Annealed Softmax Greedy in Many-Armed Bayesian Bandits

William Overman and Mohsen Bayati

Keywords: Many-armed bandits, Bayesian bandits, exploration, Boltzmann exploration, RLVR.

Summary

RLVR methods such as GRPO reweight sampled completions by reward without explicitly tracking epistemic uncertainty. We study a stylized explanation for when such updates can nevertheless work by analyzing annealed softmax over empirical means in a many-armed Bayesian Bernoulli bandit. Under a β-regular upper-tail condition, we prove Bayes regret $\tilde{O}(m+T/m)$, yielding $\tilde{O}(\sqrt{T})$ when $m = \Theta(\sqrt{T})$ and matching empirical-mean greedy up to logarithmic factors. The key mechanism is the abundance of near-optimal arms, which makes softmax probability assigned away from the empirical best relatively inexpensive. This contrasts with the fixed-K Boltzmann failure mode and suggests a structural analogy to RLVR when the base policy already assigns non-negligible probability to correct or near-optimal completions.

Contribution(s)

1. We prove that annealed softmax over empirical means achieves Bayes regret $\tilde{O}(m + T/m)$ in many-armed Bayesian Bernoulli bandits under a β-regular upper-tail condition, yielding $\tilde{O}(\sqrt{T})$ when $m = \Theta(\sqrt{T})$.
   Context: This provides a positive regret guarantee for uncertainty-agnostic softmax reweighting in a many-armed Bayesian regime, matching the near-greedy benchmark of Bayati et al. (2020) up to logarithmic factors, without using uncertainty bonuses or per-arm confidence intervals.
2. The mechanism behind this guarantee is the abundance of near-optimal arms under β-regularity: with high probability, many arms maintain empirical means close to the optimum throughout learning, so when softmax samples an arm other than the empirically best, that arm tends to be another near-optimal one rather than a clearly inferior one.
   Context: This contrasts with the fixed-K failure mode of monotone-temperature Boltzmann exploration studied by Cesa-Bianchi et al. (2017), where the same softmax policy can suffer linear regret. The many-armed Bayesian regime changes its behavior because the prior puts substantial mass near the optimum.
3. We identify a structural condition—β-regularity—under which uncertainty-agnostic softmax achieves near-greedy Bayes regret, and note a natural analogue in RLVR: a base policy with nontrivially high pass@k already assigns non-negligible probability to correct or near-optimal completions.
   Context: This is a structural analogy rather than a characterization of GRPO. Our theorem is proved only in the non-contextual Bayesian bandit setting, and extending the analysis to the contextual or sequential settings in which RLVR methods operate remains an open question (Shao et al., 2024; Liu et al., 2025b; Yue et al., 2025; Cui et al., 2025).

Annealed Softmax Greedy in Many-Armed Bayesian Bandits

William Overman and Mohsen Bayati
wpo@stanford.edu, bayati@stanford.edu

Stanford University

Abstract

Reinforcement learning with verifiable rewards (RLVR) and group-based policy optimization methods such as GRPO update a stochastic policy by sampling multiple completions per prompt and increasing the policy's probability on those with higher reward, regularized by a KL penalty toward a reference policy. These updates do not include explicit mechanisms that track epistemic uncertainty. This paper studies a stylized explanation for why such uncertainty-agnostic updates can nevertheless be effective. We analyze an annealed softmax (Boltzmann) policy that selects actions according to a softmax of empirical mean rewards in a many-armed Bayesian Bernoulli bandit. Under a β-regular upper-tail condition on the prior, which implies an abundance of near-optimal arms, we prove that annealed softmax greedy achieves Bayes regret $\tilde{O}(m + T/m)$, and in particular $\tilde{O}(\sqrt{T})$ when the number of active arms scales as $m = \Theta(\sqrt{T})$, matching the rate that empirical-mean greedy achieves in the same regime. Under β-regularity, many arms maintain empirical means close to the optimum throughout learning, so when softmax samples an arm other than the empirically best, that arm tends to be another near-optimal one rather than a clearly inferior one. By contrast, with a small number of arms, the same kind of softmax policy can suffer linear regret (Cesa-Bianchi et al., 2017). The result also provides a structural analogy to RLVR, where a base policy with a non-negligible probability of producing a correct completion plays the role of β-regularity.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) is now a common component of language-model post-training pipelines. In RLVR, a model is treated as a stochastic policy over candidate solutions and is trained using rewards that can be checked automatically, such as exact-match accuracy in mathematics or unit-test success in code. This makes it possible to optimize behavior at scale without relying on human preference labels. Group-based policy optimization methods such as Group Relative Policy Optimization (GRPO) are prominent examples of this paradigm (Shao et al., 2024).

RLVR sits somewhat outside the classical exploration picture in reinforcement learning and bandit theory. In those settings, performance is often tied to *epistemic uncertainty*: the learner must gather information to distinguish promising actions from poor ones. Group-based RLVR pipelines, by contrast, sample multiple completions per prompt, raise the policy probability of those with higher reward, and apply a KL penalty toward a reference policy; the randomness comes from repeated sampling rather than from an uncertainty-aware exploration rule. Under verifiable rewards, such updates can be viewed as iteratively amplifying the probability of already successful outputs, with the gain coming largely from redistribution within the model's existing support (Liu et al., 2025b).

This raises a basic question: when can uncertainty-agnostic reweighting work well in problems that seem to require exploration?

A related discussion concerns the base model's *coverage* of good solutions. Here pass@k denotes the probability that at least one of k independent samples from the model is correct, with pass@1 being the per-sample success rate. One line of work reports that RLVR often improves pass@1 while making little progress on pass@k for large k, suggesting that rewarded reasoning paths may already be present in the base model's output distribution (Yue et al., 2025). Other recent work reports settings where RLVR *does* extend reasoning performance, indicating that the empirical picture depends on training details and evaluation choices (Cui et al., 2025). We do not try to resolve that broader debate. We instead study a simplified model that isolates one mechanism by which reweighting alone can be effective.

1.1 A stylized lens: many-armed bandits and annealed softmax

We consider a stochastic many-armed bandit model with Bernoulli rewards and Beta priors, where each arm represents a "completion mode" and reward corresponds to a verifiable success event. Within this model, we analyze an annealed softmax policy that, at each time step t, selects an arm with probability proportional to $\exp(\eta_t m_{i,t})$, where $m_{i,t}$ is the empirical mean reward of arm i at time t and η_t is an inverse-temperature schedule that increases over time, a setup referred to as *annealing*. The policy does not use an optimism term, a posterior sample, or per-arm confidence intervals.

A classical result of Cesa-Bianchi et al. (2017) shows that this type of exploration (also referred to as Boltzmann exploration) can suffer linear regret with a small number of arms under any monotone temperature schedule, and that strong guarantees in that regime typically require schedules that adapt to arm-specific information. The standard failure mode is "cooling too fast": early noise gets amplified, causing premature concentration on suboptimal actions. At first glance, such results argue against exactly the kind of uncertainty-agnostic annealing that appears in RLVR practice.

The many-armed Bayesian setting allows a different possibility. If the prior places enough mass near the optimum, the action space contains many arms whose rewards are already close to optimal. Even if the algorithm spreads probability across several arms, many of them are good enough that the regret cost remains small. This idea is studied in the many-armed bandit literature, where greedy-style procedures achieve Bayesian regret guarantees under upper-tail regularity conditions on the prior (Bayati et al., 2020).

1.2 Main results and perspective

The question we study is therefore: can annealed softmax, despite ignoring epistemic uncertainty, inherit the regret behavior of greedy methods in the many-armed Bayesian regime?

The answer is yes for the Beta-Bernoulli model under an upper-tail regularity condition on the prior. Informally, this condition implies that near-optimal arms are sufficiently abundant: with many active arms, many of them are close to the optimum. Under this condition, a time-only annealing schedule suffices for annealed softmax to achieve the same Bayes regret scaling as the many-armed greedy benchmark, up to logarithmic factors.

Annealed softmax differs from greedy only because it places some probability mass on arms that are not currently empirically best. Under the upper-tail regularity condition, this leakage is cheap: most of the diverted probability lands on arms that are themselves near-optimal.

1.3 Connection to RLVR and pass@k

Although the formal results are proved in a bandit model, they map to one structural view of RLVR with group sampling. In language modeling, "actions" are not literal discrete arms, but the base policy still induces a distribution over candidate completions of varying quality. Metrics such as pass@k probe the upper tail of that distribution. If correct completions occur with non-negligible probability, repeated sampling surfaces them reliably, and subsequent reweighting can improve pass@1 without an explicit uncertainty-aware exploration rule.

Our result is far from a model of what happens in GRPO. The setting is a standard multi-armed bandit, with no context and no sequential structure. The connection it suggests is structural, and extending the analysis to the contextual or sequential settings where GRPO operates is an open problem.

1.4 Paper organization

Section 2 reviews related work. Section 3 introduces the many-armed Bayesian Bernoulli bandit model, the β-regular prior assumption, and notation. Section 4 presents the greedy baseline and the annealed softmax greedy (ASG) algorithm. Section 5 states the main regret theorem and its extensions. Appendix A contains the proofs, and Section 6 concludes with discussion and limitations. We complement the theoretical analysis with simulations on Bernoulli bandits in Appendix B.

2 Related Work

Our paper connects three strands of literature: the theory of Boltzmann exploration in bandits, the many-armed bandit framework with "free exploration," and the empirical RLVR/GRPO pipeline for LLM post-training. We give a concise overview here and defer a comprehensive survey to Appendix C.

Boltzmann exploration and its limitations. Boltzmann (softmax) action selection is a standard randomized alternative to greedy or optimistic policies in bandits and RL. Cesa-Bianchi et al. (2017) prove that, for fixed-K stochastic bandits, *any monotone* temperature schedule can be forced into suboptimal behavior—either exploring too long or committing too early—and propose per-arm learning rates that explicitly track uncertainty as a remedy. This negative result is the starting point of our analysis: we identify a structural regime (many arms, thick-tailed prior) that circumvents the impossibility without requiring uncertainty-aware schedules.

Many-armed bandits and free exploration. When the number of arms is large relative to the horizon, the distribution of arm qualities (rather than per-arm estimation) governs achievable regret. In the Bayesian setting, Bayati et al. (2020) show that under an upper-tail regularity condition on the prior ("β-regularity"), a subsampled greedy policy achieves Bayes regret $\tilde{O}(\max\{m, T/m\})$, yielding $\tilde{O}(\sqrt{T})$ with $m = \Theta(\sqrt{T})$ arms. The key mechanism is *free exploration*: discarding a poorly performing arm still leaves many near-optimal alternatives. Our work extends this viewpoint from greedy to annealed softmax policies, showing that the same prior-tail structure suppresses the "softmax leakage" that causes failures in the fixed-K regime. Related infinite-armed formulations include Berry et al. (1997); Wang et al. (2008); Carpentier & Valko (2015).

