FINE-TUNING ON SELF-GENERATED AND REWARD-WEIGHTED DATA: LEARNING DYNAMICS, CONVERGENCE RATES, AND BENEFITS OF OFF-POLICYNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the learning dynamics of fine-tuning a policy model on self-generated and reward-weighted rollout data, with particular focus on a generalized version of classic REINFORCE — referred to as RE(S)— that updates the rollout distribution once every $S \geq 1$ gradient steps. Prior work in bandits and reinforcement learning has developed rich theory for REINFORCE and policy gradient methods, and on-policy sampling (i.e., a small value of S, ideally 1) is often viewed as crucial to their success; yet in prominent application like post-training large language models, reward-guided self-training has proved to be effective even when the rollout distribution is updated infrequently, but theoretical understanding remains limited for the learning dynamics and convergence properties of these off-policy methods. To bridge these gaps, we develop a unified theory for RE(S) that covers the full spectrum of $S \geq 1$: it can be interpreted as a stage-wise iterative optimization process, where each stage takes S gradient steps for minimizing the Kullback–Leibler distance to a fixed reward-weighted rollout distribution. For multi-arm bandits with softmax policies, our in-depth analysis and numerical experiments reveal three key findings: (1) for any fixed S, RE(S) enjoys global convergence to the optimal policy as the number of rollout distribution updates $B = \lfloor T/S \rfloor \to \infty$, where T denotes the total number of gradient steps; (2) we prove tight two-sided bounds for the convergence rate of RE(S), and show that its suboptimality gap achieves an asymptotic $\Theta(1/T)$ convergence rate, while S only affects the length of a burn-in phase; (3) when initialized at a weak policy with a small optimal-action probability, RE(1) gets trapped around suboptimal policies for a long period, whereas RE(S) with a suitable S avoids the detour and achieves significantly faster convergence to the global optimum, highlighting the benefits of off-policyness in this case.

1 INTRODUCTION

Multi-armed bandits and reinforcement learning (RL) have a long history (Lattimore & Szepesvári, 2020; Sutton & Barto, 1998) and have achieved substantial empirical success, including the recent use of RL in post-training large language models (LLMs) (Ouyang et al., 2022; Touvron et al., 2023; OpenAI, 2024; DeepSeek-AI, 2025; Kimi-Team, 2025). Many algorithms in these areas can be viewed as fine-tuning a policy model π_θ on self-generated and reward-weighted data. One prominent example is the classical REINFORCE algorithm (Williams, 1992) that, in the bandit setting, estimates the policy gradient by $(1/N)\sum_{i=1}^{N} r_i \nabla_\theta \log \pi_\theta(a_i)$, where $a_i \sim \pi_\theta$ is a sampled action and r_i is the reward for taking action a_i.

This perspective highlights a gap between two major lines of methodology, practice, and theory:

(1) In the bandit and RL literature, there exists a rich theory for policy gradient and policy iteration methods, including convergence and sample-complexity guarantees (Sutton et al., 1999; Agarwal et al., 2021; Mei et al., 2020b; 2023; Lu et al., 2024). Much of this theory assumes on-policy sampling, where the rollout distribution matches the current policy under optimization. Methods that account for (a limited degree of) off-policyness often require explicit algorithmic modifications, for example through trust-region regularization or clipping importance-sampling ratios (Schulman et al., 2015; 2017; Espeholt et al., 2018). In practice, keeping the rollout and training policies closely aligned is

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

also an important design consideration for these methods, motivating substantial efforts to support frequent and efficient synchronization of LLM model weights in RL infrastructure (Sheng et al., 2024; Hu et al., 2024; Pan et al., 2025).

(2) However, in the broader reward-guided self-training paradigm, it is also common to update the rollout distribution *infrequently* while fine-tuning the policy model on rollout samples weighted or filtered by their rewards, using standard log-likelihood objectives. Successful examples in the LLM domain include STaR (Zelikman et al., 2022), RAFT (Dong et al., 2023), ReST (Gulcehre et al., 2023), ReST^{EM} (Singh et al., 2024), rejection-sampling fine-tuning in Llama 2 (Touvron et al., 2023), AsymRE (Arnal et al., 2026), and more recent industrial practice. Such methods can be implemented with simple infrastructure: an LLM inference engine (Kwon et al., 2023; Zheng et al., 2024) generates rollout data and periodically loads an updated policy checkpoint, while a standard supervised fine-tuning (SFT) engine performs gradient updates on the reward-weighted or reward-filtered data. Despite their empirical success, theoretical understanding and guarantees for these off-policy methods remain limited.

