Improved regret for Stochastic Gradient Bandits

Dorian Baudry

Research note, version 1
16 November 2026

Abstract

We present a detailed analysis of discrete-time stochastic gradient bandits based on the Lyapunov argument of Jia and Ouyang. Keeping the drift margin and the gap differences explicit improves the admissible learning-rate range and the bound on failure regret. For identical suboptimal gaps, the resulting failure bound is of order $\frac{K}{\eta}$, improving over the previous $(1+C)^{K-1}$ bound in this setting. A separate analysis of the post-convergence term establishes the exact expected asymptotic regret coefficient $\frac{K-1}{K\eta}$ for each fixed instance and learning rate satisfying the stated moment and learning-rate conditions.

Suggested citation. Dorian Baudry. *Improved regret for Stochastic Gradient Bandits*. Research note, version 1, 16 November 2026. https://dbaudry.github.io/research-notes/sgb-regret/

Introduction

This note builds on the finite-time expected-regret analysis of Jia and Ouyang (2026, Theorem 3) for discrete-time stochastic gradient bandits. We develop their Lyapunov argument in detail, explaining the role of the subset potentials and tracking the dependence of the bounds on the reward gaps and the learning rate.

In doing so, we also modify several arguments to improve the final result or make its parameter dependence explicit. We identify these improvements throughout the note. All other arguments and derivations follow Jia and Ouyang's work. In particular, we retain their subset Lyapunov functions and exponential one-step estimate. The regret decomposition into a post-convergence term and a failure term is due to Baudry et al. (2025, equation (7)) and also underlies Jia and Ouyang's proof.

For K arms and a fixed learning rate η, the most notable improvements are the following.

(1) *Sharper post-convergence analysis.* We prove sharper finite-time bounds in expectation and identify the exact logarithmic coefficient $\frac{K-1}{K\eta}$ in the expected asymptotics under the stated learning-rate condition (Lemma 3.1). For fixed learning rates satisfying (2.1), this establishes the cumulative-regret coefficient predicted by Baudry et al. (2025, Conjecture 3, Appendix G.3).

(2) *Direct control of failure regret.* By retaining the individual gaps and the full singleton drift, the Lyapunov analysis bounds the failure-regret term itself. This avoids Jia and Ouyang's intermediate upper bound through a sum of squared suboptimal sampling probabilities.

(3) *Interpolation between gap and learning-rate regimes.* We retain the drift margin and the gap spread in the subset-potential weights. The resulting bounds interpolate between identical and unequal suboptimal gaps. In particular, for identical gaps and $\eta K\Gamma_K(\eta) \leq \Delta_{\min}$, where $\Gamma_K(\eta)$ is the moment bound defined below, the additive failure bound is of order $\frac{K}{\eta}$ (Theorem 2.1). This recovers a logarithmic regret bound comparable in form to the identical-gap result of Baudry et al. (2025, Theorem 4).

For general gap configurations, the bounds in the current version of this work can still depend exponentially on the number of arms through the initial value of the subset Lyapunov function.

The main improvements are stated in Theorem 2.1 and Lemma 3.1. Section 1 recalls the model and Jia–Ouyang's result. Section 2 develops the refined failure bound, and Section 3 proves the sharper post-convergence bounds and the matching asymptotic lower bound.

1 A finite-time upper bound for discrete-time SGB

In this section we present the upper bound of Jia and Ouyang (2026, Theorem 3). We first recall the algorithm and introduce the notation used throughout this paper. For completeness, we restate exactly the theorem obtained by Jia & Ouyang with our notation, before introducing our improved result and its detailed proof.

1.1 Algorithm, notation, and assumption

Here $t \in \{1, 2, \ldots\}$ indexes discrete rounds. We use a filtration $(\mathcal{F}_t)_{t\geq 1}$, where $\mathcal{F}_t$ contains all actions and rewards observed before round t and any other information available when the policy and baseline for that round are chosen. The logit vector $\theta_t \in \mathbb{R}^K$ and the real baseline b_t are $\mathcal{F}_t$-measurable. Stochastic Gradient Bandits (`SGB`) uses the softmax policy

$$\pi_{t,i} = \frac{\exp(\theta_{t,i})}{\sum_{j=1}^{K} \exp(\theta_{t,j})}, \quad \text{and we write} \quad \pi_t = (\pi_{t,1}, \ldots, \pi_{t,K})^\top. \tag{1.1}$$

At each round, `SGB` samples A_t with $\mathbb{P}(A_t = i \mid \mathcal{F}_t) = \pi_{t,i}$ and observes a reward Y_t that satisfies $\mathbb{E}[Y_t \mid \mathcal{F}_t, A_t = i] = \mu_i$. The conditional reward law may depend on the past. The baseline may also depend on the past, but is chosen before the current action and reward. For $X_t = Y_t - b_t$ and a fixed $\eta > 0$, the update is

$$\theta_{t+1} = \theta_t + \eta X_t(e_{A_t} - \pi_t), \qquad \theta_1 = 0, \tag{1.2}$$

where e_i is the ith coordinate vector and $\mu = (\mu_1, \ldots, \mu_K)^\top$ is the vector of reward means. We assume $K \geq 2$ and $\mu_1 > \max_{i>1} \mu_i$. For $i > 1$, we define the suboptimality gap $\Delta_i := \mu_1 - \mu_i$, and we write $\Delta_{\min} = \min_{i>1} \Delta_i$ and $\Delta_{\max} = \max_{i>1} \Delta_i$. At round t, we define the instantaneous pseudo-regret $\Delta_t^\pi = \sum_{i=2}^{K} \Delta_i \pi_{t,i}$, which is the expected suboptimality of the selected arm conditional on $\mathcal{F}_t$. For each integer $T \geq 1$, the expected cumulative pseudo-regret is

$$R_T := \mathbb{E}\left[\sum_{t=1}^{T} (\mu_1 - \mu_{A_t})\right] = \mathbb{E}\left[\sum_{t=1}^{T} \Delta_t^\pi\right]. \tag{1.3}$$

Moment assumption. The proofs of Baudry et al. (2025) and Jia and Ouyang (2026) use second-order bounds for exponential increments. In particular, a bound on the second-order Taylor remainder for an exponential increment $\eta q X_t$ involves $X_t^2 \exp(\eta q |X_t|)$ (see Lemma 2.6). We therefore assume that, for each $q \in \{1, K\}$, there is a finite deterministic constant $\Gamma_q(\eta) \geq 0$ such that, for every round t and arm i,

$$\mathbb{E}\Big[X_t^2 \exp(\eta q |X_t|) \mid \mathcal{F}_t, A_t = i\Big] \leq \Gamma_q(\eta). \tag{1.4}$$

Thus $\Gamma_q(\eta)$ is a uniform second-moment bound for the baseline-adjusted reward, with an exponential factor controlling the Taylor remainder.

$\Gamma_q(\eta)$ is finite if the conditional distribution of X_t is *light-tailed*. More precisely, the moment-generating function of X_t needs to be finite in a neighborhood of zero (Foss et al., 2013, Definition 2.3).

For (1.4), we need uniform control of the conditional laws: if there are deterministic constants $\lambda > 0$ and $C < \infty$ such that $\mathbb{E}[\exp(\lambda|X_t|) \mid \mathcal{F}_t, A_t = i] \le C$ for every t and i, then (1.4) holds whenever $\eta K < \lambda$. Indeed, for $q \in \{1, K\}$, the function $x \mapsto x^2 \exp(-(\lambda - \eta q)|x|)$ is bounded.

The proof of Jia and Ouyang (2026) uses $q = 1$ for $\exp(-\theta_{t,i})$ and $q = K$ for the subset potentials, whose logarithms change by at most $\eta K|X_t|$ in one update. The bound at $q = K$ also controls the moment at $q = 1$, so we may take $\Gamma_1(\eta) = \Gamma_K(\eta)$. Keeping a smaller bound at $q = 1$ can reduce the logarithmic term in the regret bound.

For example, if $|X_t| \le M$ almost surely for a deterministic $M > 0$, the exponential factor lies between 1 and $\exp(\eta q M)$, which tends to 1 as $\eta \downarrow 0$ for fixed q. We may then take $\Gamma_q(\eta) = M^2 \exp(\eta q M)$, which tends to M^2. This includes rewards and predictable baselines in $[0, 1]$, with $M = 1$. In such case, (1.4) does not impose any constraint on the learning rate: Γ is always finite even though it grows exponentially with η.

1.2 Upper bound of Jia and Ouyang

We restate Jia and Ouyang (2026, Theorem 3) using the notation and moment assumption of Section 1.1.

Theorem 1.1 (Jia and Ouyang)**.** *Consider the* `SGB` *update* (1.2) *under the assumptions of Section 1.1. Assume that the fixed learning rate satisfies*

$$\eta K \Gamma_K(\eta) \le \Delta_{\min}. \tag{1.5}$$

Then, as $t \to \infty$*,*

$$\pi_{t,1} \longrightarrow 1, \qquad \pi_{t,i} \longrightarrow 0 \quad (2 \le i \le K) \qquad \textit{almost surely.} \tag{1.6}$$

Moreover, for every integer $T \ge 1$*,*

$$R_T \le \frac{K-1}{\eta} \log\left(1 + \frac{\eta\Delta_{\max} + \frac{\eta^2 \Gamma_1(\eta)}{2}}{K} T\right) + \frac{\Delta_{\max}(K-1)}{\eta\Delta_{\min}} \frac{(1+C)^{K-1} - 1}{C}, \tag{1.7}$$

where $C = \frac{2\Delta_{\max}}{\Delta_{\min}} + \frac{1}{K}$.

The notation corresponds to $d = K$, $\ell = \eta$, $\phi_{t-1} = \theta_t$, and $r_{t-1} = \Delta_t^\pi$ in Jia and Ouyang (2026), where rounds start at zero. Their auxiliary constants called K and C in equation (65) are $\Delta_{\max} + \frac{\Delta_{\min}}{2K}$ and $\frac{2\Delta_{\max}}{\Delta_{\min}} + \frac{1}{K}$, respectively.

For fixed problem parameters and learning rate, the asymptotic scaling in the horizon is $\frac{K-1}{\eta} \log(T)$ while the constant term can depend exponentially on K, even for equal gaps. In the next section, we focus our improvements on the failure regret term, and in Section 3 we propose a new analysis technique to reach the $\frac{K-1}{K\eta} \log(T)$ asymptotic scaling conjecture by Baudry et al. (2025).

2 Improved result and detailed analysis

We keep the drift margin explicit in Jia–Ouyang's proof to obtain the admissible range $\eta < \frac{2\Delta_{\min}}{K}$ (ignoring second-moment factors), as conjectured. The choice $\varepsilon = \frac{1}{2}$ below recovers their learning-rate condition. Our bound on the cross terms gives smaller potential weights, and we control the failure term directly. The improvements in this section follow from keeping track of the drift margin

and the cross terms throughout the proof. The next section develops a separate argument for the post-convergence term. The proof sketch below explains the potential construction and the estimates that determine the bound.