RLVR, GRPO, and the "bounded-by-base" phenomenon. Reinforcement learning with verifiable rewards has become a core post-training technique for reasoning-oriented LLMs. GRPO Shao et al. (2024) performs group-sampled policy updates under KL regularization, and recent analyses relate its dynamics to iterative soft reweighting of completions Mroueh (2025); Liu et al. (2025b). A central empirical observation is that RLVR often improves pass@1 while failing to improve, or even degrading, large-k pass@k, suggesting that optimization primarily redistributes probability mass within the base model's existing support rather than discovering new reasoning paths Yue et al. (2025); though subsequent work argues this picture is nuanced and depends on training details and evaluation choices Cui et al. (2025); Wen et al. (2025); Liu et al. (2025a). Our bandit model provides a stylized formalization of this "bounded-by-base" effect: the β-regular tail condition translates "good pass@k" into "many near-optimal arms," under which softmax reweighting suffices without explicit exploration.

3 Problem Setup

This section sets up the model and the notation. More specifically, it formalizes the bandit model, defines the prior regularity condition and the Bayes regret criterion, and introduces the empirical-mean notation used in the algorithm. Throughout the paper we assume $T \geq m$.

3.1 Many-armed Bayesian Bernoulli bandit

Fix a time horizon $T \in \mathbb{N}$ and consider $m \in \mathbb{N}$ arms indexed by $i \in [m] := \{1, \dots, m\}$. Each arm has an unknown mean $\mu_i \in [0, 1]$ drawn i.i.d. from a prior Γ on $[0, 1]$. Conditional on μ_i, each arm i has an i.i.d. sequence of Bernoulli rewards

$$X_{i,s} \mid \mu_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(\mu_i) \qquad (s = 1, 2, \dots),$$

where $X_{i,s}$ is the reward from the s-th pull of arm i (s is arm-specific, not the global time index). At each round $t \in [T]$, a policy selects an arm $A_t \in [m]$ (possibly randomized and history-dependent) and observes $X_{A_t, N_{A_t}(t)}$, where

$$N_i(t) := \sum_{s=1}^{t} \mathbf{1}\{A_s = i\}$$

is the number of pulls of arm i up to time t.

3.2 Prior regularity: β-regular upper tail

We assume Γ has a polynomially thick upper tail near 1.

Definition 3.1 (β-regular prior near 1)**.** A distribution Γ on $[0, 1]$ is β*-regular* if there exist constants $0 < c_0 \leq C_0 < \infty$ and $\varepsilon_0 \in (0, 1)$ such that, for all $\varepsilon \in (0, \varepsilon_0]$,

$$c_0\, \varepsilon^{\beta} \;\leq\; \Gamma([1 - \varepsilon, 1]) \;\leq\; C_0\, \varepsilon^{\beta}.$$

We will present the main theorem first for the *linear tail* case $\beta = 1$ (which is the regime treated in Bayati et al. (2020) for simplicity), and then state a clean general-β extension as a corollary/remark.

3.3 Bayes regret and a convenient surrogate

Let $\mu_* := \max_{i \in [m]} \mu_i$. The (frequentist) regret of a policy π given $\boldsymbol{\mu} = (\mu_1, \dots, \mu_m)$ is

$$R_T(\pi \mid \boldsymbol{\mu}) \;:=\; \sum_{t=1}^{T} \bigl(\mu_* - \mu_{A_t}\bigr).$$

The *Bayes regret* is

$$\mathrm{BR}_{T,m}(\pi) \;:=\; \mathbb{E}\bigl[R_T(\pi \mid \boldsymbol{\mu})\bigr],$$

where the expectation is over $\boldsymbol{\mu} \sim \Gamma^{\otimes m}$ (i.e., $\mu_1, \dots, \mu_m$ are i.i.d. draws from Γ, so $\Gamma^{\otimes m}$ is the m-fold product measure on $[0, 1]^m$), reward randomness, and any internal randomness of π.

We will also use the standard surrogate "regret-to-1"

$$\widetilde{R}_T(\pi \mid \boldsymbol{\mu}) \;:=\; \sum_{t=1}^{T} (1 - \mu_{A_t}) \;=\; \sum_{i=1}^{m} (1 - \mu_i)\, N_i(T), \tag{1}$$

which satisfies the exact identity

$$R_T(\pi \mid \boldsymbol{\mu}) \;=\; \widetilde{R}_T(\pi \mid \boldsymbol{\mu}) \;-\; T(1 - \mu_*). \tag{2}$$

Thus

$$\mathrm{BR}_{T,m}(\pi) \;=\; \mathbb{E}\bigl[\widetilde{R}_T(\pi \mid \boldsymbol{\mu})\bigr] \;-\; T\,\mathbb{E}[1 - \mu_*] \;\leq\; \mathbb{E}\bigl[\widetilde{R}_T(\pi \mid \boldsymbol{\mu})\bigr]. \tag{3}$$

Under 1-regularity, $\mathbb{E}[1 - \mu_*]$ is of order $1/m$ (Lemma A.9 below), so the exact identity can sharpen constants, but our main upper bound only needs $\mathrm{BR}_{T,m}(\pi) \leq \mathbb{E}[\widetilde{R}_T(\pi \mid \boldsymbol{\mu})]$.

3.4 Empirical means

Let $S_i(t)$ be the number of observed successes of arm i up to time t, so $0 \leq S_i(t) \leq N_i(t)$. Because every arm is pulled once during initialization in Algorithms 1–2, we have $N_i(t) \geq 1$ throughout the greedy/softmax phase. We therefore use the empirical mean

$$\widehat{\mu}_{i,t} \;:=\; \frac{S_i(t)}{N_i(t)} \tag{4}$$

as the score of arm i at time t.

Asymptotic notation. Throughout the paper, $\tilde{O}(\cdot)$ hides polylogarithmic factors in T and m.

4 Algorithms

This section presents the greedy baseline and the annealed softmax greedy (ASG) policy that is the main object of analysis.

4.1 Greedy baseline on empirical means

We begin with the natural greedy benchmark, shown in Algorithm 1. After pulling each arm once so that every arm has an initial estimate, the policy always selects an arm with the largest empirical mean. This is exactly the many-armed greedy rule analyzed by Bayati et al. (2020), and it serves as the baseline throughout the paper.

4.2 Annealed Softmax Greedy on empirical means

Our main object of study is a randomized analogue of greedy, given in Algorithm 2. Instead of deterministically choosing the empirical-best arm at each round, the policy samples from a softmax distribution over empirical means. Arms with higher empirical means are more likely to be selected, but lower-ranked arms still receive some probability mass. The amount of randomness is controlled by a nondecreasing inverse-temperature schedule $\{\eta_t\}_{t\geq 1}$.

Algorithm 1 Greedy (Empirical-Mean Greedy)

Require: Arms $[m]$, horizon T

Initialization:
for $t = 1, \ldots, m$ **do**
  Pull arm $A_t = t$; observe reward $X_{t,1}$
end for
Greedy phase:
for $t = m+1, \ldots, T$ **do**
  Compute empirical means $\widehat{\mu}_{i,t-1} = \frac{S_i(t-1)}{N_i(t-1)}$ for all $i \in [m]$
  Pull arm $A_t \in \arg\max_{i\in[m]} \widehat{\mu}_{i,t-1}$ (ties broken arbitrarily)
  Observe reward $X_{A_t,N_{A_t}(t)}$
end for

This policy can be viewed as interpolating between uniform sampling and pure greedy behavior. When η_t is small, the softmax distribution is relatively flat, so the algorithm spreads probability more evenly across all arms. As η_t grows, the distribution becomes more concentrated, and the policy increasingly resembles Greedy. The key question for the rest of the paper is whether this softmax sampling rule, which is agnostic to epistemic uncertainty in arm quality, can achieve near-greedy regret in the many-armed Bayesian regime.

ASG uses exactly the same empirical information as the greedy baseline; it neither estimates uncertainty explicitly nor uses per-arm temperatures that depend on the pull counts $N_i(t)$. Cesa-Bianchi

Algorithm 2 Annealed Softmax Greedy (ASG)

Require: Arms $[m]$, horizon T, nonnegative nondecreasing inverse-temperature schedule $\{\eta_t\}_{t\geq 1}$ with $\eta_t \to \infty$

Initialization:
for $t = 1, \dots, m$ **do**
  Pull arm $A_t = t$; observe reward $X_{t,1}$
end for
Softmax phase:
for $t = m+1, \dots, T$ **do**
  Compute empirical means $\widehat{\mu}_{i,t-1}$ for all $i \in [m]$
  Sample arm $A_t = i$ with probability

$$p_t(i) \;:=\; \frac{\exp(\eta_t\, \widehat{\mu}_{i,t-1})}{\sum_{j=1}^{m} \exp(\eta_t\, \widehat{\mu}_{j,t-1})} \tag{6}$$

  Observe reward $X_{A_t, N_{A_t}(t)}$
end for

et al. (2017) prove that any monotone temperature schedule shared across arms can suffer linear regret in some small-arm instance, and propose per-arm scalings as a remedy: their Boltzmann–Gumbel Exploration uses $\sigma/\sqrt{N_i(t)}$, so less-pulled arms have noisier scores and are favored for exploration. We study the unadjusted version because it matches the structure used in group-based methods such as GRPO, where the policy does not adapt its per-completion sampling temperature to historical counts.

Concrete schedules. For the analysis, it is convenient to choose a logarithmically increasing schedule of the form

$$\eta_t \;=\; \frac{c_\eta}{\delta}\log(t \vee 2), \tag{5}$$

where $c_\eta > 1$ is a fixed constant and δ is a threshold parameter that will be tuned as a function of (T, m) in the regret bound. Intuitively, δ sets the scale of what counts as a meaningfully suboptimal arm in the proof. With this choice, $\exp(-\eta_t\delta) = (t \vee 2)^{-c_\eta}$, so the probability weight placed on arms that are worse by at least δ decays polynomially in time. In particular,

$$\sum_{t=1}^{T} \exp(-\eta_t \delta) = O(1),$$

which is exactly what we need to make the total softmax "leakage" summable. In the linear-tail case ($\beta = 1$), for example, the relevant choice is $\delta \asymp (\log T)/m$, leading to the $\tilde{O}(m + T/m)$ Bayes regret rate (Theorem 5.2).