These observations raise intriguing research questions. What are the connections and differences between these two lines of on-policy versus off-policy methodology? How does rollout staleness affect policy learning? What convergence guarantees can be established when the rollout distribution is synchronized with the trained policy infrequently? In particular, can off-policyness possibly lead to faster convergence than on-policy learning?

Contributions and outline. Towards answering these questions, we focus on a simple algorithm termed RE(S) and formalized in Algorithm 1, which is a generalized version of vanilla REINFORCE that updates the rollout distribution once every $S \geq 1$ gradient steps. The parameter S denotes the maximum staleness of a rollout sample with respect to the most updated policy model, and RE(1) reduces to standard on-policy REINFORCE. We show in Section 2 that RE(S) can be interpreted as a stage-wise iterative optimization process, where each stage effectively takes S gradient steps for minimizing the Kullback-Leibler (KL) distance to the reward-weighted rollout distribution. Unlike on-policy REINFORCE — which moves along the direction of standard policy gradient — our interpretation shows that RE(S) takes an alternative path towards the optimal policy.

In Section 3, we further focus on K-arm bandits and policies with softmax parameterization, for in-depth analysis and formal theoretical guarantees. Our key findings include the following:

1. Theorem 3.1: RE(S) enjoys global convergence to the optimal policy as the total number of gradient steps $T \to \infty$ regardless of S, as long as the number of rollout distribution updates $B = \lfloor T/S \rfloor$ also grows unbounded;
2. Theorems 3.3 and 3.4: We prove tight two-sided bounds showing that the suboptimality gap of RE(S) achieves an asymptotic $\Theta(1/T)$ convergence rate with respect to the number of gradient updates, while S only affects the length of a burn-in phase;
3. Theorem 3.6: When initialized at a weak policy with a small optimal-action probability x, RE(1) provably gets trapped around suboptimal policies for $\Omega(x^{-(K-1)})$ steps, whereas RE(S) with $S = \Theta(x^{-1})$ avoids such a detour and approaches the global optimum within $O(x^{-1}\log(1/x))$ gradient steps, highlighting the benefits of off-policyness in this case.

Discussion on prior literature and its connections to our work can be found in Section 4. Limitations of our work and directions for future research are discussed in Section 5. We hope that our theoretical study and numerical experiments can offer new insights for reward-guided self-training methodology, and inspire further research or practice in this area.

2 A HIGH-LEVEL UNDERSTANDING OF RE(S) LEARNING DYNAMICS

We formalize the problem setting as follows. Consider a multi-arm bandit with action space $\mathcal{A}$. Let θ denote the parameters of a policy, and π_θ denote its output distribution. Every time an action $a \in \mathcal{A}$ is sampled from the policy distribution $a \sim \pi_\theta$, it receives a (stochastic) reward $r \sim P_R(a)$ with expectation $\mathbb{E}[r] = \mu(a) \geq 0$. We assume non-negative reward means μ throughout this work. The mean reward of a policy π is denoted by $J(\pi) \coloneqq \mathbb{E}_{a\sim\pi}\mu(a)$.

The RE(S) algorithm is formalized in Algorithm 1. It is a double-loop process: in the b-th outer iteration, a total of $N \times S$ action-reward pairs are sampled from the same rollout distribution $\pi_{\theta_{b,0}}$; then, the inner loop takes S gradient steps of reward-weighted fine-tuning, each consuming N rollout samples. It is obvious that the special case $S = 1$ reduces to standard on-policy REINFORCE.