Theorem 2.1 (Regret bound with an explicit drift margin)**.** *Consider the* `SGB` *update* (1.2) *under the assumptions above, including* (1.4)*. Fix* $\varepsilon \in (0,1)$ *and assume that the fixed learning rate satisfies*

$$\eta K \Gamma_K(\eta) \le 2(1-\varepsilon)\Delta_{\min}. \tag{2.1}$$

We define the constants

$$D_\eta = \Delta_{\max} - \Delta_{\min} + \frac{\eta \Gamma_K(\eta)}{2}, \qquad C_\varepsilon = \frac{D_\eta}{\varepsilon \Delta_{\min}}. \tag{2.2}$$

Then every integer $T \ge 1$ *satisfies*

$$R_T \le \frac{K-1}{\eta} \log\left(1 + \frac{\eta\Delta_{\max} + \frac{\eta^2 \Gamma_1(\eta)}{2}}{K} T\right) + \frac{1}{\eta\varepsilon} \frac{(1+C_\varepsilon)^{K-1} - 1}{C_\varepsilon}. \tag{2.3}$$

The same fixed η *and* ε *are used at every horizon. The suboptimal means need not be distinct.*

Allowing any $\varepsilon \in (0,1)$ covers the range $\eta K \Gamma_K(\eta) < 2\Delta_{\min}$. The deterministic moment bounds in (1.4) may depend on η, so the learning-rate condition (2.1) is generally implicit. For example, if $|X_t| \le M$ almost surely for a deterministic $M > 0$, a sufficient learning-rate condition is

$$\eta K M^2 \exp(\eta K M) \le 2(1-\varepsilon)\Delta_{\min},$$

which holds, for example, whenever

$$0 < \eta \le \frac{2(1-\varepsilon)\Delta_{\min}}{KM^2} \exp\left(-\frac{2(1-\varepsilon)\Delta_{\min}}{M}\right).$$

Indeed, this choice gives $\eta K M \le \frac{2(1-\varepsilon)\Delta_{\min}}{M}$, so $\eta K M^2 \exp(\eta K M) \le 2(1-\varepsilon)\Delta_{\min}$.

Short discussion of the result. This result partially resolves the conjecture of Baudry et al. (2025, Appendix D.3): the admissible range is $\eta < \frac{2\Delta_{\min}}{K}$ (ignoring second-moment factors). The gap spread in D_η improves the constant from Jia–Ouyang's Lyapunov analysis. For equal gaps and $\varepsilon = \frac{1}{2}$, we have $C_\varepsilon = \frac{\eta\Gamma_K(\eta)}{\Delta_{\min}} \le \frac{1}{K}$, so the additive term is at most $\frac{2\exp(1)(K-1)}{\eta}$. For general gaps, the potential factor can still depend exponentially on the number of arms. When $\Delta_{\max} \gg \Delta_{\min}$, it retains a dependence in $\frac{1}{\Delta_{\min}^{K-2}}$ through $V_{C_\varepsilon}(0)$. It is left open by the authors to determine whether this dependence is intrisic or can be improved with a sharper analysis.

A second source of potential untighteness, athough less severe, it the $K-1$ factor in front of the log term. We recall that Baudry et al. (2025, Conjecture 3, Appendix G.3) conjectured that the leading term in the logarithmic regime is $\frac{K-1}{K\eta}\log(T)$, which should provide a more standard scaling $\frac{K-1}{\Delta_{\min}}\log(T)$ close the largest admissible learning rate. This would be near optimal in instances with $\Delta_{\min} = \Delta_{\max}$.

Breaking down the potential construction The regret decomposition (2.4) below isolates the failure term $\mathbb{E}\left[\sum_{t=1}^{T}(1-\pi_{t,1})\Delta_t^\pi\right]$. To bound this term uniformly in T, we seek a nonnegative potential whose decrease in conditional expectation is at least $\eta\varepsilon(1-\pi_{t,1})\Delta_t^\pi$ at each round. Summing these decreases will bound the accumulated failure regret by the initial value of the potential. We explain how the probability ratio used by Baudry et al. (2025, proof of Theorem 1) for two arms leads to the subset potential of Jia and Ouyang (2026, equations (65) and (75)–(76)).

For two arms, consider $h_{\{2\}}(\theta_t) = \exp(\theta_{t,2} - \theta_{t,1}) = \frac{\pi_{t,2}}{\pi_{t,1}}$. This ratio is large when the optimal arm receives little probability. The update (1.2) gives a conditional mean increment $-2\eta\Delta_2\pi_{t,1}\pi_{t,2}$ for its logarithm, with the baseline cancelling. With the moment bound, Taylor's formula as in Lemma 2.6 then gives

$$\mathbb{E}\left[h_{\{2\}}(\theta_{t+1}) \mid \mathcal{F}_t\right] - h_{\{2\}}(\theta_t) \le -2\eta\Delta_2\pi_{t,2}^2 + 2\eta^2\Gamma_2(\eta)\pi_{t,2}^2 = -2\eta\pi_{t,2}^2(\Delta_2 - \eta\Gamma_2(\eta)).$$

The first term is the decrease due to the reward gap, and the second bounds the rewards second moment, including a Taylor's approximation remainder $\Gamma_2(\eta)$. Below, we absorb these second-order terms using the learning-rate condition. Both contain the same factor $\pi_{t,2}^2$: multiplying the factor $\pi_{t,1}\pi_{t,2}$ in the logit calculation by $h_{\{2\}}(\theta_t)$ cancels $\pi_{t,1}$. Moreover, $\Delta_2\pi_{t,2}^2 = (1-\pi_{t,1})\Delta_t^\pi$ is exactly the instantaneous failure regret. Under (2.1), $\Delta_2 - \eta\Gamma_2(\eta) \ge \varepsilon\Delta_2$. The gap term therefore dominates the remainder and provides the required decrease.

For several arms, the natural first attempt is to sum the ratios $h_{\{i\}}(\theta_t) = \frac{\pi_{t,i}}{\pi_{t,1}}$ over $i > 1$. For each singleton, Lemma 2.6 gives

$$\frac{\mathbb{E}\left[h_{\{i\}}(\theta_{t+1}) \mid \mathcal{F}_t\right] - h_{\{i\}}(\theta_t)}{\eta} \le \underbrace{-\pi_{t,i}\left(\Delta_i\pi_{t,i} + \Delta_t^\pi\right)}_{\text{negative drift}} + \underbrace{h_{\{i\}}(\theta_t)\sum_{\substack{j>1\\ j\neq i}}\pi_{t,i}\pi_{t,j}(\Delta_j - \Delta_i)}_{\text{cross terms}}$$
$$+ \underbrace{\frac{\eta\Gamma_K(\eta)}{2}h_{\{i\}}(\theta_t)\left(4\pi_{t,1}\pi_{t,i} + (\pi_{t,1}+\pi_{t,i})\sum_{\substack{j>1\\ j\neq i}}\pi_{t,j}\right)}_{\text{second-order term}}.$$

The identity $h_{\{i\}}(\theta_t)\pi_{t,1} = \pi_{t,i}$ produces the displayed negative drift. Retaining its full expression is our refinement: summing $\pi_{t,i}\Delta_t^\pi$ over $i > 1$ gives exactly $(1-\pi_{t,1})\Delta_t^\pi$. Under (2.1), for small enough learning rate, we can absorb the second-order terms involving arm 1 into this negative drift as in the two-arms case. The remaining cross terms have coefficients at most D_η. Summing over the singletons and dropping additional negative squared probability terms therefore gives

$$\sum_{i=2}^{K}\left(\mathbb{E}\left[h_{\{i\}}(\theta_{t+1}) \mid \mathcal{F}_t\right] - h_{\{i\}}(\theta_t)\right) \le -\eta\varepsilon\underbrace{(1-\pi_{t,1})\Delta_t^\pi}_{\text{failure regret}} + \eta D_\eta\underbrace{\sum_{2\le i<j\le K}\frac{\pi_{t,i}\pi_{t,j}(\pi_{t,i}+\pi_{t,j})}{\pi_{t,1}}}_{\text{cross terms}}.$$

The desired failure regret already appears with a negative sign, but for $K \ge 3$ we have a positive remainder, that can become arbitrarily large as $\pi_{t,1}$ tends to zero while two suboptimal arms keep positive probability. Since $(1-\pi_{t,1})\Delta_t^\pi \le \Delta_{\max}$, failure regret cannot always compensate this term. We therefore need additional potentials whose negative drift has the same form as this remainder. The form of the cross terms suggests adding products of two singleton ratios, then repeating the

construction for larger subsets. Let $I = \{2, \dots, K\}$ and write $A_I^c = I \setminus A$ for the complement of $A \subseteq I$ within the suboptimal arms. For each nonempty $A \subseteq I$, define

$$h_A(\theta_t) = \prod_{i \in A} \frac{\pi_{t,i}}{\pi_{t,1}} = \exp\left(\sum_{i \in A} (\theta_{t,i} - \theta_{t,1}) \right).$$

The reason for using these particular products is the identity

$$h_A(\theta_t)\pi_{t,i}\pi_{t,j} = h_{A \cup \{j\}}(\theta_t)\pi_{t,1}\pi_{t,i}, \qquad i \in A, \quad j \in A_I^c.$$

A positive interaction involving an arm outside A therefore has exactly the probability factors appearing in the negative drift of the potential for $A \cup \{j\}$. For example, the terms left by $h_{\{i\}}$ and $h_{\{j\}}$ can be absorbed by adding a sufficiently weighted $h_{\{i,j\}}$. Its drift can in turn involve an arm outside $\{i, j\}$, which calls for a size-three potential. Continuing this procedure leads to every nonempty subset of suboptimal arms. It stops at $A = I$, since there is then no outside suboptimal arm.

It remains to give each larger subset enough weight to absorb the terms arriving from smaller subsets. Jia–Ouyang use geometric weights and define, for $C > 0$,

$$V_C(\theta) = \sum_{\varnothing \neq A \subseteq I} C^{|A|-1} h_A(\theta).$$

To see how the weights work, fix a subset D of size $r \geq 2$ and $i \in D$. A term proportional to $h_D(\theta_t)\pi_{t,1}\pi_{t,i}$ arrives from each $A = D_{\setminus\{j\}}$ with $j \in D_{\setminus\{i\}}$, giving $r - 1$ incoming terms. The negative drift supplied by h_D has a factor $(r+1)\Delta_i$ after the Taylor remainder is absorbed using the learning-rate condition. Passing from size $r - 1$ to size r multiplies the weight by C. With the bounds used in our proof, the coefficient to control is consequently $D_\eta(r-1) - \varepsilon C(r+1)\Delta_i$. The choice $C = C_\varepsilon = \frac{D_\eta}{\varepsilon \Delta_{\min}}$ makes every such coefficient nonpositive.

The larger subsets thus absorb the cross terms while leaving the negative singleton drift available to control the failure regret. The detailed proof below establishes

$$\mathbb{E}[V_{C_\varepsilon}(\theta_{t+1}) \mid \mathcal{F}_t] - V_{C_\varepsilon}(\theta_t) \leq -\eta\varepsilon(1 - \pi_{t,1})\Delta_t^\pi.$$

Taking expectations and summing gives the desired bound $\mathbb{E}\left[\sum_{t=1}^T (1 - \pi_{t,1})\Delta_t^\pi\right] \leq \frac{V_{C_\varepsilon}(0)}{\eta\varepsilon}$. The subset structure and the absorption of successive subset sizes are Jia–Ouyang's construction. Our direct control of the failure term and smaller choice of weights use the same mechanism. Finally, every $h_A(0)$ equals one, so $V_C(0) = \frac{(1+C)^{K-1}-1}{C}$: the initial sum over subsets is where this argument can incur an exponential dependence on K if C is bounded away from zero.

Proof of Theorem 2.1. Throughout the proof, we write $\Gamma_q = \Gamma_q(\eta)$ and $\mathbb{E}_t[W] = \mathbb{E}[W \mid \mathcal{F}_t]$ for simplicity. We first prove the theorem by separating the two terms in the regret. The exponential estimate and the finite-round integrability used in this proof are established in Section 2.2 below.

We fix an integer $T \geq 1$. As a preliminary step, Lemma 2.7 justifies the expectations and finite sums below.