5 Main Results

This section presents the main Bayes regret bound for ASG.

Before stating our main theorem, we recall the result for the greedy policy (Algorithm 1). In the Bernoulli many-armed setting with a linear upper tail, Bayati et al. (2020) show that empirical-mean greedy achieves near-optimal Bayes regret.

Proposition 5.1 (Greedy benchmark in the Bernoulli, $\beta = 1$ regime (Bayati et al., 2020, Theorem (Bernoulli)))**.** *Assume Bernoulli rewards and a* 1*-regular prior* Γ *(Definition 3.1 with* $\beta = 1$*). Then Greedy (Algorithm 1) achieves*

$$\mathrm{BR}_{T,m}(\textit{Greedy}) \;\leq\; \tilde{O}\left(m + \frac{T}{m}\right),$$

and in particular for $m = \Theta(\sqrt{T})$ *one has* $\mathrm{BR}_{T,m}(\textit{Greedy}) = \tilde{O}(\sqrt{T})$.

5.1 ASG matches greedy up to a leakage term

We now turn to our main result for ASG. At a high level, the theorem shows that ASG behaves like greedy plus an additional *softmax leakage* penalty: the probability mass assigned away from the empirically best arms. The key technical point is that in the many-armed, thick-tail regime, this leakage remains controlled because when softmax samples an arm other than the empirically best, that arm tends to be another near-optimal one rather than a clearly inferior one. The extra randomization introduced by softmax therefore does not change the leading Bayes regret scaling.

The bound below decomposes regret into several interpretable terms. The terms m and $T\delta$ are the same coarse baseline contributions that already appear in greedy-style many-armed analyses. The logarithmic term $m(1+\log(1/\delta))$ reflects the cost of integrating over the upper tail. The summation term $\frac{1}{\delta}\sum_{t=1}^{T}\exp(-\eta_t\delta)$ is the softmax-specific leakage term, controlled by the cooling schedule. Finally, the exponentially small term $T\exp(-cm\delta)$ corresponds to the bad event that too few of the m arms are near-optimal. In this sense, ASG is essentially greedy up to a leakage penalty that becomes small under suitable annealing.

Theorem 5.2 (ASG Bayes regret, Bernoulli, linear tail)**.** *Assume Bernoulli rewards, and assume the prior* Γ *is* 1*-regular (Definition 3.1 with* $\beta = 1$*) with regularity constants* $(c_0, C_0, \varepsilon_0)$*. Fix any* $\delta \in (0, \delta_0]$ *with* $\delta_0 \le \min\{1/8, \varepsilon_0/8\}$ *(which also ensures* $\delta < 1/6$*, the condition required by Lemma A.2). Run ASG (Algorithm 2) with any nonnegative nondecreasing schedule* $\{\eta_t\}_{t\ge1}$.

Then there exist constants $c, C > 0$ *(depending only on* c_0, C_0, ε_0 *and on the Bernoulli crossing constant in Lemma A.2) such that*

$$\mathrm{BR}_{T,m}(\textit{ASG}) \;\le\; C\left[m \;+\; T\,\delta \;+\; m\big(1+\log(1/\delta)\big) \;+\; \frac{1}{\delta}\sum_{t=1}^{T}\exp(-\eta_t\,\delta) \;+\; T\,\exp(-c\,m\,\delta)\right]. \tag{7}$$

In particular, choose

$$\delta \;=\; \min\left\{\delta_0,\, A\frac{\log(T\vee 2)}{m}\right\}, \qquad \eta_t \;=\; \frac{c_\eta}{\delta}\log(t\vee 2) \quad (c_\eta > 1), \tag{8}$$

with $A > 1/c$*. Then*

$$\mathrm{BR}_{T,m}(\textit{ASG}) \;=\; \tilde{O}\left(m + \frac{T}{m}\right),$$

and for $m = \Theta(\sqrt{T})$ *one obtains* $\mathrm{BR}_{T,m}(\textit{ASG}) = \tilde{O}(\sqrt{T})$.

The proof is given in Section A. The main takeaway is that ASG inherits the same leading-order regret scaling as greedy in the linear-tail many-armed regime. In this sense, the additional softness of the policy preserves the near-greedy rate; it only introduces a leakage term that can be made summable by a suitable logarithmic cooling schedule.

Remark 5.3 (General β (statement-level)). The linear-tail case $\beta = 1$ is the cleanest to state, but the same proof strategy extends to general $\beta > 0$, paralleling the corresponding generalization of the greedy benchmark in Bayati et al. (2020). Repeating the proof with Lemma A.3 replaced by the general-β analogue $p_\delta \ge c\,\delta^\beta$ (which follows from the same crossing argument together with the β-regularity bound $\Gamma([1-\delta,1]) \ge c_0\delta^\beta$), and with the corresponding many-good-arms bound $\mathbb{P}(M(\delta) < r) \le \exp(-cm\delta^\beta)$, yields a bound of the form

$$\mathrm{BR}_{T,m}(\mathrm{ASG}) \;\le\; \tilde{O}\left(m \;+\; T\delta \;+\; m\cdot\mathfrak{M}_\beta(\delta) \;+\; \frac{1}{\delta^\beta}\sum_{t=1}^{T}e^{-\eta_t\delta} \;+\; Te^{-cm\delta^\beta}\right),$$

where $\mathfrak{M}_\beta(\delta)$ captures the tail-moment scaling from Lemma A.8. In particular,

$$\mathfrak{M}_1(\delta) = 1 + \log(1/\delta), \qquad \mathfrak{M}_\beta(\delta) = O(1) \text{ for } \beta > 1, \qquad \mathfrak{M}_\beta(\delta) = \Theta(\delta^{-(1-\beta)}) \text{ for } \beta \in (0,1).$$

Optimizing this bound with $\delta \asymp (\log T/m)^{1/\beta}$ (paralleling the $\beta = 1$ specialization) yields rates of the form $\tilde{O}\big(m + T(\log T/m)^{1/\beta}\big)$ for $\beta \geq 1$, matching the corresponding generalization of the greedy benchmark in Bayati et al. (2020). For $\beta < 1$, the same procedure yields a similar expression with an additional $\mathfrak{M}_\beta$ contribution. The optimized rate is *not* $\tilde{O}(m + T/m)$ outside $\beta = 1$. The qualitative message is preserved: thicker upper tails make softmax leakage cheaper, and thinner tails give correspondingly slower rates.

6 Discussion and Limitations

Theorem 5.2 identifies a regime in which *uncertainty-agnostic* reweighting can still be statistically effective. The key mechanism is multiplicity: when the prior has a sufficiently thick upper tail, there are many arms whose means lie close to optimal, and a nontrivial fraction of them remain statistically competitive over time. As a result, the softmax denominator contains many near-optimal competitors, so probability mass assigned away from the current best arm is not necessarily wasted on clearly inferior actions. In this regime, "leakage" is cheap because it is spread across many nearly optimal arms.

This perspective helps reconcile our result with classical negative results for Boltzmann exploration. In fixed-K settings, annealing can amplify early noise and lead to persistent selection of suboptimal arms. Our theorem shows that this failure mode is not universal: in a many-armed Bayesian regime with sufficient upper-tail mass, the same uncertainty-agnostic softmax mechanism achieves near-greedy Bayes regret. The difference is not just the policy class, but the geometry of the action space induced by the prior.

The result also suggests a stylized interpretation for group-based methods such as GRPO and, more broadly, RLVR. We do *not* analyze language models, sequence generation, or policy-gradient training. Rather, we isolate one mechanism by which simple reweighting can work well: if the base policy already has strong upper-tail coverage, then reweighting among sampled candidates may improve top performance without requiring explicit uncertainty bonuses or directed exploration.

Several limitations are important. First, our model is narrow: we study i.i.d. Bayesian Bernoulli bandits, with no context, no shared structure across actions, and no sequential state dynamics. Second, the guarantees are Bayesian rather than minimax, and rely essentially on upper-tail regularity of the prior; if near-optimal arms are rare, the multiplicity mechanism weakens. Third, we do not claim optimality of annealed softmax even in this regime, only that it matches the near-greedy benchmark up to logarithmic factors under our assumptions. Finally, the connection to GRPO and pass@k is conceptual rather than literal, since modern RLVR systems involve shared parametric policies, dependent samples, and non-bandit credit assignment.

A natural next step is to extend this analysis to structured or contextual action spaces, and to sequential RL settings where "many near-optimal actions" might be replaced by many near-optimal trajectories or completions. More broadly, an important open question is to characterize the boundary of the phenomenon: precisely when does upper-tail multiplicity make uncertainty-agnostic reweighting effective, and when do the classical Boltzmann failure modes re-emerge?

Broader Impact Statement

This work is primarily theoretical and studies fundamental properties of exploration strategies in multi-armed bandits. We do not foresee specific negative societal impacts from the results presented here.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2312–2320, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.html.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23 of *Proceedings of Machine Learning Research*, pp. 39.1–39.26. PMLR, 2012. URL https://proceedings.mlr.press/v23/agrawal12.html.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory (ALT)*, volume 4754 of *Lecture Notes in Computer Science*, pp. 150–165. Springer, 2007. DOI: 10.1007/978-3-540-75225-7_15. URL https://doi.org/10.1007/978-3-540-75225-7_15.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002. DOI: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Hamsa Bastani, Mohsen Bayati, and Khashayar Khosravi. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2020. DOI: 10.1287/mnsc.2020.3605. URL https://doi.org/10.1287/mnsc.2020.3605.

Mohsen Bayati, Nima Hamidi, Ramesh Johari, and Khashayar Khosravi. Unreasonable effectiveness of greedy algorithms in multi-armed bandit with many arms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/12d16adf4a9355513f9d574b76087a08-Abstract.html.

Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp. Bandit problems with infinitely many arms. *The Annals of Statistics*, 25(5):2103–2116, 1997. DOI: 10.1214/aos/1069362389. URL https://doi.org/10.1214/aos/1069362389.

Thomas Bonald and Alexandre Proutière. Two-target algorithms for infinite-armed bandits with Bernoulli rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2184–2192, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/fc2c7c47b918d0c2d792a719dfb602ef-Abstract.html.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. DOI: 10.1561/2200000024. URL https://doi.org/10.1561/2200000024.

Alexandra Carpentier and Michal Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1133–1141. PMLR, 2015. URL https://proceedings.mlr.press/v37/carpentier15.html.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, and Gergely Neu. Boltzmann exploration done right. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1705.10257.