Algorithm 1 RE(S), a general version

Input: initial policy parameters $\theta_{0,0}$, staleness $S \in \mathbb{N}_+$, learning rate $\eta > 0$.
for $b = 0, 1, \ldots, B-1$ **do**
  Sample $a_{s,i} \sim \pi_{\theta_{b,0}}$ and its reward $r_{s,i} \sim P_R(a_{s,i})$, $0 \le s < S$, $1 \le i \le N$.
  for $s = 0, 1, \ldots, S-1$ **do**
    $\theta_{b,s+1} \leftarrow \theta_{b,s} + \frac{\eta}{N} \sum_{i=1}^{N} r_{s,i} \nabla_\theta \log \pi_\theta(a_{s,i})|_{\theta=\theta_{b,s}}$.
  end
  Set $\theta_{b+1,0} \leftarrow \theta_{b,S}$
end

While standard policy gradient theory says $\mathbb{E}_{a\sim\pi_\theta}[\mu(a)\nabla_\theta \log \pi_\theta(a)] = \nabla_\theta J(\pi_\theta)$, such unbiasedness is clearly not true for the gradient estimate in RE(S) when off-policyness occurs, namely $S > 1$. Fortunately, we can show that RE(S) takes an alternative path towards the optimal policy. Indeed, RE(S) can be viewed as a stage-wise optimization process, where each stage takes S gradient-descent steps for minimizing the KL distance to the reward-weighted rollout distribution $\tilde{\pi}_b^\mu$:

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi_{\theta_{b,0}}, r\sim P_R(a)}[r \cdot \nabla_\theta \log \pi_\theta(a)] = \mathbb{E}_{a\sim\pi_{\theta_{b,0}}}[\mu(a) \cdot \nabla_\theta \log \pi_\theta(a)] \\
&\quad = Z_b \cdot \int_{a\in\mathcal{A}} \frac{\pi_{\theta_{b,0}}(a)\mu(a)}{Z_b} \nabla_\theta \log \pi_\theta(a)\, \mathrm{d}a = Z_b \cdot \mathbb{E}_{a\sim\tilde{\pi}_b^\mu} \nabla_\theta \log \pi_\theta(a) \\
&\quad = -Z_b \cdot \nabla_\theta \, \mathrm{D}_{\mathrm{KL}}\Big(\tilde{\pi}_b^\mu \,\|\, \pi_\theta\Big),
\end{aligned}$$

where $\tilde{\pi}_b^\mu$ and Z_b are defined by

$$\tilde{\pi}_b^\mu(a) := \frac{\pi_{\theta_{b,0}}(a)\mu(a)}{Z_b}, \quad Z_b := \int_{a\in\mathcal{A}} \pi_{\theta_{b,0}}(a)\mu(a)\, \mathrm{d}a = \mathbb{E}_{a\sim\pi_{\theta_{b,0}}} \mu(a).$$

If S is sufficiently large, then by the end of the inner loop, $\pi_{\theta_{b+1,0}} = \pi_{\theta_{b,S}}$ will ideally get close to $\tilde{\pi}_b^\mu$, which is a higher-reward policy than $\pi_{\theta_{b,0}}$. Thus RE(S) could converge to the optimal policy as the rollout distribution gets updated for $B \to \infty$ times. In the next section, we formalize this intuition and provide in-depth convergence analysis in concrete settings.

3 In-depth analysis for bandits and softmax policies

In this section, we present in-depth analysis for the learning dynamics and convergence rates of RE(S) in bandits with softmax policies. Our key findings include global convergence of RE(S) to the optimal policy (Section 3.1), two-sided bounds for convergence rates that reveal two distinct time scales in the learning dynamics of RE(S) (Section 3.2), and the benefits of off-policyness in accelerating convergence under certain conditions (Section 3.3). We conclude with numerical validation of these theoretical findings (Section 3.4).