1. Regret decomposition. The proofs follows the post-convergence/failure regret decomposition of Baudry et al. (2025, equation (7)). The first term can be expressed through the expected logit

of the optimal arm. To see this, we write $\mu_t^\pi = \sum_i \pi_{t,i}\mu_i$. Since the baseline is chosen before the action, the update gives

$$\begin{aligned}\mathbb{E}_t[\theta_{t+1,1} - \theta_{t,1}] &= \eta \sum_i \pi_{t,i}(\mu_i - b_t)(\mathbf{1}\{i=1\} - \pi_{t,1}) \\ &= \eta\pi_{t,1}(\mu_1 - \mu_t^\pi) = \eta\pi_{t,1}\Delta_t^\pi.\end{aligned}$$

Taking expectations, summing over $t = 1, \dots, T$, and using $\theta_{1,1} = 0$, we obtain the exact regret decomposition

$$\begin{aligned}R_T &= \mathbb{E}\left[\sum_{t=1}^T \pi_{t,1}\Delta_t^\pi\right] + \mathbb{E}\left[\sum_{t=1}^T (1-\pi_{t,1})\Delta_t^\pi\right] \\ &= \frac{\mathbb{E}[\theta_{T+1,1}]}{\eta} + \mathbb{E}\left[\sum_{t=1}^T (1-\pi_{t,1})\Delta_t^\pi\right]. \end{aligned} \tag{2.4}$$

We call these two summands the post-convergence term and the failure term, respectively. We will bound the first logarithmically in T and the second uniformly in T.

Jia–Ouyang further bound the failure term using $\Delta_t^\pi \le \Delta_{\max}(1-\pi_{t,1})$ and Cauchy–Schwarz, obtaining (Jia and Ouyang, 2026, proof of Theorem 3, after equation (82))

$$\mathbb{E}\left[\sum_{t=1}^T (1-\pi_{t,1})\Delta_t^\pi\right] \le (K-1)\Delta_{\max}\,\mathbb{E}\left[\sum_{t=1}^T \sum_{i=2}^K \pi_{t,i}^2\right]. \tag{2.5}$$

Our refinement keeps the full failure term in (2.4). By retaining the individual gaps and the negative cross terms in the singleton drift, we will control it directly, without using (2.5). We therefore work directly on the regret equality above.

2. The post-convergence term. In this segment of the proof we follow Jia-Ouyang's analysis, and we keep our improvement for next section. Following their proof, conservation of the sum of the logits lets us bound $\mathbb{E}[\theta_{T+1,1}]$ through $\mathbb{E}[\exp(-\theta_{T+1,i})]$ for the suboptimal arms. Indeed, summing the coordinates of (1.2) and using $\theta_1 = 0$ shows, pathwise,

$$\sum_{i=1}^K \theta_{t+1,i} = \sum_{i=1}^K \theta_{t,i} + \eta X_t\left(1 - \sum_{i=1}^K \pi_{t,i}\right) = 0, \quad \text{so} \quad \theta_{T+1,1} = -\sum_{i>1}\theta_{T+1,i}. \tag{2.6}$$

For $i > 1$, the vector $v = -e_i$ has oscillation one, so we can apply Lemma 2.6 with $q = 1$. We write $\mathsf{C}(\pi_t) = \operatorname{diag}(\pi_t) - \pi_t\pi_t^\top$. We have

$$-e_i^\top \mathsf{C}(\pi_t)\mu = \pi_{t,i}(\mu_t^\pi - \mu_i) \le \Delta_i\pi_{t,i}, \qquad e_i^\top \mathsf{C}(\pi_t)e_i = \pi_{t,i}(1-\pi_{t,i}) \le \pi_{t,i}.$$

Consequently, for any step $t \ge 1$ it holds that

$$\mathbb{E}_t[\exp(-\theta_{t+1,i})] - \exp(-\theta_{t,i}) \le \exp(-\theta_{t,i})\pi_{t,i}\left(\eta\Delta_i + \frac{\eta^2\Gamma_1}{2}\right). \tag{2.7}$$

We write $Z_t = \sum_{j=1}^K \exp(\theta_{t,j})$. The arithmetic–geometric mean inequality and (2.6) imply

$$Z_t \ge K\exp\left(\frac{1}{K}\sum_{j=1}^K \theta_{t,j}\right) = K, \qquad \text{so} \qquad \exp(-\theta_{t,i})\pi_{t,i} = \frac{1}{Z_t} \le \frac{1}{K}.$$

Taking expectations in (2.7) and summing over $t = 1, \dots, T$, with $\exp(-\theta_{1,i}) = 1$, gives

$$\mathbb{E}[\exp(-\theta_{T+1,i})] \le 1 + a_\eta T, \qquad \text{with} \qquad a_\eta := \frac{\eta \Delta_{\max} + \frac{\eta^2 \Gamma_1}{2}}{K}. \tag{2.8}$$

Since log is concave and $\theta_{T+1,i}$ is integrable,

$$\mathbb{E}[-\theta_{T+1,i}] = \mathbb{E}\big[\log(\exp(-\theta_{T+1,i}))\big] \le \log\left(\mathbb{E}[\exp(-\theta_{T+1,i})]\right) \le \log(1 + a_\eta T).$$

Finally, $\theta_{T+1,1} = -\sum_{i=2}^{K} \theta_{T+1,i}$ gives

$$\mathbb{E}[\theta_{T+1,1}] \le (K-1)\log(1 + a_\eta T). \tag{2.9}$$

Using the identity established in part 1, we conclude that

$$\mathbb{E}\left[\sum_{t=1}^{T} \pi_{t,1} \Delta_t^\pi\right] = \frac{\mathbb{E}[\theta_{T+1,1}]}{\eta} \le \frac{K-1}{\eta} \log(1 + a_\eta T). \tag{2.10}$$

In the next section we show that this term can be sharpened to $\frac{K-1}{K\eta}\log(T)$, keeping only the leading asymptotic scaling here, under a different proof technique that bounds $\mathbb{E}[\theta_{1,T+1}]$ directly.

3. The failure term. By (2.4), it remains to bound the full failure term. We retain the individual gaps in the drift of Jia–Ouyang's potentials to recover this term directly. The identity $\exp(\theta_{t,i} - \theta_{t,1})\pi_{t,1} = \pi_{t,i}$ suggests using exponentials of logit differences. With more than two arms, their drifts contain cross terms involving other suboptimal arms. Jia–Ouyang's subset potentials absorb these terms into the negative drift of potentials associated with larger subsets. For $I = \{2, \dots, K\}$, each nonempty $A \subseteq I$, and $C > 0$, we define

$$h_A(\theta) = \exp\left(\sum_{i \in A} (\theta_i - \theta_1)\right), \qquad V_C(\theta) = \sum_{\varnothing \ne A \subseteq I} C^{|A|-1} h_A(\theta). \tag{2.11}$$

If $m = |A|$, then $h_A = f_{v_A}$ for $v_{A,1} = -m$ and $v_{A,i} = \mathbf{1}\{i \in A\}$ for $i > 1$. In particular, $\operatorname{osc} v_A = m + 1 \le K$, so Lemma 2.7 gives integrability of every $h_A(\theta_t)$ and of $V_C(\theta_t)$.

We fix a round t and abbreviate $h_A = h_A(\theta_t)$. Applying Lemma 2.6 with $q = K$ and $v = v_A$ gives

$$\mathbb{E}_t[h_A(\theta_{t+1})] - h_A \le h_A \left(\eta v_A^\top \mathsf{C}(\pi_t)\mu + \frac{\eta^2 \Gamma_K}{2} v_A^\top \mathsf{C}(\pi_t) v_A\right).$$

Since $\eta h_A > 0$, we can divide by this factor to obtain

$$\frac{\mathbb{E}_t[h_A(\theta_{t+1})] - h_A}{\eta h_A} \le \underbrace{v_A^\top \mathsf{C}(\pi_t)\mu}_{\text{Drift term}} + \underbrace{\frac{\eta \Gamma_K}{2} v_A^\top \mathsf{C}(\pi_t) v_A}_{\text{Variance term}}.$$

We expand these terms in Lemma 2.3 below. The result show that failure regret is captured by the drift of the singleton subset, and that positive cross terms are captured by the drift of higher-order subsets. This motivates the multi-level Lyapunov function. The resulting bound results from an appropriate grouping of all the terms, and a telescoping argument. We capture these steps in the following lemma, that we prove in the next section.

Lemma 2.2 (Failure regret bound)**.** *Under the assumptions of Theorem 2.1, for every integer* $T \geq 1$,

$$\mathbb{E}\left[\sum_{t=1}^{T}(1-\pi_{t,1})\Delta_t^{\pi}\right] \leq \frac{V_{C_\varepsilon}(0)}{\eta\varepsilon} = \frac{1}{\eta\varepsilon}\cdot\frac{(1+C_\varepsilon)^{K-1}-1}{C_\varepsilon}.$$

We prove this lemma in Section 2.1, using a subset drift estimate whose coordinate computations are deferred to Section 2.1.1. Combining Lemma 2.2 with (2.4) and (2.10) proves (2.3). □

2.1 Proof of the failure regret bound

We first state the one-step estimate for each subset potential, keeping the gap differences before replacing them by D_η. We then sum the estimates over subsets to prove Lemma 2.2. In this subsection, we write $\Gamma_q = \Gamma_q(\eta)$ and $\mathbb{E}_t[W] = \mathbb{E}[W \mid \mathcal{F}_t]$. We use the potentials in (2.11), with $I = \{2, \dots, K\}$ and $A_I^c = I \setminus A$. For $S \subseteq I$, we write $\pi_{t,S} = \sum_{i\in S}\pi_{t,i}$ and $\Delta_t(S) = \sum_{i\in S}\pi_{t,i}\Delta_i$. For nonempty S, the gap under the policy conditioned on S is $\overline{\Delta}_t(S) = \frac{\Delta_t(S)}{\pi_{t,S}}$. We set $\overline{\Delta}_t(\varnothing) = 0$.

We now introduce the main result motivating the choice of the potential.

Lemma 2.3 (Subset drift estimate)**.** *Under the assumptions of Theorem 2.1, for every* $t \geq 1$ *and nonempty* $A \subseteq I$*, set* $m = |A|$ *and* $h_A = h_A(\theta_t)$*. Then,*

$$\begin{aligned}\frac{\mathbb{E}_t[h_A(\theta_{t+1})] - h_A}{\eta h_A} \leq{}& -\varepsilon\pi_{t,1}\left(m\Delta_t^{\pi} + \Delta_t(A)\right) - \frac{\eta m\Gamma_K}{2}\pi_{t,1}\pi_{t,A_I^c} \\ &+ \pi_{t,A}\pi_{t,A_I^c}\left(\overline{\Delta}_t(A_I^c) - \overline{\Delta}_t(A) + \frac{\eta\Gamma_K}{2}\right).\end{aligned}$$

The proof of Lemma 2.3 is postponed to Section 2.1.1, and we now use this result to prove the bound on the failure regret.

Proof of Lemma 2.2. Fix an integer $T \geq 1$. At each round t, we abbreviate $h_A = h_A(\theta_t)$ and $m = |A|$. We use Lemma 2.3 to bound the interactions between subsets and match them to the negative drift of larger subsets. Factorizing the η terms gives

$$\pi_{t,A_I^c}(\pi_{t,A} - m\pi_{t,1})\,\frac{\eta\Gamma_K}{2},$$

which is nonpositive when arm 1 has the largest probability. We discard the negative term outside the bracket for the simplified bound (2.12), and retain it in (2.15).