Arghya Roy Chaudhuri and Shivaram Kalyanakrishnan. Quantile-regret minimisation in infinitely many-armed bandits. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 425–434, 2018. URL http://auai.org/uai2018/proceedings/papers/169.pdf.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL https://arxiv.org/abs/2505.22617.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948. Published in Nature 645, 633–638 (2025), doi:10.1038/s41586-025-09422-z.

John C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979. DOI: 10.1111/j.2517-6161.1979.tb01068.x. URL https://doi.org/10.1111/j.2517-6161.1979.tb01068.x.

Botao Hao, Tor Lattimore, and Csaba Szepesvári. Adaptive exploration in linear contextual bandit. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3536–3545. PMLR, 2020. URL https://proceedings.mlr.press/v108/hao20b.html.

Sampath Kannan, Jamie H. Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2227–2236, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/2cfd4560539f887a5e420412b370b361-Abstract.html.

Emilie Kaufmann. On Bayesian index policies for sequential resource allocation. *The Annals of Statistics*, 46(2):842–865, 2018. DOI: 10.1214/17-AOS1569. URL https://doi.org/10.1214/17-AOS1569.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. DOI: 10.1016/0196-8858(85)90002-8. URL https://doi.org/10.1016/0196-8858(85)90002-8.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. DOI: 10.1017/9781108571401. URL https://doi.org/10.1017/9781108571401.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models, 2025a. URL https://arxiv.org/abs/2505.24864.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective, 2025b. URL https://arxiv.org/abs/2503.20783.

Youssef Mroueh. Reinforcement learning with verifiable rewards: GRPO's effective loss, dynamics, and success amplification, 2025. URL https://arxiv.org/abs/2503.06639.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155. Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. Simko: Simple pass@K policy optimization, 2025. URL https://arxiv.org/abs/2510.14807.

Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, and Zhiwei Steven Wu. The externalities of exploration and how data diversity helps exploitation. In *Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1724–1738. PMLR, 2018. URL https://proceedings.mlr.press/v75/raghavan18a.html.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1583–1591, 2014a. URL https://papers.nips.cc/paper/5463-learning-to-optimize-via-information-directed-sampling.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014b. DOI: 10.1287/moor.2014.0650. URL https://doi.org/10.1287/moor.2014.0650.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016. URL https://jmlr.org/papers/v17/14-087.html.

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning, 2018. URL https://arxiv.org/abs/1803.02855.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards, 2025. URL https://arxiv.org/abs/2505.24760. NeurIPS 2025 Spotlight.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. DOI: 10.2307/2332286. URL https://doi.org/10.2307/2332286.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In *Advances in Neural Information Processing Systems 21 (NeurIPS)*, pp. 1729–1736, 2008. URL https://proceedings.neurips.cc/paper/2008/hash/49ae49a23f67c759bf4fc791ba842aa2-Abstract.html.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs, 2025. URL https://arxiv.org/abs/2506.14245.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yang Yu. Pass@k metric for RLVR: A diagnostic tool of exploration, but not an objective, 2025. URL https://arxiv.org/abs/2511.16231.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? 2025. URL https://arxiv.org/abs/2504.13837.

Guanning Zeng, Zhaoyi Zhou, Daman Arora, and Andrea Zanette. Shrinking the variance: Shrinkage baselines for reinforcement learning with verifiable rewards, 2025. URL https://arxiv.org/abs/2511.03710.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A survey of reinforcement learning for large reasoning models, 2025. URL https://arxiv.org/abs/2509.08827.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for KL-regularized contextual bandits and RLHF, 2025. URL https://openreview.net/forum?id=TE63KPCXWt. NeurIPS 2025 poster; arXiv:2411.04625.

A Proofs

The proof technique follows the many-armed analysis of Bayati et al. (2020), adapted to control the softmax leakage probability. Even when the score vector is already informative, Boltzmann exploration still assigns some probability mass to inferior arms (Cesa-Bianchi et al., 2017). Under a 1-regular prior and many available arms, we show that the presence of many persistently good arms suppresses this leakage enough to recover the many-armed regret rate.

Throughout, $\mathcal{F}_t$ denotes the history σ-field generated by actions and observations up to time t.

A.1 A Bernoulli "never-crossing" tail event

Definition A.1 (Never-crossing probability $q_\theta(\mu)$)**.** Fix $\theta \in (0,1)$ and $\mu \in [0,1]$. Let $\{X_s\}_{s\geq 1}$ be i.i.d. Bernoulli(μ) and let $\widehat{\mu}(n) := \frac{1}{n}\sum_{s=1}^{n} X_s$. Define

$$q_\theta(\mu) \;:=\; \mathbb{P}\big(\widehat{\mu}(n) > \theta \text{ for all } n \geq 1\big),$$

with the convention $q_\theta(\mu) = 0$ for $\mu \leq \theta$.

Lemma A.2 (Bernoulli crossing bound (Bayati et al., 2020, Lemma 4.2))**.** *Let $X_s \sim \mathrm{Bernoulli}(\mu)$ i.i.d. and fix $\theta > 2/3$. If $\mu \geq (1+\theta)/2$, then*

$$q_\theta(\mu) \;\geq\; C_{\mathrm{Bern}} \;:=\; \frac{e^{-1/2}}{3}.$$

A.2 Many always-good arms under a β-regular tail

We now formalize the event that the prior produces many "always-good" arms, which will be used to control softmax leakage.

Fix $\delta \in (0, 1/8]$ and define two thresholds

$$\theta := 1 - 2\delta, \qquad \theta' := 1 - 3\delta.$$

For each arm $i \in [m]$, define the *always-good* event

$$\mathcal{G}_i(\delta) \;:=\; \Big\{\mu_i \geq 1-\delta\Big\} \,\cap\, \Big\{\widehat{\mu}_i(n) > \theta \;\; \forall n \geq 1\Big\}, \tag{9}$$

where $\widehat{\mu}_i(n)$ denotes the empirical mean of the *first n rewards* of arm i (a property of the arm's reward sequence, independent of the policy).

Let

$$M(\delta) := \sum_{i=1}^{m} \mathbf{1}\{\mathcal{G}_i(\delta)\}$$

be the number of always-good arms.

Lemma A.3 (Expected mass of always-good arms)**.** *Assume Γ is 1-regular (Definition 3.1 with $\beta = 1$). There exists $\delta_0 > 0$ and a constant $c > 0$ such that for all $\delta \in (0, \delta_0]$,*

$$p_\delta := \mathbb{P}\big(\mathcal{G}_i(\delta)\big) = \mathbb{E}_{\mu\sim\Gamma}\big[\mathbf{1}\{\mu \geq 1-\delta\}\, q_{1-2\delta}(\mu)\big] \geq c\,\delta.$$

Proof. By Lemma A.2, for $\delta < 1/6$ and $\mu \geq 1-\delta$, we have $q_{1-2\delta}(\mu) \geq C_{\text{Bern}}$ since $1-\delta \geq (1+(1-2\delta))/2$. Therefore

$$p_\delta \geq C_{\text{Bern}} \cdot \mathbb{P}(\mu \geq 1-\delta).$$

By 1-regularity, $\mathbb{P}(\mu \geq 1-\delta) \geq c_0\,\delta$ for all sufficiently small δ, hence $p_\delta \geq C_{\text{Bern}} c_0\,\delta$. □

Lemma A.4 (Many always-good arms with high probability)**.** *Under the assumptions of Lemma A.3, there exist constants $c, c' > 0$ such that for all sufficiently small δ,*

$$\mathbb{P}\Big(M(\delta) \geq r(\delta)\Big) \geq 1-\exp(-c'\,m\,\delta), \qquad r(\delta) := \max\Big\{1, \Big\lfloor \frac{mp_\delta}{2} \Big\rfloor\Big\}.$$

Proof. Across arms, the pairs $(\mu_i, (X_{i,s})_{s\geq 1})$ are i.i.d., so the indicators $\mathbf{1}\{\mathcal{G}_i(\delta)\}$ are i.i.d. Bernoulli with mean p_δ. Thus $M(\delta) \sim \text{Binomial}(m, p_\delta)$.

If $mp_\delta \geq 2$, then $r(\delta) = \lfloor mp_\delta/2 \rfloor$ and a multiplicative Chernoff bound gives

$$\mathbb{P}\big(M(\delta) < r(\delta)\big) \leq \mathbb{P}\big(M(\delta) < mp_\delta/2\big) \leq \exp(-mp_\delta/8).$$

If $mp_\delta < 2$, then $r(\delta) = 1$ and

$$\mathbb{P}\big(M(\delta) < r(\delta)\big) = \mathbb{P}\big(M(\delta) = 0\big) = (1-p_\delta)^m \leq \exp(-mp_\delta).$$

In either case,

$$\mathbb{P}\big(M(\delta) < r(\delta)\big) \leq \exp(-c\,mp_\delta).$$

Lemma A.3 gives $p_\delta \geq c'\delta$, which yields the stated bound after adjusting constants. □

A.3 A key inequality: softmax leakage controlled by r good arms

We next bound how often a suboptimal arm can be sampled when many good arms maintain a score margin.

Lemma A.5 (Pointwise leakage bound for softmax)**.** *Fix time t, scores $\{s_j\}_{j=1}^m \subset \mathbb{R}$, an inverse temperature $\eta_t \geq 0$, and an integer $r \in \{1, \ldots, m\}$. Let $p_t(i) \propto \exp(\eta_t s_i)$. If there is a subset $\mathcal{J} \subset [m]$ with $|\mathcal{J}| = r$ such that $s_j \geq s^\star$ for all $j \in \mathcal{J}$, then for any i,*

$$p_t(i) \leq \frac{\exp(\eta_t s_i)}{r\,\exp(\eta_t s^\star)} = \frac{1}{r}\exp\big(-\eta_t(s^\star - s_i)\big).$$

In particular, if $s^\star - s_i \geq \Delta > 0$, then $p_t(i) \leq \frac{1}{r}e^{-\eta_t\Delta}$.

Proof. Immediate from $\sum_{j=1}^m \exp(\eta_t s_j) \geq \sum_{j\in\mathcal{J}} \exp(\eta_t s_j) \geq r\exp(\eta_t s^\star)$. □

A.4 Bounding pull counts of suboptimal arms: spikes + leakage

We now use a simple "spikes + leakage" decomposition for empirical means.

Fix δ and thresholds $\theta = 1 - 2\delta$, $\theta' = 1 - 3\delta$ as above. For a fixed arm i, let

$$\widehat{\mu}_i(n) \;:=\; \frac{1}{n}\sum_{s=1}^{n} X_{i,s}$$

be the empirical mean of the first n rewards of arm i.