Problem setup. Following the formulation in Section 2, we further assume a discrete action space $\mathcal{A} = [K] := \{1, \ldots, K\}$, and a reward-mean vector $\mu = [\mu(1), \ldots, \mu(K)]$. Assume without loss of generality that action 1 is the unique optimal action. We let $\mu_{\max}/\mu_{\min}$ and Δ denote the maximum/minimum reward and the reward gap between the top-2 actions, respectively:

$$\mu_{\max} := \mu(1) > \mu(a) \ge \mu_{\min} > 0, \quad 2 \le a \le K; \qquad \Delta := \mu_{\max} - \max_{2\le a\le K} \mu(a) > 0.$$

We focus on policies with softmax parameterization throughout this section:

$$\theta = \big[\theta(1), \ldots, \theta(K)\big] \in \mathbb{R}^K, \quad \pi_\theta(a) = \frac{e^{\theta(a)}}{\sum_{j=1}^K e^{\theta(j)}}, \quad 1 \le a \le K.$$

Suppose that every coordinate of the initial logit vector θ is finite, hence the initial policy has full support over $\mathcal{A}$. The learner aims to find the policy π_θ that maximizes the objective function:

$$J(\theta) = \mathbb{E}_{a\sim\pi_\theta}\mu(a) = \langle \pi_\theta, \mu \rangle. \tag{1}$$

For notational simplicity, we will use $\pi_{b,s}$ and $\pi_{\theta_{b,s}}$ interchangeably. In this concrete setting, the infinite-sample limit of the gradient estimate in Algorithm 1 can be calculated by

$$g_{b,s} \coloneqq \pi_{b,0} \odot \mu - J(\pi_{b,0})\pi_{b,s},$$

where $\odot$ denotes element-wise multiplication; see Lemma A.1 in the appendix for the derivation. This leads us to the specialized version of RE(S) in Algorithm 2, which we investigate in this section.

Algorithm 2 RE(S), softmax parameterization and infinite-sample limit

Input: initial logits $\theta_{0,0} \in \mathbb{R}^K$, staleness $S \geq 1$, learning rate $\eta > 0$.
for $b = 0, 1, \ldots, B-1$ **do**
  for $s = 0, 1, \ldots, S-1$ **do**
    $\theta_{b,s+1} \leftarrow \theta_{b,s} + \eta\,(\pi_{b,0} \odot \mu - J(\pi_{b,0})\pi_{b,s})$.
  end
  Set $\theta_{b+1,0} \leftarrow \theta_{b,S}$.
end

3.1 Global convergence to the optimal policy

The following theorem guarantees global convergence of RE(S) to the optimal policy $\pi^\star = e_1$, namely probability 1 on the optimal action and 0 elsewhere.

Theorem 3.1. *Suppose that $\theta_{0,0} \in \mathbb{R}^K$ is finite, and the learning rate $\eta > 0$ satisfies $\eta \cdot \mu_{max} < 4$. Then for any fixed integer $S \geq 1$, RE(S) converges to the optimal policy asymptotically:*

$$\pi_{b,s} \to e_1 \quad \textit{for all} \quad 0 \leq s \leq S \quad \textit{as} \quad b \to \infty.$$

Proof sketch. The proof combines improvement of the surrogate objective with bounds on changes in the logits. The step-size condition and smoothness ensure that $J_b \coloneqq J(\pi_{b,0})$ is nondecreasing in b, so it converges to some $J_\infty \leq \mu(1)$. The remaining work is to rule out convergence to a suboptimal action. To prove this by contradiction, suppose that $J_\infty < \mu(1)$. For fixed S, logit changes during each stage tend to zero. Actions with rewards above J_∞ eventually have increasing logits, and those below it have decreasing logits. The estimate $\|g_{b,0}\|_2 \leq C(J_\infty - J_b)$ implies $\sum_b (J_\infty - J_b) = \infty$, since a finite sum would make all logits converge to finite values and prevent the optimal action's gradient from tending to zero. Keeping $J_b \leq J_\infty$ therefore requires an action a with $\mu(a) < J_\infty$ and $\sum_b \pi_{b,0}(a) = \infty$, therefore its logit tends to $-\infty$. Since the optimal logit is bounded below, the optimal action eventually has greater probability, which implies $\sum_b \pi_{b,0}(1) = \infty$ and its logit tends to $+\infty$. The probability of every action with $\mu(a) < J_\infty$, divided by the optimal action's probability, then tends to zero. These actions can no longer offset the optimal action's positive contribution to $J_b - J_\infty$, giving $J_b > J_\infty$, which is a contradiction. Therefore J_∞ must be $\mu(1) = \mu_{\max}$, which means convergence to the optimal policy. See Appendix A for the full proof. □