Simplification of Lemma 2.3. After exhibiting a similar recursion to the one displayed in the lemma, Jia–Ouyang proceed by upper bounding the gap difference by $\Delta_{\max}$ uniformly across all subsets. We introduce a modest improvement by remarking that $\Delta_{\max} - \Delta_{\min}$ can be used instead, which largely improves the regret bound in equal-gap instances, as discussed in Section 2. Let

$$D_\eta \coloneqq \Delta_{\max} - \Delta_{\min} + \frac{\eta\Gamma_K}{2}.$$

We then obtain the simplified bound

$$\frac{\mathbb{E}_t[h_A(\theta_{t+1})] - h_A}{\eta\varepsilon} \leq -h_A\pi_{t,1}\left(\sum_{i\in A}\Delta_i\pi_{t,i} + m\Delta_t^{\pi}\right) + \frac{D_\eta}{\varepsilon}h_A\sum_{\substack{i\in A\\ j\in A_I^c}}\pi_{t,i}\pi_{t,j}. \tag{2.12}$$

Isolating the failure term. For all sets $A = \{i\}$, the weight $C^{|A|-1}$ equals one and $h_{\{i\}}\pi_{t,1} = \pi_{t,i}$. Their cumulative negative drift captures the full failure term, since

$$\sum_{i=2}^{K} h_{\{i\}}\pi_{t,1}(\Delta_i\pi_{t,i} + \Delta_t^\pi) = \sum_{i=2}^{K} \Delta_i\pi_{t,i}^2 + (1-\pi_{t,1})\Delta_t^\pi \geq (1-\pi_{t,1})\Delta_t^\pi.$$

This is the refinement of Jia–Ouyang's argument: retaining the individual gaps and the full singleton drift produces the failure term itself. For each $i > 1$, $\pi_{t,i} \leq 1 - \pi_{t,1}$, so

$$0 \leq \sum_{i=2}^{K} \Delta_i\pi_{t,i}^2 \leq (1-\pi_{t,1}) \sum_{i=2}^{K} \Delta_i\pi_{t,i} = (1-\pi_{t,1})\Delta_t^\pi.$$

Thus the sum of the two negative singleton terms has magnitude between one and two times the instantaneous failure regret. Dropping the square term loses at most a factor two in this part of the drift.

The positive terms in (2.12) still remain. Summing the terms of Lemma 2.3 over singletons, bounding the interaction terms by D_η, and pairing the terms indexed by (i,j) and (j,i) gives the normalized drift bound

$$\frac{1}{\eta\varepsilon}\sum_{i=2}^{K}\Big(\mathbb{E}_t\Big[h_{\{i\}}(\theta_{t+1})\Big] - h_{\{i\}}\Big) \leq -(1-\pi_{t,1})\Delta_t^\pi + \frac{D_\eta}{\varepsilon}\sum_{2\leq i<j\leq K} h_{\{i,j\}}\pi_{t,1}(\pi_{t,i}+\pi_{t,j}).$$

This remainder bounds the variance terms involving two suboptimal arms together with the drift terms involving gap differences. For $K \geq 3$, the positive remainder cannot be bounded uniformly by a constant multiple of the failure term. Indeed, $h_{\{i,j\}}\pi_{t,1}(\pi_{t,i}+\pi_{t,j}) = \frac{\pi_{t,i}\pi_{t,j}(\pi_{t,i}+\pi_{t,j})}{\pi_{t,1}}$ can diverge as $\pi_{t,1} \downarrow 0$ while $\pi_{t,i}$ and $\pi_{t,j}$ stay positive, whereas $(1-\pi_{t,1})\Delta_t^\pi \leq \Delta_{\max}$. Thus the singleton estimate alone does not give the required negative drift.

Absorbing the cross terms with larger subsets. For a positive term indexed by A, $i \in A$, and $j \in A_I^c$, we put $D = A \cup \{j\}$. The softmax identity $\pi_{t,j} = \exp(\theta_{t,j} - \theta_{t,1})\pi_{t,1}$ gives

$$h_A\ \pi_{t,i}\pi_{t,j} = h_D\ \pi_{t,1}\pi_{t,i}.$$

This relation explains the role of Jia–Ouyang's subset potentials: The remainder of the sum of singleton potentials can be rewritten as a sum of level-two sets, $\sum_{2\leq i<j\leq K} h_{\{i,j\}}\pi_{t,1}(\pi_{t,i}+\pi_{t,j})$. Adding the size-two potentials supplies negative drift terms containing $h_{\{i,j\}}\pi_{t,1}(\pi_{t,i}+\pi_{t,j})$, with gap-dependent coefficients. Any cross terms involving a third suboptimal arm can in turn be absorbed by size-three potentials. We now make this relation between successive subset sizes explicit, continuing to divide each drift by $\eta\varepsilon$.

Exact relation. A positive term from a subset of size $r-1$ has the same form as a negative drift term from a subset D of size r. To see the latter, we split Δ_t^π into the terms in D and its complement in (2.12). The normalized drift of $C^{r-1}h_D$ then contains

$$-C^{r-1}(r+1)h_D\pi_{t,1}\sum_{i\in D}\Delta_i\pi_{t,i}.$$

The incoming positive terms have weight C^{r-2}, so the extra factor C in this negative term lets us absorb them. The process stops at $D = I$, whose drift has no cross term involving an outside suboptimal arm.

To count the positive terms arriving at each subset, we follow Jia–Ouyang's regrouping in their equation (76). For a fixed set D of size $r \geq 2$ and a fixed $i \in D$, each $j \in D_{\setminus\{i\}}$ gives exactly one preimage $A = D_{\setminus\{j\}}$. There are $r-1$ such choices. Therefore

$$\sum_{\varnothing \neq A \subseteq I} C^{|A|-1} h_A \sum_{\substack{i \in A \\ j \in A_I^c}} \pi_{t,i}\pi_{t,j} = \sum_{r=2}^{K-1} C^{r-2}(r-1) \sum_{\substack{D \subseteq I \\ |D|=r}} h_D \pi_{t,1} \sum_{i \in D} \pi_{t,i}. \tag{2.13}$$

After factoring out $C^{r-2} h_D \pi_{t,1}\pi_{t,i}$, the $r-1$ incoming terms have total coefficient $\frac{D_\eta}{\varepsilon}(r-1)$, while the negative term has coefficient $C(r+1)\Delta_i$. This is the comparison that will determine C.

We now use Lemma 2.3 with the same D_η bound on the interaction terms, multiply by $\frac{C^{|A|-1}h_A}{\varepsilon}$, and sum over nonempty A. In addition to the terms just matched, splitting Δ_t^π for $|A| \geq 2$ gives the nonpositive expression

$$-\sum_{\substack{A \subseteq I \\ |A| \geq 2}} C^{|A|-1}|A| h_A \pi_{t,1} \sum_{j \in A_I^c} \Delta_j \pi_{t,j} \leq 0. \tag{2.14}$$

We retain it in the combined bound. Together with the full singleton estimate, the preceding identities give, for every $C > 0$,

$$\begin{aligned}
\frac{\mathbb{E}_t[V_C(\theta_{t+1})] - V_C(\theta_t)}{\eta\varepsilon} &\leq -(1-\pi_{t,1})\Delta_t^\pi - \sum_{i=2}^{K} \Delta_i \pi_{t,i}^2 \\
&- \sum_{\substack{A \subseteq I \\ |A| \geq 2}} C^{|A|-1}|A| h_A \pi_{t,1} \sum_{j \in A_I^c} \Delta_j \pi_{t,j} - \frac{\eta \Gamma_K}{2\varepsilon} \sum_{\varnothing \neq A \subseteq I} C^{|A|-1}|A| h_A \pi_{t,1} \pi_{t,A_I^c} \\
&+ \sum_{r=2}^{K-1} C^{r-2} \sum_{\substack{D \subseteq I \\ |D|=r}} h_D \pi_{t,1} \times \sum_{i \in D} \pi_{t,i} \left(\frac{D_\eta}{\varepsilon}(r-1) - C(r+1)\Delta_i \right).
\end{aligned} \tag{2.15}$$

Remark 2.4 (Uniform control of subset interactions). Replacing the gap differences in Lemma 2.3 by the common bound D_η discards information about the arms inside and outside each subset. This step determines the geometric weights used below and can lead to a large initial potential for unequal gaps. Retaining the individual interactions may improve the bound for particular gap configurations.

Choosing the weights. We now choose C so that every bracket in the last sum is nonpositive, leaving the full singleton drift to control the failure term. For $C = C_\varepsilon := \frac{D_\eta}{\varepsilon \Delta_{\min}}$, the bracket for each $i \in D$ satisfies

$$\frac{D_\eta}{\varepsilon}(r-1) - C_\varepsilon(r+1)\Delta_i \leq \frac{D_\eta}{\varepsilon}(r-1) - \frac{D_\eta}{\varepsilon}(r+1) = -\frac{2D_\eta}{\varepsilon}.$$

All the other factors are nonnegative. To make explicit the terms we bound by zero, we collect their opposites in the nonnegative remainder

$$\begin{aligned}
\mathcal{R}_t = \sum_{i=2}^{K} \Delta_i \pi_{t,i}^2 &+ \sum_{\substack{A \subseteq I \\ |A| \geq 2}} C_\varepsilon^{|A|-1}|A| h_A \pi_{t,1} \Delta_t(A_I^c) + \frac{\eta \Gamma_K}{2\varepsilon} \sum_{\varnothing \neq A \subseteq I} C_\varepsilon^{|A|-1}|A| h_A \pi_{t,1} \pi_{t,A_I^c} \\
&+ \sum_{r=2}^{K-1} C_\varepsilon^{r-2} \sum_{\substack{D \subseteq I \\ |D|=r}} h_D \pi_{t,1} \sum_{i \in D} \pi_{t,i} \underbrace{\left(C_\varepsilon(r+1)\Delta_i - \frac{D_\eta}{\varepsilon}(r-1) \right)}_{\geq 2C_\varepsilon \Delta_{\min}} \geq 0.
\end{aligned}$$

These four sums are the singleton squares, the drift involving arms outside larger subsets, the retained variance slack, and the remaining drift of larger subsets after absorbing the incoming cross terms. Equation (2.15) therefore gives

$$\begin{aligned}\frac{\mathbb{E}_t[V_{C_\varepsilon}(\theta_{t+1})] - V_{C_\varepsilon}(\theta_t)}{\eta\varepsilon} &\le -(1-\pi_{t,1})\Delta_t^\pi - \mathcal{R}_t \\ &\le -(1-\pi_{t,1})\Delta_t^\pi. \end{aligned} \tag{2.16}$$

For $K = 2$, every sum over $r = 2, \dots, K-1$ is empty, and the same inequality follows directly from the singleton term. In this case, $\Delta_2 \pi_{t,2}^2 = (1-\pi_{t,1})\Delta_t^\pi$, so the first inequality retains a factor two in front of the failure term.

Remark 2.5 (Discarded negative drift). The last inequality in (2.16) replaces $-\mathcal{R}_t$ by zero. The choice C_ε uses $\Delta_i \ge \Delta_{\min}$ for these larger subsets and makes all their coefficients strictly negative, although nonpositivity would suffice for this last estimate. The failure bound contributes a horizon-independent term to regret, so for fixed instance parameters and learning rate it is lower order than $\log(T)$. Improving this term may still affect its dependence on K and the gaps. The effect of the discarded drift on these dependencies is not quantified here.