Lemma A.6 (Spikes of empirical means are exponentially rare)**.** *For any arm mean $\mu < \theta'$ and any $n \geq 1$,*

$$\mathbb{P}\big(\widehat{\mu}_i(n) \geq \theta' \;\big|\; \mu_i = \mu\big) \;\leq\; \exp\big(-2n(\theta' - \mu)^2\big).$$

Consequently,

$$\sum_{n=1}^{\infty} \mathbb{P}\big(\widehat{\mu}_i(n) \geq \theta' \mid \mu_i = \mu\big) \;\leq\; \frac{1}{2(\theta' - \mu)^2}.$$

Proof. The first claim is Hoeffding's inequality for Bernoulli averages. For the second, write $a = \theta' - \mu > 0$ and sum the geometric bound:

$$\sum_{n=1}^{\infty} e^{-2na^2} = \frac{e^{-2a^2}}{1 - e^{-2a^2}} = \frac{1}{e^{2a^2} - 1} \leq \frac{1}{2a^2},$$

using $e^x - 1 \geq x$ for $x \geq 0$. □

Good arms provide a uniform score floor. On the event $\mathcal{G}_i(\delta)$ in (9), we have $\widehat{\mu}_i(n) > \theta$ for every $n \geq 1$. Since each arm is pulled once during initialization, it follows that for every such arm and every $t \geq m + 1$,

$$\widehat{\mu}_{i,t-1} = \widehat{\mu}_i\big(N_i(t-1)\big) > \theta.$$

Thus, if $M(\delta) \geq r$, then throughout the softmax phase there are always at least r arms whose current empirical means exceed θ. This is the key simplification in the empirical-mean proof: no maturation argument is needed.

A.5 Pull-count bound conditional on having r good arms

We now state the pull-count lemma.

Lemma A.7 (Pull-count bound under r reference arms)**.** *Fix an integer $r \in \{1, \ldots, m\}$, fix δ, and set thresholds $\theta = 1 - 2\delta$, $\theta' = 1 - 3\delta$. For each $t \in \{m+1, \ldots, T\}$, define the $\mathcal{F}_{t-1}$-measurable event*

$$\mathcal{E}_t := \Big\{\exists \mathcal{J} \subset [m] \text{ with } |\mathcal{J}| = r \text{ such that } \widehat{\mu}_{j,t-1} \geq \theta \;\forall j \in \mathcal{J}\Big\}.$$

Then for any arm i and any horizon T,

$$\mathbb{E}\Bigg[N_i(T)\,\mathbf{1}\Big\{\bigcap_{t=m+1}^{T} \mathcal{E}_t\Big\}\Bigg] \;\leq\; 1 + \sum_{n=1}^{\infty} \mathbb{P}\big(\widehat{\mu}_i(n) \geq \theta'\big) + \frac{1}{r}\sum_{t=m+1}^{T} \exp(-\eta_t\,(\theta - \theta')). \tag{10}$$

Proof. Let $\tau_{i,n}$ be the time of the n-th pull of arm i, with $\tau_{i,n} = \infty$ if the n-th pull never occurs. If $\tau_{i,n+1} \leq T$, then at time $\tau_{i,n+1} - 1$ the arm has been observed exactly n times, hence

$$\widehat{\mu}_{i,\tau_{i,n+1}-1} = \widehat{\mu}_i(n).$$

Therefore

$$N_i(T)\,\mathbf{1}\Big\{\bigcap_{t=m+1}^{T} \mathcal{E}_t\Big\} \leq 1 + \sum_{n=1}^{\infty} \mathbf{1}\Big\{\tau_{i,n+1} \leq T,\; \widehat{\mu}_i(n) \geq \theta'\Big\} + \sum_{t=m+1}^{T} \mathbf{1}\Big\{A_t = i,\; \widehat{\mu}_{i,t-1} < \theta',\; \mathcal{E}_t\Big\}.$$

Taking expectations, the first sum is bounded by

$$\sum_{n=1}^{\infty} \mathbb{P}\big(\widehat{\mu}_i(n) \geq \theta'\big).$$

For the second sum, note that $\mathcal{E}_t \in \mathcal{F}_{t-1}$. Hence

$$\mathbb{E}\big[\mathbf{1}\{A_t = i,\ \widehat{\mu}_{i,t-1} < \theta',\ \mathcal{E}_t\}\big] = \mathbb{E}\Big[\mathbf{1}\{\widehat{\mu}_{i,t-1} < \theta',\ \mathcal{E}_t\}\, \mathbb{P}(A_t = i \mid \mathcal{F}_{t-1})\Big].$$

On $\mathcal{E}_t \cap \{\widehat{\mu}_{i,t-1} < \theta'\}$, Lemma A.5 gives

$$\mathbb{P}(A_t = i \mid \mathcal{F}_{t-1}) \leq \frac{1}{r} \exp\big(-\eta_t(\theta - \theta')\big).$$

Summing over t yields (10). $\square$

A.6 A tail-moment lemma for integrating the spike bound

To convert (10) into Bayes regret, we need to integrate over μ the spike penalty. This is exactly where β-regularity yields logarithmic (for $\beta = 1$) or polynomial moment control.

Lemma A.8 (Tail-moment bound, general β)**.** *Let Γ be β-regular (Definition 3.1) and let $\mu \sim \Gamma$. Fix $\delta \in (0, \varepsilon_0/8]$ and set $U := 1 - \mu$. Then there is a constant $C < \infty$ (depending only on $\beta, c_0, C_0, \varepsilon_0$) such that*

$$\mathbb{E}\left[\mathbf{1}\{U \geq 4\delta\} \left(1 + \frac{1}{U - 3\delta}\right)\right] \ \leq\ C \cdot \mathfrak{M}_\beta(\delta),$$

where

$$\mathfrak{M}_\beta(\delta) := \begin{cases} 1 + \log(1/\delta), & \beta = 1, \\ 1, & \beta > 1, \\ \delta^{-(1-\beta)}, & \beta \in (0,1). \end{cases}$$

Proof. The proof is a straightforward integration-by-parts argument utilizing the β-regular bound $F(u) := \mathbb{P}(U \leq u) \leq C_0 u^\beta$ for $u \in (0, \varepsilon_0]$. Constants in the bounds below are allowed to depend on $(\beta, c_0, C_0, \varepsilon_0)$.

Writing

$$\mathbb{E}\left[\frac{1}{U - 3\delta}\mathbf{1}\{U \geq 4\delta\}\right] = \int_{4\delta}^{1} \frac{1}{u - 3\delta}\, dF(u),$$

integration by parts yields the boundary contribution

$$\left[\frac{F(u)}{u - 3\delta}\right]_{4\delta}^{1} = \frac{F(1)}{1 - 3\delta} - \frac{F(4\delta)}{\delta}.$$

Using the left-limit convention at the lower endpoint, the lower-boundary term is $-F((4\delta)^-)/\delta$, which is non-positive since the CDF F is non-negative; dropping it preserves the upper bound (and automatically absorbs any atom Γ may place at $u = 4\delta$). The boundary contribution is then at most $F(1)/(1 - 3\delta) = O(1)$. For the remaining integral, since $\delta \leq \varepsilon_0/8$ implies $u - 3\delta \geq \varepsilon_0/2$ on $[\varepsilon_0, 1]$, the contribution from $[\varepsilon_0, 1]$ is $O(1)$ (with constants depending on ε_0, using $F(u) \leq 1$). On $[4\delta, \varepsilon_0]$, the bound $u - 3\delta \geq u/4$ gives

$$\int_{4\delta}^{\varepsilon_0} \frac{F(u)}{(u - 3\delta)^2}\, du \ \lesssim\ \int_{4\delta}^{\varepsilon_0} \frac{u^\beta}{u^2}\, du = \int_{4\delta}^{\varepsilon_0} u^{\beta - 2}\, du,$$

which scales as $O(\log(1/\delta))$ for $\beta = 1$, $O(1)$ for $\beta > 1$, and $\Theta(\delta^{-(1-\beta)})$ for $\beta \in (0,1)$. Adding the constant 1 term from $\mathbb{E}[\mathbf{1}\{U \geq 4\delta\}] \leq 1$ completes the proof. $\square$

A.7 Order statistics: size of the subtraction term

The next lemma records the size of the subtraction term in (3). It is not needed for the upper bound in Theorem 5.2, but it quantifies how much the exact identity can sharpen constants.

Lemma A.9 (Expected gap to 1 of the best of m arms)**.** *If Γ is β-regular (Definition 3.1) and $\mu_* = \max_{i \le m} \mu_i$, then*

$$\mathbb{E}[1 - \mu_*] \;\le\; \frac{C}{m^{1/\beta}}$$

for a constant C depending only on $(\beta, c_0, C_0, \varepsilon_0)$. In particular, for $\beta = 1$, $\mathbb{E}[1 - \mu_*] \le C/m$.

Proof. Let $U_i := 1 - \mu_i$ and $U_* := \min_{i \le m} U_i = 1 - \mu_*$. By β-regularity, for small u we have $\mathbb{P}(U_i \le u) \ge c_0 u^\beta$, hence

$$\mathbb{P}(U_* > u) = \mathbb{P}(U_1 > u)^m \le (1 - c_0 u^\beta)^m \le \exp(-c_0 m u^\beta) \quad (u \le \varepsilon_0).$$

Integrating,

$$\mathbb{E}[U_*] = \int_0^\infty \mathbb{P}(U_* > u)\, du \le \int_0^{\varepsilon_0} e^{-c_0 m u^\beta}\, du + (1 - \varepsilon_0)\mathbb{P}(U_* > \varepsilon_0) \le C m^{-1/\beta},$$

using the change of variables $v = c_0 m u^\beta$. ☐

A.8 Proof of Theorem 5.2

Proof. Let

$$\theta := 1 - 2\delta, \qquad \theta' := 1 - 3\delta, \qquad p_\delta := \mathbb{P}\big(\mathcal{G}_i(\delta)\big), \qquad r := \max\Big\{1, \Big\lfloor \frac{m p_\delta}{2} \Big\rfloor\Big\}.$$

Define the event

$$\mathcal{E} := \big\{M(\delta) \ge r\big\}.$$

By Lemma A.4,

$$\mathbb{P}(\mathcal{E}^c) \;\le\; \exp(-c\, m\, \delta).$$

For each $t \in \{m+1, \dots, T\}$ let $\mathcal{E}_t$ be the event from Lemma A.7. If $\mathcal{E}$ occurs, then there are at least r arms j such that $\mathcal{G}_j(\delta)$ holds. For each such arm and each $t \ge m+1$,

$$\widehat{\mu}_{j,t-1} = \widehat{\mu}_j\big(N_j(t-1)\big) > \theta,$$

so $\mathcal{E} \subseteq \bigcap_{t=m+1}^T \mathcal{E}_t$. Hence Lemma A.7 implies that for every arm i,

$$\mathbb{E}\big[N_i(T)\mathbf{1}\{\mathcal{E}\}\big] \;\le\; 1 + \sum_{n=1}^\infty \mathbb{P}\big(\widehat{\mu}_i(n) \ge \theta'\big) + \frac{1}{r} \sum_{t=m+1}^T e^{-\eta_t \delta}. \tag{11}$$