3.2 Two-sided bounds for convergence rates

We next characterize the convergence rate of RE(S) and its dependence on S. Denote the suboptimality gaps of policies by

$$D(\pi) \coloneqq \mu_{\max} - J(\pi), \qquad d_b \coloneqq D(\pi_{b,0}),$$

and define the surrogate loss function L as

$$L(\pi_{b,s}; \pi_{b,0}) \coloneqq \mathbb{E}_{a\sim\pi_{b,0}}\mu(a) \log \pi_{b,s}(a).$$

Our analysis critically relies on the following lemma that quantifies progress within each inner loop.

Lemma 3.2 (Progress and suboptimality gaps within an inner loop). *Consider the b-th outer iteration of Algorithm 2. For every $1 \le s \le S$, we have*

$$\lambda\big(\pi_{b,0}(1)\big)\min\{\eta s d_b^2, d_b\} \le L(\pi_{b,s};\pi_{b,0}) - L(\pi_{b,0};\pi_{b,0}) \le J(\pi_{b,s}) - J(\pi_{b,0}) \le A\min\{\eta s d_b^2, d_b\}, \tag{2}$$

$$\text{and} \quad \rho d_b \le D(\pi_{b,s}) \le d_b, \tag{3}$$

where the specific forms of $\lambda(\cdot)$, ρ and A can be found in Appendix B.

Proof sketch. We focus on explaining the proof idea of the most important lower bound in Eq. (2), and proof of the other results can be found in Appendix B. It is known that for on-policy softmax policy gradient, smoothness and the non-uniform Łojasiewicz inequality give an $\Omega(\eta d_b^2)$ improvement per step (Mei et al., 2020b). For fixed inner-loop step s, when $\eta s d_b \le 1$, a small suboptimality gap d_b limits how far the current policy can move over the inner loop. Therefore, the accumulated progress is $\Omega(s\eta d_b^2)$ in this inner loop. This also explains the on-policy order of progress near convergence. But when $\eta s d_b > 1$, the gradient direction can be very different from on-policy gradient during the inner loop. Nonetheless, we can still show that the current policy remains close enough to the on-policy gradient direction for about $1/(\eta d_b)$ steps. Combining this with the on-policy result, we see that the progress in this inner loop is at least $\Omega(\eta d_b^2 \cdot 1/(\eta d_b)) = \Omega(d_b)$. □

Eq. (2) in the above lemma suggests that $J(\pi_{b,s}) - J(\pi_{b,0}) = \Theta\big(\min\{\eta s d_b^2, d_b\}\big)$. With the progress of the optimization, decreasing d_b changes this scale from $\Theta(d_b)$ to $\Theta(\eta s d_b^2)$. This change essentially causes the two time scales that we will soon describe in the main theorems below. The reason behind the lower bound in Eq. (3) is that, increasing the likelihood of the reward-weighted rollout distribution prevents an arbitrarily large reduction of the suboptimality gap within the same inner loop.

Building on Lemma 3.2, we can now establish upper and lower bounds for the convergence rates of RE(S) in Theorems 3.3 and 3.4 respectively. We introduce a new notation $p_{\min} := \inf_{b\ge 0} \pi_{b,0}(1) > 0$, whose non-negativity[1] originates from Theorem 3.1. In the theorems below, we use notation $t = bS + s$ to denote the total number of gradient steps for achieving $\pi_{b,s}$ in RE(S).