Summing the drift. We take expectations in (2.16) and sum from $t = 1$ to T. The tower property and Lemma 2.7 give

$$\mathbb{E}\left[\sum_{t=1}^{T} (1-\pi_{t,1})\,\Delta_t^\pi\right] \le \frac{V_{C_\varepsilon}(0) - \mathbb{E}[V_{C_\varepsilon}(\theta_{T+1})]}{\eta\varepsilon} \le \frac{V_{C_\varepsilon}(0)}{\eta\varepsilon}.$$

At $\theta = 0$, every h_A equals one. Counting subsets and applying the binomial formula gives

$$V_C(0) = \sum_{m=1}^{K-1} \binom{K-1}{m} C^{m-1} = \frac{(1+C)^{K-1} - 1}{C}. \tag{2.17}$$

We have therefore proved

$$\mathbb{E}\left[\sum_{t=1}^{T} (1-\pi_{t,1})\Delta_t^\pi\right] \le \frac{1}{\eta\varepsilon}\frac{(1+C_\varepsilon)^{K-1} - 1}{C_\varepsilon}. \tag{2.18}$$

□

2.1.1 Proof of the subset drift estimate

Proof of Lemma 2.3. Fix $t \ge 1$ and a nonempty $A \subseteq I$, and set $m = |A|$ and $h_A = h_A(\theta_t)$. As in part 3 of the theorem proof, we apply Lemma 2.6 with $q = K$ and $v = v_A$. The coordinates $v_{A,1} = -m$ and $v_{A,i} = \mathbf{1}\{i \in A\}$ for $i > 1$ give $\operatorname{osc} v_A = m + 1 \le K$. We expand the computations below for any fixed subset A, detailing what each term provides to the final bound.

Drift We first compute the drift term coordinate by coordinate. Writing $\odot$ for the Hadamard product and $\mathbf{1} \in \mathbb{R}^K$ for the vector of ones, we have

$$\begin{aligned}\mathsf{C}(\pi_t)\mu &= \operatorname{diag}(\pi_t)\mu - \pi_t(\pi_t^\top \mu) \\ &= \pi_t \odot \mu - \mu_t^\pi \pi_t = \pi_t \odot (\mu - \mu_t^\pi \mathbf{1}).\end{aligned}$$

Thus its ith coordinate is $\pi_{t,i}(\mu_i - \mu_t^\pi)$. Since $1 \notin A$, $v_{A,1} = -m$, $v_{A,i} = 1$ for $i \in A$, and all other coordinates of v_A vanish, the scalar product splits as

$$\begin{aligned} v_A^\top \mathsf{C}(\pi_t)\mu &= v_{A,1}\,\pi_{t,1}(\mu_1 - \mu_t^\pi) + \sum_{i\in A} v_{A,i}\,\pi_{t,i}(\mu_i - \mu_t^\pi) \\ &= -m\pi_{t,1}(\mu_1 - \mu_t^\pi) + \sum_{i\in A} \pi_{t,i}(\mu_i - \mu_t^\pi) \\ &= -m\pi_{t,1}\Delta_t^\pi + \sum_{i\in A} \pi_{t,i}(\Delta_t^\pi - \Delta_i). \end{aligned}$$

We expand these two terms separately. For coordinate 1, rewrite $\sum_{k=1}^K \pi_{t,k} = 1$ gives

$$-m\pi_{t,1}\Delta_t^\pi = -m\pi_{t,1}\sum_{i\in A}\pi_{t,i}\Delta_i - m\pi_{t,1}\sum_{j\in A_I^c}\pi_{t,j}\Delta_j.$$

For the coordinates in A, we split the inner sum into coordinate 1, the coordinates in A, and those in A_I^c:

$$\begin{aligned} \sum_{i\in A}\pi_{t,i}(\mu_i - \mu_t^\pi) &= \sum_{i\in A}\pi_{t,i}\sum_{k=1}^K \pi_{t,k}(\mu_i - \mu_k) \\ &= -\pi_{t,1}\sum_{i\in A}\pi_{t,i}\Delta_i + \sum_{i,k\in A}\pi_{t,i}\pi_{t,k}(\Delta_k - \Delta_i) + \sum_{\substack{i\in A\\ j\in A_I^c}}\pi_{t,i}\pi_{t,j}(\Delta_j - \Delta_i). \end{aligned}$$

The sum involving two indices in A cancels exactly. We now add the term for coordinate 1 to the terms indexed by A. Summing remaining terms yields

$$v_A^\top \mathsf{C}(\pi_t)\mu = -(m+1)\pi_{t,1}\sum_{i\in A}\pi_{t,i}\Delta_i - m\pi_{t,1}\sum_{j\in A_I^c}\pi_{t,j}\Delta_j + \sum_{\substack{i\in A\\ j\in A_I^c}}\pi_{t,i}\pi_{t,j}(\Delta_j - \Delta_i).$$

The interaction between A and A_I^c becomes

$$\begin{aligned} \sum_{\substack{i\in A\\ j\in A_I^c}}\pi_{t,i}\pi_{t,j}(\Delta_j - \Delta_i) &= \pi_{t,A}\Delta_t(A_I^c) - \pi_{t,A_I^c}\Delta_t(A) \\ &= \pi_{t,A}\pi_{t,A_I^c}\left(\overline{\Delta}_t(A_I^c) - \overline{\Delta}_t(A)\right). \end{aligned}$$

Thus the drift also reads

$$\begin{aligned} v_A^\top \mathsf{C}(\pi_t)\mu = &-(m+1)\pi_{t,1}\pi_{t,A}\overline{\Delta}_t(A) - m\pi_{t,1}\pi_{t,A_I^c}\overline{\Delta}_t(A_I^c) \\ &+ \pi_{t,A}\pi_{t,A_I^c}\left(\overline{\Delta}_t(A_I^c) - \overline{\Delta}_t(A)\right). \end{aligned}$$

Variance term We compute the quadratic term with the same coordinate expansion. For the vector v_A, we write $\bar{v}_{A,t} = \pi_t^\top v_A = -m\pi_{t,1} + \sum_{i\in A}\pi_{t,i}$. The Hadamard form is

$$\begin{aligned} \mathsf{C}(\pi_t)v_A &= \mathrm{diag}(\pi_t)v_A - \pi_t(\pi_t^\top v_A) \\ &= \pi_t \odot v_A - \bar{v}_{A,t}\pi_t = \pi_t \odot \left(v_A - \bar{v}_{A,t}\mathbf{1}\right). \end{aligned}$$

Using the values of v_A at coordinate 1 and on A, we obtain

$$v_A^\top \mathsf{C}(\pi_t)v_A = v_{A,1}\,\pi_{t,1}(v_{A,1} - \bar{v}_{A,t}) + \sum_{i\in A} v_{A,i}\,\pi_{t,i}(v_{A,i} - \bar{v}_{A,t}).$$

We expand the term for coordinate 1 using $\sum_{k=1}^{K} \pi_{t,k} = 1$:

$$-m\pi_{t,1} \sum_{k=1}^{K} \pi_{t,k}(v_{A,1} - v_{A,k}) = m(m+1)\pi_{t,1} \sum_{i \in A} \pi_{t,i} + m^2 \pi_{t,1} \sum_{j \in A_I^c} \pi_{t,j}.$$

Here $v_{A,1} - v_{A,i} = -(m+1)$ for $i \in A$ and $v_{A,1} - v_{A,j} = -m$ for $j \in A_I^c$. For the terms indexed by A, we again split the inner sum into coordinate 1, the coordinates in A, and those in A_I^c:

$$\begin{aligned} \sum_{i \in A} \pi_{t,i}(v_{A,i} - \bar{v}_{A,t}) &= \sum_{i \in A} \pi_{t,i} \sum_{k=1}^{K} \pi_{t,k}(v_{A,i} - v_{A,k}) \\ &= (m+1)\pi_{t,1} \sum_{i \in A} \pi_{t,i} + \sum_{\substack{i \in A \\ j \in A_I^c}} \pi_{t,i}\pi_{t,j}, \end{aligned}$$

where products involving $i, j \in A$ were removed because they give zero. The sum with both indices in A vanishes because every such coordinate of v_A equals 1. We add these two expressions. The coefficient of $\pi_{t,1} \sum_{i \in A} \pi_{t,i}$ is $m(m+1) + (m+1) = (m+1)^2$. We therefore obtain

$$v_A^\top \mathsf{C}(\pi_t) v_A = (m+1)^2 \pi_{t,1} \sum_{i \in A} \pi_{t,i} + m^2 \pi_{t,1} \sum_{j \in A_I^c} \pi_{t,j} + \sum_{\substack{i \in A \\ j \in A_I^c}} \pi_{t,i}\pi_{t,j}.$$

Resulting bound. Substituting these expressions into the one-step bound, we obtain

$$\begin{aligned} \frac{\mathbb{E}_t[h_A(\theta_{t+1})] - h_A}{\eta h_A} \le &-(m+1)\pi_{t,1} \sum_{i \in A} \pi_{t,i}\Delta_i + \frac{\eta \Gamma_K}{2}(m+1)^2 \pi_{t,1} \sum_{i \in A} \pi_{t,i} \\ &- m\pi_{t,1} \sum_{j \in A_I^c} \pi_{t,j}\Delta_j + \frac{\eta \Gamma_K}{2} m^2 \pi_{t,1} \sum_{j \in A_I^c} \pi_{t,j} \\ &+ \sum_{\substack{i \in A \\ j \in A_I^c}} \pi_{t,i}\pi_{t,j}(\Delta_j - \Delta_i) + \frac{\eta \Gamma_K}{2} \sum_{\substack{i \in A \\ j \in A_I^c}} \pi_{t,i}\pi_{t,j}. \end{aligned}$$

We group the two terms for each type of pair. In the first line, we factor $-(m+1)\pi_{t,1}\pi_{t,i}$ for each $i \in A$. In the second line, we factor $-m\pi_{t,1}\pi_{t,j}$ for each $j \in A_I^c$. In the last line, we factor $\pi_{t,i}\pi_{t,j}$. This gives

$$\begin{aligned} \frac{\mathbb{E}_t[h_A(\theta_{t+1})] - h_A}{\eta h_A} \le &-(m+1)\pi_{t,1} \sum_{i \in A} \pi_{t,i} \left(\Delta_i - \frac{\eta(m+1)\Gamma_K}{2} \right) \\ &- m\pi_{t,1} \sum_{j \in A_I^c} \pi_{t,j} \left(\Delta_j - \frac{\eta m \Gamma_K}{2} \right) \\ &+ \sum_{\substack{i \in A \\ j \in A_I^c}} \pi_{t,i}\pi_{t,j} \left(\Delta_j - \Delta_i + \frac{\eta \Gamma_K}{2} \right). \end{aligned} \tag{2.19}$$

To retain the individual gaps, we use $m+1 \le K$ and (2.1). For the first two brackets, these give

$$\begin{aligned} \Delta_i - \frac{\eta(m+1)\Gamma_K}{2} &\ge \Delta_i - (1-\varepsilon)\Delta_{\min} \ge \varepsilon \Delta_i, && i \in A, \\ \Delta_j - \frac{\eta m \Gamma_K}{2} &= \Delta_j - \frac{\eta(m+1)\Gamma_K}{2} + \frac{\eta \Gamma_K}{2} \ge \varepsilon \Delta_j + \frac{\eta \Gamma_K}{2}, && j \in A_I^c. \end{aligned}$$

The second margin retains an additional $\frac{\eta\Gamma_K}{2}$. Before bounding the interaction between the subsets, we therefore have

$$\frac{\mathbb{E}_t[h_A(\theta_{t+1})] - h_A}{\eta h_A} \le -\varepsilon\pi_{t,1}\left(m\Delta_t^\pi + \Delta_t(A)\right) + \pi_{t,A}\pi_{t,A_I^c}\left(\overline{\Delta}_t(A_I^c) - \overline{\Delta}_t(A) + \frac{\eta\Gamma_K}{2}\right) - \frac{\eta m\Gamma_K}{2}\pi_{t,1}\pi_{t,A_I^c}.$$

□

2.2 Supporting results

The main proof uses one estimate for the drift of exponential potentials. We derive it from Taylor's formula, then record the covariance identity used for the subset potentials and verify integrability at every finite round.