Write $U_i := 1 - \mu_i$. Decompose the surrogate regret on $\mathcal{E}$ as

$$\widetilde{R}_T\, \mathbf{1}\{\mathcal{E}\} = \sum_{i=1}^m U_i N_i(T) \mathbf{1}\{\mathcal{E}\}\mathbf{1}\{U_i < 4\delta\} + \sum_{i=1}^m U_i N_i(T) \mathbf{1}\{\mathcal{E}\}\mathbf{1}\{U_i \ge 4\delta\}.$$

For the near-optimal arms,

$$\sum_{i=1}^m U_i N_i(T) \mathbf{1}\{\mathcal{E}\}\mathbf{1}\{U_i < 4\delta\} \le 4\delta \sum_{i=1}^m N_i(T) = 4\delta T,$$

hence

$$\mathbb{E}\Big[\sum_{i=1}^{m} U_i N_i(T)\mathbf{1}\{\mathcal{E}\}\mathbf{1}\{U_i < 4\delta\}\Big] \leq 4T\delta. \tag{12}$$

For the remaining arms, we redo the pathwise decomposition from the proof of Lemma A.7, this time keeping the factor $U_i\mathbf{1}\{U_i \geq 4\delta\}$. On $\mathcal{E} \subseteq \bigcap_{t=m+1}^{T} \mathcal{E}_t$, that proof gives the pathwise inequality

$$N_i(T)\,\mathbf{1}\{\mathcal{E}\} \;\leq\; 1 + \sum_{n=1}^{\infty} \mathbf{1}\{\widehat{\mu}_i(n) \geq \theta'\} + \sum_{t=m+1}^{T} \mathbf{1}\{A_t = i,\ \widehat{\mu}_{i,t-1} < \theta',\ \mathcal{E}_t\}.$$

Multiply both sides by the nonnegative quantity $U_i\mathbf{1}\{U_i \geq 4\delta\}$, sum over i, and take expectation. The first two pathwise terms give the spike part on the right-hand side of (13) below; these will be handled by conditioning on μ_i. For the leakage part (the last sum), use the pathwise bound $U_i\mathbf{1}\{U_i \geq 4\delta\} \leq 1$ *before* taking expectation, and then repeat the conditioning step from the proof of Lemma A.7: on $\mathcal{E}_t \cap \{\widehat{\mu}_{i,t-1} < \theta'\}$, Lemma A.5 bounds $\mathbb{P}(A_t = i \mid \mathcal{F}_{t-1}) \leq \frac{1}{r}e^{-\eta_t\delta}$. The result is

$$\mathbb{E}\Big[\sum_{i=1}^{m} U_i N_i(T)\mathbf{1}\{\mathcal{E}\}\mathbf{1}\{U_i \geq 4\delta\}\Big] \leq \sum_{i=1}^{m}\mathbb{E}\Big[\mathbf{1}\{U_i \geq 4\delta\}\,U_i\Big(1+\sum_{n=1}^{\infty}\mathbf{1}\{\widehat{\mu}_i(n) \geq \theta'\}\Big)\Big] + \frac{m}{r}\sum_{t=m+1}^{T} e^{-\eta_t\delta}. \tag{13}$$

Now condition on μ_i. Since the summands in the spike series are nonnegative, Tonelli's theorem justifies interchanging the expectation and the infinite sum. If $U_i \geq 4\delta$, then $\mu_i \leq 1 - 4\delta < \theta'$, so Lemma A.6 gives

$$\sum_{n=1}^{\infty}\mathbb{P}\big(\widehat{\mu}_i(n) \geq \theta' \mid \mu_i\big) \leq \frac{1}{2(\theta'-\mu_i)^2} = \frac{1}{2(U_i - 3\delta)^2}.$$

Therefore

$$\begin{aligned}\mathbb{E}\Big[\mathbf{1}\{U_i \geq 4\delta\}\,U_i\Big(1+\sum_{n=1}^{\infty}\mathbf{1}\{\widehat{\mu}_i(n) \geq \theta'\}\Big)\Big] &= \mathbb{E}\Big[\mathbf{1}\{U_i \geq 4\delta\}\,U_i\Big(1+\sum_{n=1}^{\infty}\mathbb{P}(\widehat{\mu}_i(n) \geq \theta' \mid \mu_i)\Big)\Big] \\ &\leq \mathbb{E}\Big[\mathbf{1}\{U_i \geq 4\delta\}\,U_i\Big(1+\frac{1}{2(U_i-3\delta)^2}\Big)\Big].\end{aligned}$$

On $\{U_i \geq 4\delta\}$ we have $U_i - 3\delta \geq \delta$ and $U_i \leq 4(U_i - 3\delta)$, hence

$$U_i\Big(1+\frac{1}{2(U_i-3\delta)^2}\Big) \leq 4(U_i-3\delta) + \frac{2}{U_i-3\delta} \leq 6\Big(1+\frac{1}{U_i-3\delta}\Big).$$

Applying Lemma A.8 with $\beta = 1$ yields

$$\mathbb{E}\Big[\mathbf{1}\{U_i \geq 4\delta\}\,U_i\Big(1+\sum_{n=1}^{\infty}\mathbf{1}\{\widehat{\mu}_i(n) \geq \theta'\}\Big)\Big] \leq C\big(1+\log(1/\delta)\big). \tag{14}$$

Summing (14) over i and combining with (13) gives

$$\mathbb{E}\Big[\sum_{i=1}^{m} U_i N_i(T)\mathbf{1}\{\mathcal{E}\}\mathbf{1}\{U_i \geq 4\delta\}\Big] \leq C\,m\big(1+\log(1/\delta)\big) + \frac{m}{r}\sum_{t=m+1}^{T} e^{-\eta_t\delta}. \tag{15}$$

By Lemma A.3, $p_\delta \geq c\delta$ for all sufficiently small δ. If $mp_\delta \geq 2$, then $r = \lfloor mp_\delta/2 \rfloor$ and therefore

$$\frac{m}{r} \leq \frac{4}{p_\delta} \leq \frac{C}{\delta}.$$

If $mp_\delta < 2$, then $r = 1$ and
$$m < \frac{2}{p_\delta} \le \frac{C}{\delta},$$
so again $m/r \le C/\delta$. Combining (12) and (15),
$$\mathbb{E}\big[\widetilde{R}_T\,\mathbf{1}\{\mathcal{E}\}\big] \le C\left[T\delta + m\big(1+\log(1/\delta)\big) + \frac{1}{\delta}\sum_{t=1}^{T} e^{-\eta_t\delta}\right],$$
where we enlarged the sum from $t = m+1$ to $t = 1$.

Finally,
$$\mathbb{E}[\widetilde{R}_T] = \mathbb{E}\big[\widetilde{R}_T\,\mathbf{1}\{\mathcal{E}\}\big] + \mathbb{E}\big[\widetilde{R}_T\,\mathbf{1}\{\mathcal{E}^c\}\big] \le \mathbb{E}\big[\widetilde{R}_T\,\mathbf{1}\{\mathcal{E}\}\big] + T\,\mathbb{P}(\mathcal{E}^c),$$
so
$$\mathbb{E}[\widetilde{R}_T] \le C\left[m + T\delta + m\big(1+\log(1/\delta)\big) + \frac{1}{\delta}\sum_{t=1}^{T} e^{-\eta_t\delta} + Te^{-cm\delta}\right].$$
Using (3), we have
$$\mathrm{BR}_{T,m}(\mathrm{ASG}) \le \mathbb{E}[\widetilde{R}_T],$$
which proves (7).

For the specialization (8), set $\delta = \min\{\delta_0, A\log(T\vee 2)/m\}$ with $A > 1/c$, and $\eta_t = (c_\eta/\delta)\log(t\vee 2)$ with $c_\eta > 1$. Then $\sum_{t=1}^{T} e^{-\eta_t\delta} = O(1)$.

If $A\log(T\vee 2)/m \le \delta_0$, then $\delta = A\log(T\vee 2)/m$ and $Te^{-cm\delta} = T^{1-cA} \le 1$ since $cA > 1$. The other terms in (7) are m, $T\delta = A\log(T\vee 2)\cdot T/m$, $m(1+\log(1/\delta)) = \tilde{O}(m)$, and $(1/\delta)\sum_t e^{-\eta_t\delta} = \tilde{O}(m)$, summing to $\tilde{O}(m + T/m)$.

If $A\log(T\vee 2)/m > \delta_0$ (i.e., $m < A\log(T\vee 2)/\delta_0$), the cap binds and $T/m > \delta_0 T/(A\log(T\vee 2)) = \tilde{\Omega}(T)$, so $\tilde{O}(m+T/m)$ already absorbs the trivial bound $\mathrm{BR}_{T,m}(\mathrm{ASG}) \le T$.

In both cases, $\mathrm{BR}_{T,m}(\mathrm{ASG}) = \tilde{O}(m + T/m)$. $\square$

B Simulation Experiments

We complement the theoretical analysis with simulations on Bernoulli bandits. The experiments are designed to (i) empirically validate the multiplicity mechanism of Theorem 5.2, (ii) probe the role of the sharpness (inverse temperature) parameter, and (iii) stress-test the assumptions underlying the Bayes regret guarantee of Theorem 5.2 by varying the informativeness and accuracy of the prior. Throughout, we compare four algorithm classes:

- **Classic Thompson Sampling (TS).** At each step, sample $\theta_i \sim \mathrm{Beta}(\alpha_0 + S_i, \beta_0 + F_i)$ independently for each arm and pull $\arg\max_i \theta_i$.
- **Pure Greedy (emirical-mean greedy).** Pull $\arg\max_i m_{i,t}$, corresponding to the baseline of Section 4.
- **Softmax empirical mean.** Sample arms with probability $\pi(i) \propto \exp(\eta \cdot m_{i,t})$, where $\eta > 0$ is a fixed sharpness parameter. This is a constant-temperature analogue of the ASG policy in (6).
- **Softmax + KL penalty.** A KL-regularized variant that anchors to a reference policy $\pi_0(i) \propto \exp(s\cdot m_{i,0})$ derived from the prior means, selecting arms via $\pi(i) \propto \pi_0(i)\exp(s\cdot m_{i,t})$. This models the KL-regularized objective structure common in GRPO-style updates (Section 1).