Theorem 3.3 (Upper Bound). *Suppose that $\eta \cdot \mu_{max} < 4$. There exist constant $c > 0$ (depending only on K, μ, η, and $p_{\min}$) such that, if we define*

$$\bar{b} = \begin{cases} 0 & \text{if} \quad \eta S d_0 \le 1, \\ \left\lceil \dfrac{\log(\eta S d_0)}{\log(1+c)} \right\rceil & \text{if} \quad \eta S d_0 > 1, \end{cases} \tag{4}$$

then for $0 \le b \le \bar{b}$, we have

$$D(\pi_{b,s}) \le d_0(1+c)^{-b}, \tag{5}$$

whereas for every $b \ge \bar{b}$, we have

$$D(\pi_{b,s}) \le \frac{1}{d_{\bar{b}}^{-1} + c\eta(t-\bar{t})}, \quad \text{where} \quad t = bS + s, \quad \bar{t} = \bar{b}S. \tag{6}$$

Theorem 3.4 (Lower Bound). *Suppose that $\eta \cdot \mu_{max} < 4$. Let A and $\rho \in (0,1)$ be the constants in Lemma 3.2, and define*

$$\bar{b}' = \begin{cases} 0 & \text{if} \quad \eta S d_0 \le (1-\rho)/A, \\ \left\lceil \dfrac{\log(A\eta S d_0/(1-\rho))}{\log(1/\rho)} \right\rceil & \text{if} \quad \eta S d_0 > (1-\rho)/A, \end{cases} \tag{7}$$

then for $0 \le b \le \bar{b}'$, we have

$$D(\pi_{b,S}) \ge d_0 \rho^{b+1}, \tag{8}$$

[1] For an arbitrary finite initialization, $p_{\min} > 0$ but its value depends on the optimization trajectory and could be very small (Mei et al., 2020b). Fortunately, if we impose stronger conditions on the initial policy, the lower bound of $p_{\min}$ can be significantly improved. For example, if η satisfies $0 < \eta\mu_{\max} \le 2$ and $\pi_{0,0}(1) \ge \pi_{0,0}(a)$ for every $a > 1$, then this ordering will be preserved at every inner iterate, as shown in Appendix B.3. Hence $p_{\min} \ge 1/K$ in this case, and the progress coefficient can be bounded below by $\lambda(1/K)$.

whereas for every $b \geq \bar{b}'$, we have

$$D(\pi_{b,s}) \geq \frac{1}{d_0^{-1}\rho^{-\bar{b}'} + (A/\rho)\eta(t - \bar{t}')}, \quad \textit{where} \quad t = bS + s, \quad \bar{t}' = \bar{b}'S. \tag{9}$$

Proof sketch. Using Lemma 3.2 and dividing Eq. (2) therein by $d_b D(\pi_{b,s})$, we have

$$\lambda(p_{\min}) \min\{\eta s, \frac{1}{d_b}\} \leq \lambda\big(\pi_{b,0}(1)\big) \min\{\eta s, \frac{1}{d_b}\} \leq \frac{1}{D(\pi_{b,s})} - \frac{1}{d_b} \leq \min\left\{\frac{A}{\rho}\eta s, \frac{1-\rho}{\rho}\frac{1}{d_b}\right\}.$$

Denote $a_b \coloneqq 1/d_b$. For $s = S$, replacing each bound by equality gives two sequences of the form $a_{b+1} = a_b + \min\{x\eta S, y a_b\}$, starting at $a_0 = 1/d_0$. The coefficients x and y come from the corresponding side of the inequality. This update is increasing in a_b, so the two sequences remain lower and upper bounds on $1/d_b$ at every inner loop. When $y a_b < x\eta S$, the update is $a_{b+1} = (1+y)a_b$. Once $y a_b \geq x\eta S$, it becomes $a_{b+1} = a_b + x\eta S$ and remains in this case. See Appendix B.2 for the complete proof. □

Remark 3.5. Several previous works studied similar problems but with different theoretical techniques. Mei et al. (2020b) lower bound improvement by a multiple of the squared gap using smoothness and a non-uniform Łojasiewicz inequality; such a bound is not guaranteed to hold for every step of an inner loop in the RE(S) algorithm that we study, since its gradient update $g = q \odot \mu - J(q)\pi$ (where q denotes the rollout distribution) can vanish. Liu et al. (2024) use the nonnegativity of $(u(a) - u(a'))(e^{\eta u(a)} - e^{\eta u(a')})$, where $u = \nabla_\theta J$; for RE(S), the corresponding terms are $(u(a) - u(a'))(e^{\eta g(a)} - e^{\eta g(a')})$, and these two factors need not have the same sign. Some works consider stage-wise algorithms similar to RE(S), where each inner loop uses a fixed rollout distribution q: Arnal et al. (2026) identify the inner-loop convergence limit through a Lyapunov function, and then iterate this limiting map across outer iterations, while Štrupl et al. (2022) use exact weighted-likelihood maximization, nonnegative action-value variance, and probability ratios to prove improvement and convergence. For bandits with positive rewards, both yield the updated rollout distribution $q^+ = q \odot \mu / J(q)$, but RE(S) with a finite S value need not obey this identity. Adapting their arguments to RE(S) would require additional control of how incomplete fitting within a stage changes these ratios. Our proposed analysis instead estimates improvement over a finite-step stage in terms of the current suboptimality gap, including cases when the gap approaches zero and when it does not.