Lemma 2.6 (Exponential one-step approximation)**.** *For every real x,*

$$\exp(x) = 1 + x + x^2\int_0^1 (1-u)\exp(ux)\,\mathrm{d}u \le 1 + x + \frac{x^2}{2}\exp(|x|). \tag{2.20}$$

Under (1.2) and (1.4), we write $\mathsf{C}(\pi_t) = \mathsf{C}(\pi_t) = \mathrm{diag}(\pi_t) - \pi_t\pi_t^\top$. For a deterministic $v \in \mathbb{R}^K$, we define $f_v(\theta) = \exp(v^\top\theta)$, $\bar{v}_t = \sum_i \pi_{t,i}v_i$, and $\mathrm{osc}(v) = \max_i v_i - \min_i v_i$. If $\mathrm{osc}(v) \le q$ for $q \in \{1, K\}$, then

$$\mathbb{E}_t[f_v(\theta_{t+1})] - f_v(\theta_t) \le f_v(\theta_t)\left(\eta v^\top\mathsf{C}(\pi_t)\mu + \frac{\eta^2\Gamma_q}{2}v^\top\mathsf{C}(\pi_t)v\right). \tag{2.21}$$

This is the approximation of Jia and Ouyang (2026, equation (73)).

Proof. The equality in (2.20) is Taylor's formula with integral remainder. Bounding $\exp(ux)$ by $\exp(|x|)$ for $u \in [0,1]$ gives the inequality.

For the one-step estimate, the update gives $\frac{f_v(\theta_{t+1})}{f_v(\theta_t)} = \exp(\eta X_t(v_{A_t} - \bar{v}_t))$. Since $\bar{v}_t$ is a convex combination of the entries of v, $|v_i - \bar{v}_t| \le \mathrm{osc}(v) \le q$. Applying (2.20) to this exponential yields

$$\frac{f_v(\theta_{t+1}) - f_v(\theta_t)}{f_v(\theta_t)} \le \eta X_t(v_{A_t} - \bar{v}_t) + \frac{\eta^2}{2}X_t^2(v_{A_t} - \bar{v}_t)^2\exp(\eta|X_t||v_{A_t} - \bar{v}_t|). \tag{2.22}$$

The terms on the right are integrable by (1.4), so we can take conditional expectations. The baseline cancels from the first-order term because $\sum_i \pi_{t,i}(v_i - \bar{v}_t) = 0$:

$$\begin{aligned}\mathbb{E}_t[X_t(v_{A_t} - \bar{v}_t)] &= \sum_i \pi_{t,i}(\mu_i - b_t)(v_i - \bar{v}_t)\\ &= \sum_i \pi_{t,i}\mu_i v_i - \left(\sum_i \pi_{t,i}\mu_i\right)\left(\sum_i \pi_{t,i}v_i\right) = v^\top\mathsf{C}(\pi_t)\mu.\end{aligned}$$

For the second-order term, conditioning on the sampled arm gives

$$\begin{aligned}&\mathbb{E}_t\left[X_t^2(v_{A_t} - \bar{v}_t)^2\exp(\eta|X_t||v_{A_t} - \bar{v}_t|)\right]\\ &\quad\le \Gamma_q\sum_i \pi_{t,i}(v_i - \bar{v}_t)^2 = \Gamma_q v^\top\mathsf{C}(\pi_t)v.\end{aligned}$$

Multiplying by the positive, $\mathcal{F}_t$-measurable factor $f_v(\theta_t)$ gives (2.21). □

For the subset calculations, we also use the following pairwise form of the covariance. It makes the terms involving the optimal arm and the terms between suboptimal arms explicit. For any $v, w \in \mathbb{R}^K$, expanding the double sum yields

$$\begin{aligned} v^\top \mathsf{C}(\pi_t) w &= \frac{1}{2} \sum_{a,b} \pi_{t,a}\pi_{t,b}(v_a - v_b)(w_a - w_b) \\ &= \sum_{a<b} \pi_{t,a}\pi_{t,b}(v_a - v_b)(w_a - w_b). \end{aligned} \tag{2.23}$$

Indeed, the first expression expands to $\sum_a \pi_{t,a} v_a w_a - (\sum_a \pi_{t,a} v_a)(\sum_b \pi_{t,b} w_b)$.

Lemma 2.7 (Finite-round integrability)**.** *Under (1.2) and (1.4), every logit $\theta_{t,i}$ is integrable at each finite round $t \geq 1$. For every deterministic $v \in \mathbb{R}^K$ with $\operatorname{osc}(v) \leq K$, the exponential $f_v(\theta_t) = \exp(v^\top \theta_t)$ is also integrable.*

Proof. Conditioning on A_t in (1.4) gives $\mathbb{E}_t[X_t^2] \leq \Gamma_K$ and hence $\mathbb{E}_t[|X_t|] \leq \sqrt{\Gamma_K}$. Applying (2.20) to $x = \eta K |X_t|$, we obtain

$$\mathbb{E}_t[\exp(\eta K |X_t|)] \leq 1 + \eta K \sqrt{\Gamma_K} + \frac{\eta^2 K^2 \Gamma_K}{2} < \infty.$$

If $\operatorname{osc}(v) \leq K$, the update also gives

$$f_v(\theta_{t+1}) = f_v(\theta_t) \exp\big(\eta X_t (v_{A_t} - \bar{v}_t)\big) \leq f_v(\theta_t) \exp(\eta K |X_t|).$$

Since $f_v(\theta_t)$ is $\mathcal{F}_t$-measurable, the preceding conditional bound controls $\mathbb{E}_t[f_v(\theta_{t+1})]$ by a fixed deterministic multiple of $f_v(\theta_t)$. Starting from $f_v(\theta_1) = 1$, induction and the tower property give $\mathbb{E}[f_v(\theta_t)] < \infty$ at every finite round.

Finally, $|\theta_{t+1,i} - \theta_{t,i}| \leq \eta |X_t|$ and $\theta_1 = 0$ imply

$$\mathbb{E}[|\theta_{T+1,i}|] \leq \eta \sum_{t=1}^{T} \mathbb{E}[|X_t|] \leq T\eta \sqrt{\Gamma_K} < \infty.$$

□

Refining the failure bound. The gap spread retained in D_η reduces the weights of Jia–Ouyang's subset potentials. For equal gaps $\Delta_i = \Delta$, the differences $\Delta_j - \Delta_i$ cancel, so the cross-term coefficient is only $\frac{\eta \Gamma_K(\eta)}{2}$. If $\eta K \Gamma_K(\eta) \leq \Delta$, the theorem with $\varepsilon = \frac{1}{2}$ has $C_\varepsilon = \frac{\eta \Gamma_K(\eta)}{\Delta} \leq \frac{1}{K}$ and gives

$$\mathbb{E}\left[\sum_{t=1}^{T} (1 - \pi_{t,1}) \Delta_t^\pi\right] \leq \frac{2}{\eta} \frac{(1 + C_\varepsilon)^{K-1} - 1}{C_\varepsilon} \leq \frac{2 \exp(1)(K-1)}{\eta}.$$

Retaining (2.14) gives additional negative drift, whose effect on the failure bound is not quantified here.

3 Sharper bounds for the post-convergence term

The bound in Section 2 has logarithmic coefficient $\frac{K-1}{\eta}$. We improve this coefficient by keeping the relation between the optimal and suboptimal logits. We use the regret decomposition of Baudry

et al. (2025, equation (7)), recalled in (2.4). Write $B_T = \mathbb{E}\left[\sum_{t=1}^{T}(1-\pi_{t,1})\Delta_t^{\pi}\right]$ for the failure term and $P_T = \mathbb{E}\left[\sum_{t=1}^{T}\pi_{t,1}\Delta_t^{\pi}\right]$ for the post-convergence term. Thus

$$R_T = P_T + B_T, \qquad P_T = \frac{\mathbb{E}[\theta_{T+1,1}]}{\eta}. \tag{3.1}$$

Throughout this section, we work under the assumptions of Theorem 2.1 and fix any finite deterministic bound B on $\sup_{T\geq 1} B_T$. The failure analysis provides the choice $B = \frac{V_{C_\varepsilon}(0)}{\eta\varepsilon}$, with the explicit initial value in (2.17). We write $\Gamma_q = \Gamma_q(\eta)$, $\overline{\Gamma}_1 = \min\{\Gamma_1(\eta), \Gamma_K(\eta)\}$, and $\mathbb{E}_t[W] = \mathbb{E}[W \mid \mathcal{F}_t]$. Both moment bounds in (1.4) control the moment at $q = 1$, so $\overline{\Gamma}_1$ also bounds the conditional second moments of X_t. Lemma 2.7 ensures the finite-round logit and exponential integrability used below.

We introduce a power parameter in the gap-weighted exponential sum. Reducing this parameter makes the variance correction in its drift arbitrarily small. This gives the coefficient $\frac{K-1}{K\eta}$ with a sublogarithmic remainder, while keeping the failure term explicit. A matching lower bound identifies the exact expected asymptotic coefficient.

3.1 A bound from gap-weighted exponentials

The argument combines the negative drift of a function of the suboptimal logits with a bound on the expected growth of an exponential ratio. The power parameter α controls the curvature of the first function. It is used only in the analysis; the algorithm uses the same fixed learning rate η throughout.

Lemma 3.1 (Post-convergence regret with the exact asymptotic coefficient)**.** *Under the assumptions of Theorem 2.1, let B be a finite deterministic bound on $\sup_{T\geq 1} B_T$. For each fixed instance and fixed learning rate, every sufficiently large integer T satisfies*

$$P_T \leq \frac{K-1}{K\eta}\log\left(\Lambda_T\right) + \frac{2}{K\eta}\sqrt{\log(K-1)\log\left(\Lambda_T\right)} + B,, \tag{3.2}$$

with $\Lambda_T = (K-1)(1+4\eta\Delta_{\max}T)$. Moreover,

$$\lim_{T\to\infty}\frac{P_T}{\log(T)} = \lim_{T\to\infty}\frac{R_T}{\log(T)} = \frac{K-1}{K\eta}. \tag{3.3}$$

The first term has logarithmic coefficient $\frac{K-1}{K\eta}$. The second grows as $\sqrt{\log(T)}$ for fixed K, and the failure term enters with coefficient one. Adding B gives the corresponding upper bound on R_T. The proof first keeps the power parameter free, then optimizes it to obtain (3.2).

Proof. Fix $\alpha \in (0, 1]$, and define

$$W_\alpha(\theta) = \frac{1}{\alpha}\log\left(\sum_{i=2}^{K}(\Delta_i\exp(\theta_i))^\alpha\right), \qquad Y_\alpha(\theta) = \exp(\theta_1 - W_\alpha(\theta)).$$

Thus $\theta_1 = \log(Y_\alpha(\theta)) + W_\alpha(\theta)$. We will show that the expected increments of Y_α are bounded and that the drift of W_α contains a negative multiple of regret. Jensen's inequality will then bound the expected optimal logit through this identity.