All experiments use Beta–Bernoulli conjugacy: each arm i has an unknown probability μ_i, rewards are $\mathrm{Bernoulli}(\mu_i)$, and the empirical after observing S_i successes and F_i failures in N_i pulls is $\mathrm{Beta}(\alpha_0 + S_i, \beta_0 + F_i)$. Results are averaged over independent trials (200 for Experiment 1, 100 for Experiments 2–4).

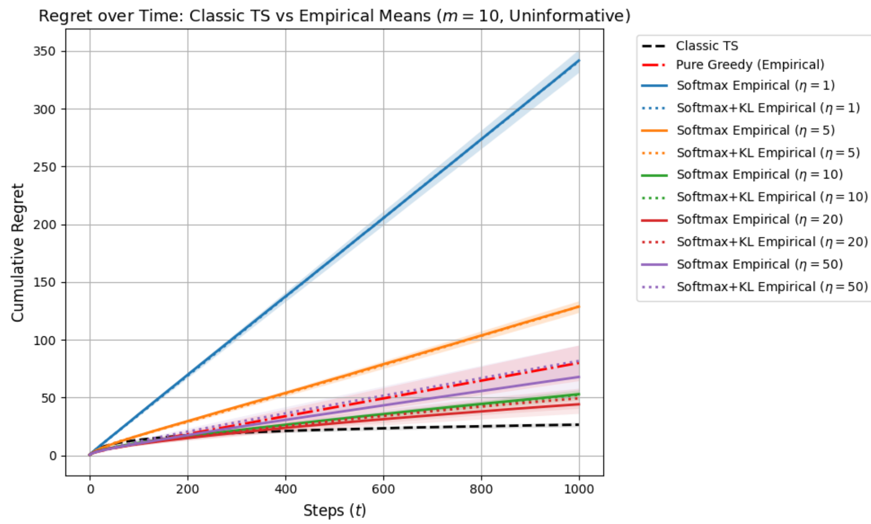

Figure 1: Cumulative regret over $T = 1{,}000$ steps with $K = 10$ arms and uninformative $\mathrm{Beta}(1, 1)$ prior. Higher sharpness s drives the Softmax policy toward greedy behavior. Under a uniform prior, Softmax+KL coincides with Softmax (overlapping curves).

B.1 Experiment 1: Effect of sharpness (uninformative prior)

We fix $K = 10$ arms, horizon $T = 1{,}000$, and an uninformative prior $\mathrm{Beta}(1, 1)$ (i.e. uniform on $[0, 1]$). We vary the sharpness parameter $\eta \in \{1, 5, 10, 20, 50\}$ for both the Softmax and Softmax+KL algorithms.

Figure 1 shows cumulative regret over time. At low sharpness ($\eta = 1$), the Softmax policy explores nearly uniformly and accumulates regret linearly, behaving essentially as a random policy. As η increases, the policy concentrates mass on arms with higher empirical means, and regret curves flatten toward the Greedy baseline. By $\eta = 50$, the Softmax algorithm is nearly indistinguishable from Pure Greedy. Classic TS achieves the lowest regret in this small-K regime, consistent with its known near-optimality for fixed, moderate K.

Under the uninformative prior, the Softmax+KL algorithm reduces to the standard Softmax algorithm (since π_0 is uniform), confirming that the KL penalty is inert when the prior carries no useful information. This is visible in the overlapping solid and dotted curves of the same color in Figure 1.

Theorem 5.2 requires $\eta_t \to \infty$ (i.e. increasing sharpness over time) to make the leakage term $\sum_t \exp(-\eta_t \delta)$ summable. Experiment 1 illustrates the static-sharpness analogue: higher η reduces leakage but with diminishing returns, and the transition from exploration-dominated to exploitation-dominated regret is smooth.

B.2 Experiment 2: Scaling with the number of arms (uninformative prior)

We vary $K \in \{10, 50, 100, 200, 500, 1{,}000\}$ with $T = 5{,}000$ and an uninformative $\mathrm{Beta}(1, 1)$ prior. Figure 2 reports the final cumulative regret R_T as a function of K.

The most striking feature is the *performance inversion* between Classic TS and the greedy-type algorithms as K grows. For small K, Classic TS dominates all other methods. However, as K increases past roughly $K \approx 200$, the Softmax algorithms with high sharpness ($\eta = 20, 50$) and Pure Greedy achieve lower regret than Classic TS, whose regret grows approximately linearly in K. In contrast, the Softmax ($\eta = 50$) and Greedy algorithms exhibit sublinear scaling, with regret plateauing or growing slowly.

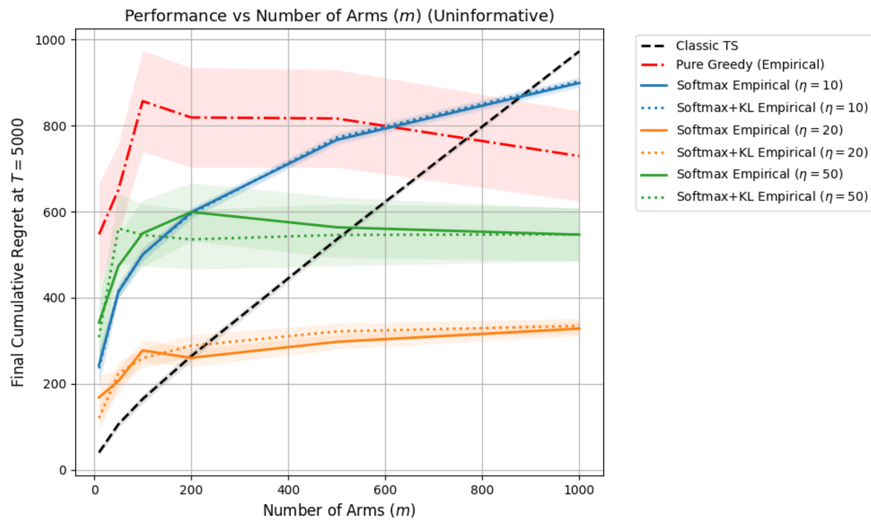

Figure 2: Final cumulative regret at $T = 5{,}000$ vs. number of arms K (uninformative prior). Greedy-type policies scale sublinearly, consistent with the $\tilde{O}(\sqrt{T})$ rate of Theorem 5.2 when $K = \Theta(\sqrt{T})$. Classic TS regret grows roughly linearly in K.

This is the empirical counterpart of the "many-armed" regime analyzed in the paper. With $K = 1{,}000$ i.i.d. $\mathrm{Uniform}[0, 1]$ arms, the best arm has $\mu_* \approx 1 - 1/K$ with high probability (Lemma A.9), and many arms cluster near μ_*. The β-regularity condition (Definition 3.1) holds with $\beta = 1$ for the uniform distribution, placing us squarely in the regime of Theorem 5.2. Classic TS wastes pulls exploring clearly suboptimal arms because its posterior sampling randomization scales with the full action space, whereas greedy and high-sharpness softmax policies exploit the multiplicity of near-optimal arms, precisely the mechanism formalized in Lemmas A.4 and A.5.

B.3 Experiment 3: Informative priors

We repeat the scaling experiment of Section B.2 but initialize the prior using information correlated with the true arm probabilities:

$$\alpha_{0,i} = 1 + c\,\mu_i, \qquad \beta_{0,i} = 1 + c\,(1 - \mu_i), \qquad c = 2.$$

This yields informative Beta priors whose means $\alpha_{0,i}/(\alpha_{0,i} + \beta_{0,i})$ approximate the true μ_i.

Figure 3 shows that informative priors dramatically improve all exploitative algorithms. Pure Greedy achieves near-zero regret across all values of K, since the prior already identifies the best arm with high accuracy. The high-sharpness Softmax algorithms ($s = 50$) and, notably, the Softmax+KL algorithms benefit substantially as well. The KL-regularized variant now *outperforms* standard Softmax at every sharpness level, because the reference policy π_0 encodes accurate prior information that the KL penalty preserves. In effect, the KL anchor "doubly exploits" the informative prior, compounding the prior-mean signal with the empirical-mean signal.

This experiment illustrates the role of the prior quality in the β-regular framework. When the prior accurately reflects the reward landscape, the "always-good" arms of Lemma A.4 are identified almost immediately.The KL-regularized variant, while not analyzed in our theoretical framework, suggests that anchoring to an informative prior can further suppress the leakage term, a direction we leave for future work.

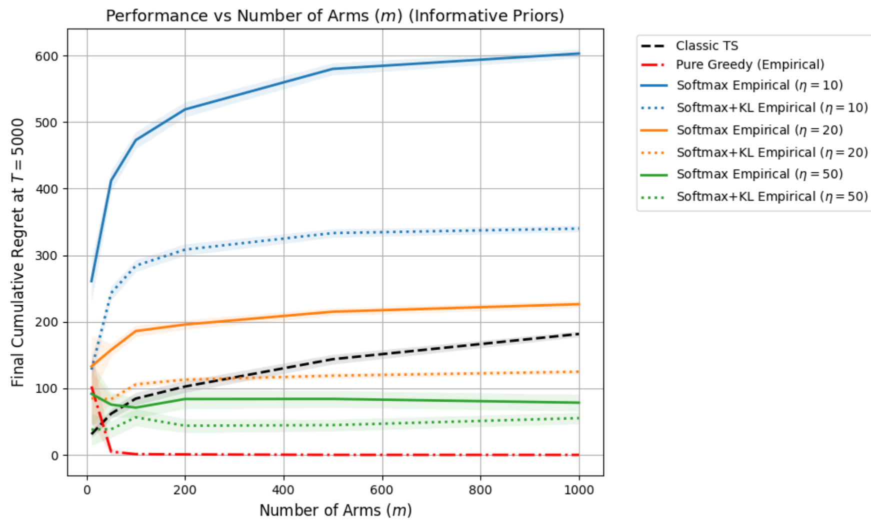

Figure 3: Final cumulative regret vs. K with informative priors ($c = 2$). Exploitative methods achieve dramatically lower regret; the KL-regularized Softmax benefits from double exploitation of the accurate prior.

B.4 Experiment 4: Misspecified priors

We stress-test the exploitative strategies by corrupting the informative prior with Gaussian noise ($\sigma = 0.1$) applied to the true probabilities before constructing the prior parameters. This misspecification means the prior's "best arm" is likely not the true best arm.

