders of magnitude smaller than the eigenvalues of J_t after warm-up (e.g., $\lambda_{\min}(J_E) \approx 0.3$ for $K = 100$, $d = 10$), so its effect on the estimates and confidence widths is negligible.

D.3 COMPUTE RESOURCES

All numerical simulations were executed on a dedicated Linux workstation equipped with an Intel(R) Core(TM) i7-11800H processor (16 cores, up to 2.3GHz) and 16GB of RAM.
In terms of computational cost, the experiment required approximately 2 hours of total execution time.

E LIMITATION AND FUTURE WORKS

Theoretically, the use of the confidence radius $C_t = O(\sqrt{m \log t \cdot \log(t^2/\delta)})$ introduces an extra $\log(1/\delta)$ factor relative to the lower bound, resulting in an upper-to-lower bound ratio of $O(K \log(1/\delta))$. Closing this gap remains an interesting direction for future research.

Another limitation is that the thresholds γ_k are fixed throughout the algorithm. While one-bit feedback does not merely lose information abstractly, its sample-complexity price is captured by the censored Fisher information in the hard gap directions. This also highlights a natural limitation of the current formulation: the fixed thresholds γ_k may differ from the instance-optimal thresholds γ_k^* appearing in the lower bound, potentially inflating the sample-complexity constant. Adaptively optimizing the thresholds jointly with arm allocation to maximize the information available for resolving hard reward gaps remains an open problem.

Finally, while our experiments provide empirical support for the theoretical analysis, the current evaluation is limited to relatively simple environment designs. An interesting direction for future work is to evaluate the algorithm in more complex environments and under a broader range of latent reward distributions.