3.3 Benefits of Off-Policyness

While the previous theorems present $\Theta(1/t)$ asymptotic convergence rates regardless of the value of S, there exist other terms in the bounds that can play a crucial role in the overall speed of convergence. In the following, we identify settings where, perhaps surprisingly, RE(S) with a suitably large S can achieve much faster convergence than on-policy RE(1).

Consider the following notations and assumptions:

$$\mu_{\max} = \mu(1) > \mu(2) > \mu(a) \geq \mu_{\min} > 0, \quad 3 \leq a \leq K; \qquad \Delta = \mu(1) - \mu(2) > 0.$$

Consider fixed weights $w(a) > 0$ for $2 \leq a \leq K$ with $\sum_{a=2}^{K} w(a) = 1$, and initialize RE(S) with the following policy $\pi_{0,0}$:

$$\pi_{0,0}(1) = x > 0, \quad \pi_{0,0}(a) = (1-x)w(a), \quad 2 \leq a \leq K. \tag{10}$$

Now we consider a family of initial conditions specified by the optimal-action probability x, while fixing other factors like K, μ, w, and η satisfying $0 < \eta\mu_{\max} < 4$. For any target accuracy $0 < \epsilon < \Delta/2$, we define $T_\epsilon(S)$ as the number of gradient steps required by RE(S) to achieve it:

$$T_\epsilon(S) \coloneqq \inf\{bS : b \geq 0,\ D(\pi_{b,0}) \leq \epsilon\}. \tag{11}$$

Focusing on asymptotic $x \to 0$, the theorem below demonstrates provable benefits of off-policyness.

Theorem 3.6. *Under the conditions stated above, we have*

$$T_\epsilon(1) = \Omega\Big(x^{-(K-1)}\Big).$$

Moreover, there exists integer $S_x = \Theta(x^{-1})$, for which

$$T_\epsilon(S_x) = O\big(x^{-1}\log(1/x)\big), \quad \textit{and thus} \quad \frac{T_\epsilon(S_x)}{T_\epsilon(1)} \longrightarrow 0 \quad \textit{as} \quad x \to 0.$$

The implicit constants in these bounds may depend on $K, \mu, w, \eta, \epsilon$, but not on x.

We provide an intuitive explanation below. On-policy policy gradient follows the direction of steepest first-order increase in expected reward, yet this local improvement might substantially reduce the optimal action's probability. Once trapped near a suboptimal vertex of the probability simplex, the small probability of the optimal action could suppresses its gradient, making recovery very slow. This slow transient behavior of softmax policy gradient has also been established theoretically and observed empirically by previous works (Mei et al., 2020a; Mei & Osband, 2026). In contrast, RE(S) fixes the rollout distribution within each stage and takes gradient steps to reduce the forward KL divergence from the reward-weighted target to the current policy. Reward weighting assigns the largest multiplicative weight to the optimal action, so once the KL divergence is sufficiently small, its probability increases relative to its value at the start of the stage. Choosing a suitable S for sufficiently accurate fitting therefore ensures progress in the optimal action's probability after each inner loop. This property holds for every full-support policy, including policies near a suboptimal vertex. Therefore, even if RE(S) has slower progress than on-policy RE(1) early in training, it can maintain steady progress in increasing the probability of the optimal action, avoiding the stagnation that may occur with RE(1) and thus achieving faster convergence eventually.