Negative drift of the logarithm of the power sum. At round t, write $z_{t,i} = \Delta_i \pi_{t,i}$ and $w_{t,i} = \frac{z_{t,i}^\alpha}{\sum_{j=2}^K z_{t,j}^\alpha}$ for $i > 1$. These weights sum to one. The inequality $\log(x) \le x - 1$ gives

$$\begin{aligned} W_\alpha(\theta_{t+1}) - W_\alpha(\theta_t) &= \frac{1}{\alpha} \log\left(\sum_{i=2}^K w_{t,i} \exp\left(\alpha(\theta_{t+1,i} - \theta_{t,i})\right)\right) \\ &\le \frac{1}{\alpha} \sum_{i=2}^K w_{t,i} \left(\exp\left(\alpha(\theta_{t+1,i} - \theta_{t,i})\right) - 1\right). \end{aligned}$$

We then apply Lemma 2.6 with $v = \alpha e_i$, and use $\mathbb{E}_t[\theta_{t+1,i} - \theta_{t,i}] = \eta \pi_{t,i}(\Delta_t^\pi - \Delta_i)$. We obtain

$$\begin{aligned} \mathbb{E}_t[W_\alpha(\theta_{t+1})] - W_\alpha(\theta_t) &\le \eta \Delta_t^\pi \sum_{i=2}^K w_{t,i}\pi_{t,i} - \eta \sum_{i=2}^K w_{t,i}\Delta_i \pi_{t,i} + \frac{\alpha\eta^2 \overline{\Gamma}_1}{2} \sum_{i=2}^K w_{t,i}\pi_{t,i}(1 - \pi_{t,i}) \\ &\le \eta(1 - \pi_{t,1})\Delta_t^\pi - \eta\left(1 - \frac{\alpha\eta\overline{\Gamma}_1}{2\Delta_{\min}}\right) \sum_{i=2}^K w_{t,i} z_{t,i}. \end{aligned}$$

The last line uses $\pi_{t,i} \le 1 - \pi_{t,1}$ and $\Delta_i \ge \Delta_{\min}$. The variance correction has a factor α, because the exponential remainder is quadratic in α and the definition of W_α divides by α.

To bound the remaining weighted sum, observe that

$$\begin{aligned} &(K-1)\sum_{i=2}^K z_{t,i}^{\alpha+1} - \left(\sum_{i=2}^K z_{t,i}\right)\left(\sum_{i=2}^K z_{t,i}^\alpha\right) \\ &\quad = \sum_{2 \le i < j \le K} (z_{t,i} - z_{t,j})(z_{t,i}^\alpha - z_{t,j}^\alpha) \ge 0. \end{aligned}$$

Every summand is nonnegative since $x \mapsto x^\alpha$ is increasing. Dividing by $(K-1)\sum_{i=2}^K z_{t,i}^\alpha$ gives

$$\sum_{i=2}^K w_{t,i} z_{t,i} \ge \frac{\Delta_t^\pi}{K-1}.$$

To express this drift bound, define $\kappa_{\eta,\alpha} = \frac{K-1}{K - \frac{\alpha\eta\overline{\Gamma}_1}{2\Delta_{\min}}}$. The learning-rate condition makes $1 - \frac{\alpha\eta\overline{\Gamma}_1}{2\Delta_{\min}} > 0$. Consequently,

$$\mathbb{E}_t[W_\alpha(\theta_{t+1})] - W_\alpha(\theta_t) \le \eta(1 - \pi_{t,1})\Delta_t^\pi - \eta\frac{1 - \kappa_{\eta,\alpha}}{\kappa_{\eta,\alpha}}\Delta_t^\pi.$$

Taking expectations and summing yields

$$\mathbb{E}[W_\alpha(\theta_{T+1})] \le W_\alpha(0) + \eta B_T - \eta\frac{1 - \kappa_{\eta,\alpha}}{\kappa_{\eta,\alpha}} R_T. \tag{3.4}$$

Expected growth of the exponential ratio. We next bound the expected increments of Y_α. Convexity of $x \mapsto x^{-\frac{1}{\alpha}}$ gives the pathwise inequality

$$\begin{aligned} \frac{Y_\alpha(\theta_{t+1})}{Y_\alpha(\theta_t)} &= \left(\sum_{i=2}^K w_{t,i} \exp\left(\alpha\big((\theta_{t+1,i} - \theta_{t+1,1}) - (\theta_{t,i} - \theta_{t,1})\big)\right)\right)^{-\frac{1}{\alpha}} \\ &\le \sum_{i=2}^K w_{t,i} \exp\big((\theta_{t+1,1} - \theta_{t+1,i}) - (\theta_{t,1} - \theta_{t,i})\big). \end{aligned}$$

For $0 < \alpha \le 1$, we have $\left(\sum_{i=2}^K z_{t,i}^\alpha\right)^{\frac{1}{\alpha}} \ge \sum_{i=2}^K z_{t,i}$. Hence cancellation of the softmax normalizing factor gives

$$Y_\alpha(\theta_t) = \frac{\pi_{t,1}}{\left(\sum_{i=2}^K z_{t,i}^\alpha\right)^{\frac{1}{\alpha}}} \le \frac{\pi_{t,1}}{\Delta_t^\pi}.$$

For each $i > 1$, the vector $e_1 - e_i$ has oscillation two, hence at most K. Its drift factor in Lemma 2.6 is $\pi_{t,1}\Delta_t^\pi - \pi_{t,i}(\Delta_t^\pi - \Delta_i)$. After dropping $-\pi_{t,i}\Delta_t^\pi$, the weighted drift factor is bounded by

$$\begin{aligned} &Y_\alpha(\theta_t)\sum_{i=2}^K w_{t,i}(\pi_{t,1}\Delta_t^\pi - \pi_{t,i}(\Delta_t^\pi - \Delta_i)) \\ &\le \pi_{t,1}^2 + \pi_{t,1}\frac{\sum_{i=2}^K w_{t,i}z_{t,i}}{\Delta_t^\pi} \le 2. \end{aligned}$$

The last inequality uses $z_{t,i} \le \Delta_t^\pi$. For the variance factor,

$$\begin{aligned} (e_1 - e_i)^\top \mathsf{C}(\pi_t)(e_1 - e_i) &= \pi_{t,1} + \pi_{t,i} - (\pi_{t,1} - \pi_{t,i})^2 \\ &= \pi_{t,1}(1 - \pi_{t,1}) + (1 + 2\pi_{t,1})\pi_{t,i} - \pi_{t,i}^2 \le 4(1 - \pi_{t,1}). \end{aligned}$$

Also, $Y_\alpha(\theta_t)(1 - \pi_{t,1}) \le \frac{\pi_{t,1}(1-\pi_{t,1})}{\Delta_t^\pi} \le \frac{1}{\Delta_{\min}}$. Lemma 2.6 therefore gives

$$\mathbb{E}_t[Y_\alpha(\theta_{t+1})] - Y_\alpha(\theta_t) \le 2\eta + \frac{2\eta^2\Gamma_K}{\Delta_{\min}}.$$

These variables are integrable: $Y_\alpha(\theta_t)$ is at most $\frac{1}{\Delta_2}\exp(\theta_{t,1} - \theta_{t,2})$, and both $W_\alpha(\theta_t)$ and $\log(Y_\alpha(\theta_t))$ grow at most linearly in the absolute logits for fixed α. To express the growth relative to the initial value $Y_\alpha(0) = \exp(-W_\alpha(0))$, define

$$L_{\eta,\alpha} = \left(\sum_{i=2}^K \Delta_i^\alpha\right)^{\frac{1}{\alpha}} \left(2\eta + \frac{2\eta^2\Gamma_K}{\Delta_{\min}}\right).$$

Summing the expected increments gives

$$\mathbb{E}[Y_\alpha(\theta_{T+1})] \le \exp(-W_\alpha(0))(1 + L_{\eta,\alpha}T).$$

Combining the estimates. Jensen's inequality and (3.4) imply

$$\begin{aligned} \mathbb{E}[\theta_{T+1,1}] &= \mathbb{E}[\log(Y_\alpha(\theta_{T+1})) + W_\alpha(\theta_{T+1})] \\ &\le \log(1 + L_{\eta,\alpha}T) + \eta B_T - \eta\frac{1 - \kappa_{\eta,\alpha}}{\kappa_{\eta,\alpha}}R_T. \end{aligned}$$

Using (3.1) and rearranging gives

$$P_T \le \frac{\kappa_{\eta,\alpha}}{\eta}\log(1 + L_{\eta,\alpha}T) + (2\kappa_{\eta,\alpha} - 1)B_T. \tag{3.5}$$

The moment bound $\overline{\Gamma}_1$ is strictly positive, since otherwise $X_t = 0$ conditional on every arm and all reward means would coincide. The learning-rate condition gives

$$0 < \frac{\alpha\eta\overline{\Gamma}_1}{2\Delta_{\min}} \le \frac{\alpha(1 - \varepsilon)}{K} < \frac{1}{K}.$$

Thus $\frac{K-1}{K} < \kappa_{\eta,\alpha} < \frac{K}{K+1}$ and $2\kappa_{\eta,\alpha} - 1 > 0$, so we may replace B_T by B.

Optimizing the bound. We now bound the terms that depend on α and choose its value. Since $2\kappa_{\eta,\alpha} - 1 < 1$ and $B_T \le B$, we first bound the failure remainder by B. The learning-rate condition also gives

$$2\eta + \frac{2\eta^2\Gamma_K}{\Delta_{\min}} \le 2\eta\left(1 + \frac{2(1-\varepsilon)}{K}\right) \le 4\eta.$$

Using $\left(\sum_{i=2}^{K}\Delta_i^{\alpha}\right)^{\frac{1}{\alpha}} \le \Delta_{\max}(K-1)^{\frac{1}{\alpha}}$ and $K - 1 \ge 1$, we obtain

$$\log(1 + L_{\eta,\alpha}T) \le \log(1 + 4\eta\Delta_{\max}T) + \frac{\log(K-1)}{\alpha}.$$

For the optimization, write $u_T = \log(1 + 4\eta\Delta_{\max}T)$ and $r_\eta = \frac{\eta\overline{\Gamma}_1}{2K\Delta_{\min}}$. Then $0 < r_\eta < \frac{1}{K^2}$, and (3.5) implies

$$P_T \le \frac{K-1}{K\eta}\cdot\frac{1}{1-r_\eta\alpha}\left(u_T + \frac{\log(K-1)}{\alpha}\right) + B. \tag{3.6}$$

For small α, the excess above $\frac{K-1}{K\eta}u_T$ has contributions proportional to αu_T and $\frac{\log(K-1)}{\alpha}$. We optimize the fraction in (3.6) directly to keep these constants explicit.

Suppose first that $K \ge 3$. Differentiating the fraction gives

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\frac{u_T + \frac{\log(K-1)}{\alpha}}{1-r_\eta\alpha}\right) = \frac{r_\eta u_T + \frac{2r_\eta\log(K-1)}{\alpha} - \frac{\log(K-1)}{\alpha^2}}{(1-r_\eta\alpha)^2}.$$

Its numerator has a unique positive zero. The minimizing power is

$$\alpha_T = \frac{\sqrt{\frac{\log(K-1)}{r_\eta}}}{\sqrt{u_T + r_\eta\log(K-1)} + \sqrt{r_\eta\log(K-1)}}.$$

Since $u_T \to \infty$, this power belongs to $(0,1]$ for every sufficiently large T. It is of order $\sqrt{\frac{\log(K-1)}{r_\eta u_T}}$, as suggested by balancing the two contributions above. The zero-derivative condition gives $r_\eta u_T\alpha_T^2 + 2r_\eta\log(K-1)\alpha_T - \log(K-1) = 0$. Consequently,

$$\begin{aligned}
\frac{u_T + \frac{\log(K-1)}{\alpha_T}}{1-r_\eta\alpha_T} &= \frac{\log(K-1)}{r_\eta\alpha_T^2} \\
&= \left(\sqrt{u_T + r_\eta\log(K-1)} + \sqrt{r_\eta\log(K-1)}\right)^2 \\
&= u_T + 2r_\eta\log(K-1) + 2\sqrt{r_\eta\log(K-1)\big(u_T + r_\eta\log(K-1)\big)} \\
&\le u_T + \log(K-1) + 2\sqrt{r_\eta\log(K-1)\big(u_T + \log(K-1)\big)}.
\end{aligned}$$

The last line uses $r_\eta < \frac{1}{K^2} \le \frac{1}{2}$. Thus the optimization gives a square-root correction and an additive term $\log(K-1)$, which we include in the logarithm of the leading term. Finally,

$$2\frac{K-1}{K\eta}\sqrt{r_\eta} \le \frac{2(K-1)}{K^2\eta} \le \frac{2}{K\eta}.$$

Combining these inequalities with $u_T + \log(K-1) = \log\left((K-1)(1+4\eta\Delta_{\max}T)\right)$ proves (3.2).