Figure 4 reveals a performance inversion relative to the informative-prior setting. Highly exploitative algorithms (Pure Greedy and Softmax with large η) suffer increased regret because they commit to the prior's (now incorrect) ranking. The Softmax+KL agent performs the worst among all methods at low sharpness, as the KL penalty actively discourages departure from the misspecified reference policy, impeding belief correction. Classic TS, while still paying a large exploration cost for high K, recovers more gracefully because empirical sampling inherently provides the variance needed to escape suboptimal lock-in.

This experiment probes the boundaries of our theoretical guarantees. Theorem 5.2 assumes a well-specified i.i.d. prior satisfying β-regularity, and the guarantee relies on many arms being genuinely near-optimal *a priori*. When the prior is misspecified, the empirical means are biased, and the always-good event $\mathcal{G}_i(\delta)$ (Equation 9) may fail to hold for the arms the policy concentrates on. The misspecified-prior experiment thus delineates the regime in which uncertainty-agnostic softmax policies are effective: they require that the prior (or the initial signal from the environment) provides a reasonable ordering of arm quality. This parallels the RLVR observation discussed in Section 1: softmax-style reweighting succeeds when the base policy already covers near-optimal completions, but cannot compensate for fundamental coverage failures.

B.5 Summary of empirical findings

The simulation experiments support three main conclusions aligned with the theoretical analysis:

1. **Softmax matches greedy without estimating uncertainty.** In the many-armed, uninformative-prior regime (Experiment 2), high-sharpness softmax and greedy policies outperform Classic TS as K grows, empirically confirming the mechanism of Theorem 5.2: the abundance of near-optimal arms under a β-regular prior makes explicit exploration unnecessary.

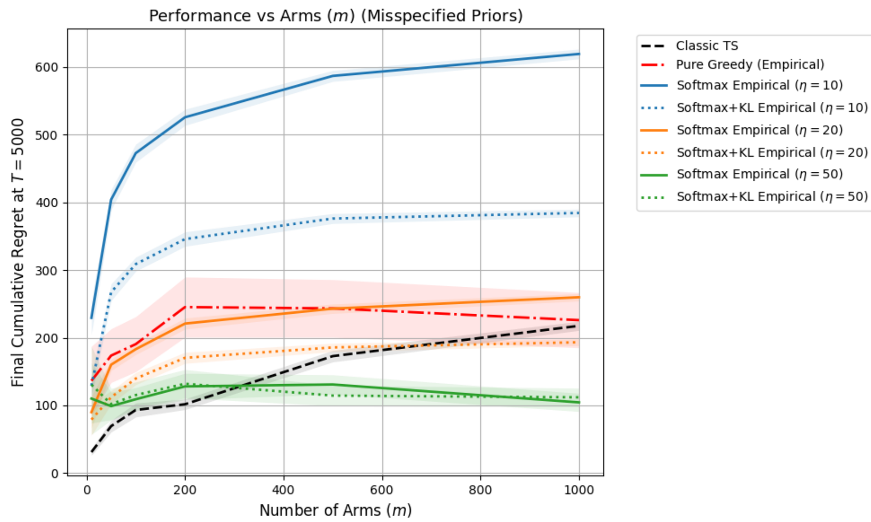

Figure 4: Final cumulative regret vs. K with misspecified priors ($\sigma = 0.1$ noise). Exploitative algorithms suffer from prior misspecification; the KL penalty amplifies this failure. Classic TS retains enough exploration to partially recover.

2. **Sharpness interpolates between exploration and exploitation.** Experiment 1 shows that increasing the inverse temperature η smoothly transitions the softmax policy from uniform exploration to greedy exploitation, corresponding to the leakage–exploitation tradeoff captured by the $\sum_t \exp(-\eta_t \delta)$ term in (7).
3. **Prior quality is the binding constraint.** Experiments 3 and 4 demonstrate that the effectiveness of uncertainty-agnostic policies hinges critically on prior accuracy. When the prior is well-specified, exploitative methods dominate; when misspecified, they fail catastrophically.

C Extended Related Work

This appendix expands on the condensed survey in Section 2, providing additional context and citations.

Stochastic bandits: optimism and posterior sampling. The modern theory of stochastic multi-armed bandits characterizes the exploration–exploitation trade-off through regret guarantees and instance-dependent lower bounds Lai & Robbins (1985); Auer et al. (2002); Audibert et al. (2007); Bubeck & Cesa-Bianchi (2012); Lattimore & Szepesvári (2020). A canonical Bayesian approach is posterior sampling (Thompson sampling), originally proposed in Thompson (1933) and analyzed in modern finite-time settings by, e.g., Agrawal & Goyal (2012); Russo & Van Roy (2016). Related Bayesian decision rules include Bayesian index policies Kaufmann (2018) and the classical Gittins index Gittins (1979). Our setting adopts the standard Beta–Bernoulli conjugate model, which enables a clean comparison between probability matching (posterior sampling) and softmax/Boltzmann action selection.

Boltzmann (softmax) exploration and its limitations. Boltzmann exploration (a.k.a. softmax or Gibbs action selection) is widely used in reinforcement learning and bandits as a simple randomized alternative to greedy choice. Despite its popularity, Cesa-Bianchi et al. (2017) show that, for stochastic K-armed bandits, *any monotone* temperature (learning-rate) schedule can be forced into suboptimal behavior: either it explores too long or it commits too early. They propose remedies including (i) tuned non-monotone schedules that require knowledge of the horizon and gaps and (ii) per-arm learning rates that explicitly track estimation uncertainty Cesa-Bianchi et al. (2017). This negative result motivates our focus on regimes where *uncertainty-aware* schedules may be unnecessary due to structural properties of the arm distribution.

Many-armed and infinite-armed bandits. A long line of work studies bandits with infinitely many arms and tail-based performance criteria, emphasizing how the distribution of arm qualities shapes achievable regret. Classical and modern examples include infinitely-many-armed formulations Berry et al. (1997); Wang et al. (2008); Bonald & Proutière (2013); Carpentier & Valko (2015); Chaudhuri & Kalyanakrishnan (2018). In the Bayesian many-armed regime, Bayati et al. (2020) formalize the idea that when the prior places substantial mass on near-optimal arms (via upper-tail regularity conditions), greedy-style policies can enjoy *free exploration*: discarding a poorly performing arm is likely to leave other near-optimal arms available. They show that subsampled greedy can achieve Bayesian regret scaling of order $\tilde{O}(\max\{m, T/m\})$ (up to prior-dependent exponents and logarithms), implying near-optimal performance with $m \asymp \sqrt{T}$ arms Bayati et al. (2020). Our analysis builds on this viewpoint, but focuses on softmax/Boltzmann policies that randomize *without* explicit epistemic-uncertainty bonuses.

Free exploration beyond non-contextual bandits. Complementary "free exploration" phenomena have been identified in contextual bandits, where exploration can be induced by natural diversity in contexts and data Bastani et al. (2020); Kannan et al. (2018); Raghavan et al. (2018); Hao et al. (2020); Abbasi-Yadkori et al. (2011). These works highlight that the need for explicit exploration is sensitive to structural assumptions (e.g., covariate diversity, smoothed analysis, or rich action sets). Our work studies an orthogonal mechanism: even *without* context diversity, the presence of many near-optimal arms (captured by prior tail regularity) can make epistemic-uncertainty-agnostic softmax policies achieve near-greedy Bayes regret.

Satisficing and multiple near-optimal actions. When near-optimal actions are plentiful, identifying the unique optimal action may be information-inefficient. This perspective is formalized in satisficing and information-theoretic approaches to bandits Russo & Van Roy (2014a;b; 2018). Conceptually, our "many near-optimal arms" regime is aligned with satisficing: regret can remain small even if the learner settles on an ε-optimal arm, provided such arms are sufficiently common under the prior.

RLVR and group-based policy optimization for LLM reasoning. The empirical motivation for this paper comes from the rapid adoption of reinforcement learning with verifiable (often binary) rewards for post-training large language models (LLMs) on reasoning-centric tasks. GRPO was introduced in DeepSeekMath as a memory-efficient alternative to PPO-style actor–critic training for verifiable rewards Shao et al. (2024); Schulman et al. (2017); it has since become a widely used baseline for RLVR training at scale DeepSeek-AI (2025); Yu et al. (2025). Several recent works analyze GRPO/RLVR training dynamics and relate them to KL-regularized reweighting. For example, Mroueh (2025) derive explicit optimal-policy forms for variants of GRPO under binary rewards and show how success probability can be amplified through iterative updates. At the same time, there is ongoing debate about whether RLVR expands a model's reasoning *support* (high-k coverage) or mainly reweights probability mass within the base model's existing support. Yue et al. (2025) report that RLVR often improves small-k performance while failing to improve—and sometimes degrading—large-k pass@k, suggesting a "bounded-by-base" effect; subsequent work revisits this question and argues that RLVR can extend reasoning boundaries under specific protocols and metrics Wen et al. (2025); Liu et al. (2025a). Additional analyses emphasize training instabilities, entropy collapse, and variance reduction techniques in RLVR pipelines Cui et al. (2025); Zeng et al. (2025); Liu et al. (2025b); Yu et al. (2025).

Pass@k as a tail metric and as an objective. Pass@k was popularized as a functional-correctness evaluation metric for code generation and is an order-statistic probe of a model's upper tail over solutions Chen et al. (2021). Recent RLVR-specific work argues that pass@k is best interpreted as a diagnostic of exploration/coverage rather than a direct optimization target Yu (2025), and proposes alternative objectives to mitigate probability concentration and improve pass@k at larger k Peng et al. (2025). Our bandit model adopts an explicit tail-regularity assumption on arm quality; this provides a stylized bridge to pass@k phenomena by translating "good pass@k" into "many near-optimal arms with non-negligible mass."

Connections to KL-regularized policy learning. A recurring theme in RLHF/RLVR is that KL regularization (to a reference or previous policy) acts as a trust-region or mirror-descent constraint, inducing stochastic policies that resemble Gibbs distributions Schulman et al. (2015; 2017); Ouyang et al. (2022). Very recent theory work in contextual bandits and RLHF argues that KL regularization alone can induce sufficient exploration under suitable coverage assumptions Zhao et al. (2025). Our contributions are complementary: we identify a distinct mechanism—prior tail mass / abundance of near-optimal arms—under which uncertainty-agnostic annealed softmax can achieve strong Bayesian regret, offering an alternative explanation for why soft reweighting can be effective even without explicit epistemic exploration.

Surveys and resources. For broader perspective on RL for large reasoning models, including RLVR training recipes, datasets, and open problems, see the recent survey Zhang et al. (2025). Reasoning Gym provides a large suite of procedurally generated, verifiable environments intended for RLVR research Stojanovski et al. (2025).