If $K = 2$, then $\log(K-1) = 0$ and $\kappa_{\eta,\alpha} \to \frac{1}{2}$ as $\alpha \downarrow 0$. Letting $\alpha \downarrow 0$ in (3.6) gives (3.2), with its square-root term equal to zero. For fixed instance and learning rate, $u_T + \log(K-1) = \log(T) + O(1)$, so the bound proves the upper limits in (3.3). The bounded difference $R_T - P_T = B_T$ gives the same upper limit for R_T. The matching lower bound is proved in Section 3.1.1. □

The free-parameter estimate (3.5) retains the individual gaps through $L_{\eta,\alpha}$ and may give smaller finite-time constants. When $K = 2$, the sum defining W_α has one term and $L_{\eta,\alpha}$ is independent of α. Letting $\alpha \downarrow 0$ directly in (3.5) then removes the failure term and gives $P_T \le \frac{1}{2\eta}\log(1 + 4\eta\Delta_2 T)$.

3.1.1 Proof of the matching asymptotic lower bound

We complete (3.3) by showing that each suboptimal arm has scaled cumulative allocation at least $\frac{1}{K}\log(T)$ asymptotically. The bounded failure term and the martingale fluctuations are negligible on this scale. We compare the allocations through their common softmax normalizing factor, then apply Fatou's lemma to their nonnegative sum.

Proof of the lower bound in Lemma 3.1. For $i > 1$, let $a_i(t) = \eta\Delta_i \sum_{s<t} \pi_{s,i}$, and let $G_t = \sum_{s<t}(1-\pi_{s,1})\Delta_s^\pi$ be the accumulated failure term. For each coordinate, let $M_i(1) = 0$ and

$$M_i(t) = \sum_{s<t} \left(\theta_{s+1,i} - \theta_{s,i} - \mathbb{E}_s[\theta_{s+1,i} - \theta_{s,i}]\right).$$

These are square-integrable martingales for the pre-round filtration $(\mathcal{F}_t)_{t\ge1}$. The conditional coordinate drifts give

$$\begin{aligned} \theta_{t,i} &= -a_i(t) + \eta \sum_{s<t} \pi_{s,i}\Delta_s^\pi + M_i(t), && i > 1, \\ \theta_{t,1} &= \sum_{i=2}^K a_i(t) - \eta G_t + M_1(t). \end{aligned} \tag{3.7}$$

Fix T and put $N_T = \max_{1\le i\le K}\max_{1\le t\le T+1}|M_i(t)|$ and $\ell_T = \eta G_{T+1} + N_T$. Since $\pi_{s,i}\Delta_s^\pi \le (1-\pi_{s,1})\Delta_s^\pi$ for $i > 1$, we have

$$|\theta_{t,i} + a_i(t)| \le \ell_T \quad (i > 1), \qquad \left|\theta_{t,1} - \sum_{i=2}^K a_i(t)\right| \le \ell_T, \quad t \le T+1. \tag{3.8}$$

Controlling the martingale fluctuations. The conditional variance of a logit increment is at most its conditional second moment. Conditioning also on A_t therefore gives

$$\mathbb{E}_t\Big[(M_i(t+1) - M_i(t))^2\Big] \le \eta^2\overline{\Gamma}_1 \sum_{j=1}^K \pi_{t,j}(\mathbf{1}\{j=i\} - \pi_{t,i})^2 = \eta^2\overline{\Gamma}_1\pi_{t,i}(1-\pi_{t,i}).$$

Doob's L^2 inequality, applied to each coordinate, implies

$$\mathbb{E}\Big[N_T^2\Big] \le 4\sum_{i=1}^K \mathbb{E}\Big[M_i(T+1)^2\Big] \le 4\eta^2\overline{\Gamma}_1 \sum_{t=1}^T \mathbb{E}\left[1 - \sum_{i=1}^K \pi_{t,i}^2\right] \le \frac{8\eta^2\overline{\Gamma}_1}{\Delta_{\min}} R_T. \tag{3.9}$$

The last inequality uses $1 - \sum_i \pi_{t,i}^2 \le 2(1-\pi_{t,1})$ and $\Delta_t^\pi \ge \Delta_{\min}(1-\pi_{t,1})$.

Comparing the cumulative allocations. By (3.8), the factor $\exp(\theta_{t,i})$ is within a factor $\exp(\ell_T)$ of $\exp(-a_i(t))$. Multiplying the allocation increment by $\exp(a_i(t))$ therefore leaves a factor $\exp(a_i(t) + \theta_{t,i})$ between $\exp(-\ell_T)$ and $\exp(\ell_T)$. We can then compare all suboptimal allocations

through $S_t = \eta \sum_{s<t} \frac{1}{Z_s}$, where $Z_s = \sum_{j=1}^{K} \exp(\theta_{s,j})$. The comparison is pathwise, so ℓ_T may depend on the whole trajectory up to $T+1$. The allocation increment satisfies

$$a_i(t+1) - a_i(t) = \frac{\eta \Delta_i \exp(\theta_{t,i})}{Z_t} \in [0, \eta\Delta_i].$$

Apply $x \le \exp(x) - 1 \le x \exp(x)$ for $x \ge 0$ to this increment and use (3.8). For $t \le T$,

$$\frac{\eta \Delta_i \exp(-\ell_T)}{Z_t} \le \exp(a_i(t+1)) - \exp(a_i(t)) \le \frac{\eta \Delta_i \exp(\ell_T + \eta\Delta_{\max})}{Z_t}.$$

Summing from round 1 to round $t-1$, with $a_i(1) = S_1 = 0$, gives

$$1 + \Delta_i \exp(-\ell_T) S_t \le \exp(a_i(t)) \le 1 + \Delta_i \exp(\ell_T + \eta\Delta_{\max}) S_t, \qquad t \le T+1. \tag{3.10}$$

Errors on the logarithmic scale. The sequence G_t is nondecreasing. Write $G_\infty = \lim_{t\to\infty} G_t$. Monotone convergence gives $\mathbb{E}[G_\infty] \le B$, so $G_\infty < \infty$ almost surely. The bound (3.5) with $\alpha = 1$, together with $R_T \le P_T + B$, implies $R_T = O(\log(T))$ for $T \ge 2$. By (3.9), there is a finite deterministic constant C such that $\mathbb{E}[N_T^2] \le C \log(T)$ for all $T \ge 2$. For $T_n = \lceil \exp(n^2) \rceil$ and any $u > 0$, Markov's inequality gives

$$\sum_{n=1}^{\infty} \mathbb{P}\{N_{T_n} > u \log(T_n)\} \le \frac{C}{u^2} \sum_{n=1}^{\infty} \frac{1}{\log(T_n)} < \infty.$$

Apply Borel–Cantelli for positive rational u. It follows that $\frac{N_{T_n}}{\log(T_n)} \to 0$ almost surely. For $T_n \le T \le T_{n+1}$, monotonicity gives $\frac{N_T}{\log(T)} \le \frac{N_{T_{n+1}}}{\log(T_n)}$. Since $\frac{\log(T_{n+1})}{\log(T_n)} \to 1$, we conclude that

$$\frac{\ell_T}{\log(T)} \longrightarrow 0 \quad \text{almost surely.} \tag{3.11}$$

A lower bound on the common sum. Fix a trajectory on which (3.11) holds and G_∞ is finite. For any $t \ge 1$, the allocation comparison with horizon $T = t$ gives

$$\exp(a_i(t)) \le \exp(\ell_t + \eta\Delta_{\max})(1 + \Delta_{\max} S_t), \qquad i > 1.$$

Using (3.8), we obtain

$$\begin{aligned} \exp(\theta_{t,1}) &\le \exp(\ell_t) \prod_{i=2}^{K} \exp(a_i(t)) \\ &\le \exp\left(K\ell_t + (K-1)\eta\Delta_{\max}\right) (1 + \Delta_{\max} S_t)^{K-1}. \end{aligned}$$

For $i > 1$, $a_i(t) \ge 0$ also gives $\exp(\theta_{t,i}) \le \exp(\ell_t)$. Since $\ell_t \ge 0$,

$$Z_t \le K \exp\left(K\ell_t + (K-1)\eta\Delta_{\max}\right) (1 + \Delta_{\max} S_t)^{K-1}.$$

For every $\xi \in (0,1)$, the fact that $\ell_t = o(\log(t))$ implies that eventually $Z_t \le t^{\xi}(1 + \Delta_{\max} S_t)^{K-1}$. Hence

$$S_{t+1} - S_t = \frac{\eta}{Z_t} \ge \frac{\eta t^{-\xi}}{(1 + \Delta_{\max} S_t)^{K-1}}.$$

To sum this inequality, define $F(s) = \frac{(1+\Delta_{\max} s)^K - 1}{K\Delta_{\max}}$. Its derivative is increasing, so eventually

$$\begin{aligned} F(S_{t+1}) - F(S_t) &= \int_{S_t}^{S_{t+1}} (1 + \Delta_{\max} u)^{K-1} \, \mathrm{d}u \\ &\geq (1 + \Delta_{\max} S_t)^{K-1} (S_{t+1} - S_t) \geq \eta t^{-\xi}. \end{aligned}$$

Summing from a sufficiently large round t_0 to T, and using $\sum_{t=t_0}^{T} t^{-\xi} \geq \int_{t_0}^{T+1} u^{-\xi} \, \mathrm{d}u = \frac{(T+1)^{1-\xi} - t_0^{1-\xi}}{1-\xi}$, gives a positive constant c_ξ, depending on the trajectory, such that eventually

$$S_{T+1} \geq c_\xi T^{\frac{1-\xi}{K}}.$$

The lower bound in (3.10) now yields

$$a_i(T+1) \geq \log(\Delta_i) - \ell_T + \log(S_{T+1}), \qquad i > 1.$$

Divide by $\log(T)$, use (3.11), and let $\xi \downarrow 0$. We conclude that almost surely

$$\liminf_{T\to\infty} \frac{a_i(T+1)}{\log(T)} \geq \frac{1}{K}, \qquad i > 1.$$

Passing to expected regret. Since $\sum_{i=2}^{K} a_i(T+1) = \eta \sum_{t=1}^{T} \Delta_t^\pi$ is nonnegative, Fatou's lemma gives

$$\liminf_{T\to\infty} \frac{\eta R_T}{\log(T)} \geq \mathbb{E}\left[\liminf_{T\to\infty} \frac{\sum_{i=2}^{K} a_i(T+1)}{\log(T)} \right] \geq \frac{K-1}{K}.$$

The upper limit was proved from (3.5). Finally, $0 \leq R_T - P_T = B_T \leq B$ gives the same lower limit for P_T, completing (3.3). □

References

Dorian Baudry, Emmeran Johnson, Simon Vary, Ciara Pike-Burke, and Patrick Rebeschini. Does stochastic gradient really succeed for bandits? In *Advances in Neural Information Processing Systems*, 2025. URL https://hal.science/hal-05446148v1.

Sergey Foss, Dmitry Korshunov, and Stan Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2013. doi: 10.1007/978-1-4614-7101-1. URL https://link.springer.com/book/10.1007/978-1-4614-7101-1.

Yanwei Jia and Du Ouyang. Convergence and regret of the policy gradient for multi-armed bandits in diffusion environment. *arXiv preprint arXiv:2607.29593v1*, 2026. URL https://arxiv.org/abs/2607.29593v1.