Proof sketch. For RE(1), analysis of inverse-logit coordinates reveal an interval in which the $K-2$ inferior actions are suppressed while the optimal-action logit remains near its initial level. Conservation of the logit sum then concentrates probability on action 2 and reduces the optimal-action probability to $O(x^{K-1})$. Controlling the time spent in this region yields the $\Omega(x^{-(K-1)})$ lower bound. For RE(S), each reward-weighted target increases the optimal action's log-odds by a fixed positive amount. Its logits differ from the stage's initial logits only by reward-dependent shifts, and standard optimization analysis gives an $O(1/S)$ KL fitting error bound for a S-step stage, with a constant term independent of x. A binary-KL bound shows that an error below a sufficiently small multiple of x preserves a fixed fraction of this log-odds gain. Thus $S_x = \Theta(x^{-1})$ suffices to ensure that the optimal-action probability at the end of a stage remains at least x throughout. Accumulating the fixed log-odds gains takes $O(\log(1/x))$ stages, leading to $O(x^{-1}\log(1/x))$ total gradient steps. See Appendix C for the complete proof. □

3.4 Numerical validation

We first conduct experiments to verify the theoretical findings in Theorems 3.3 and 3.4, namely the convergence rates of RE(S). The empirical results are shown in Figure 1, whose concrete settings are explained in the caption. In the first setting, we observe that when initialized at a strong policy, RE(S) converges to the optimal policy at a $\Theta(1/t)$ rate (where t is the number of gradient steps) after a short burn-in phase; in particular, the learning curves of RE(S) with different S values mostly overlap when using t as the X-axis. In the second setting where the initial policy is relatively weaker (though still dominated by the optimal action) and the learning rate η is larger, RE(S) with a larger S exhibits a longer burn-in phase and initially lags behind, but eventually catches up with on-policy RE(1) and achieves the same $\Theta(1/t)$ convergence rate, as predicted by our main theorems.

We further conduct experiments to verify Theorem 3.6, namely the benefits of off-policyness in accelerating the convergence of RE(S) when the initial policy has a small optimal-action probability. The empirical results are shown in Figure 2, whose concrete settings are explained in the caption. In the first plot, we observe that the suboptimality gap of on-policy RE(1) decays quickly at the beginning, but then plateaus at 0.3 — the suboptimality gap of the second-best action — for a long period; in contrast, RE(S) with a large $S = 512$ makes steady progress towards the global optimum and eventually reaches a suboptimality gap close to zero. The second plot visualizes the optimization trajectories of RE(S) with different S values in a minimal 3-action setting; it clearly demonstrates how RE(1) takes a long detour in its trajectory towards the optimal policy, whereas RE(S) with larger S avoids the trap and achieves faster convergence to the global optimum.

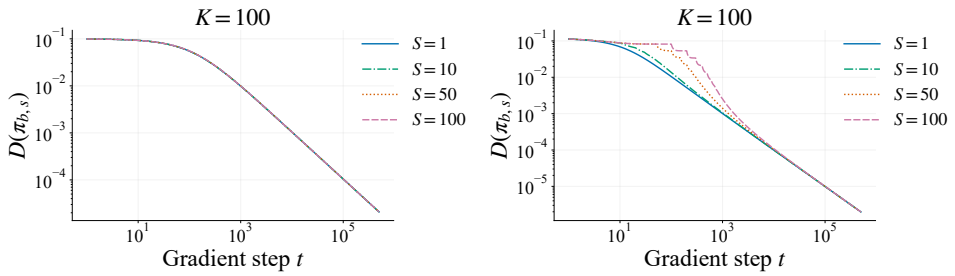

Figure 1: Empirical validation of RE(S) convergence rates in two concrete settings, both with $K = 100$. **Left:** the reward-mean vector is $\mu = [1, 0, \ldots, 0]$, the initial policy $\pi_{0,0}$ has probability 0.9 on the optimal action and equal probabilities on the remaining actions, and the learning rate $\eta = 0.095$. **Right:** $\mu = [1, 0.6, \ldots, 0.6]$, $\pi_{0,0}$ has probability 0.7 on the optimal action and equal probabilities on the remaining actions, and $\eta = 1$.

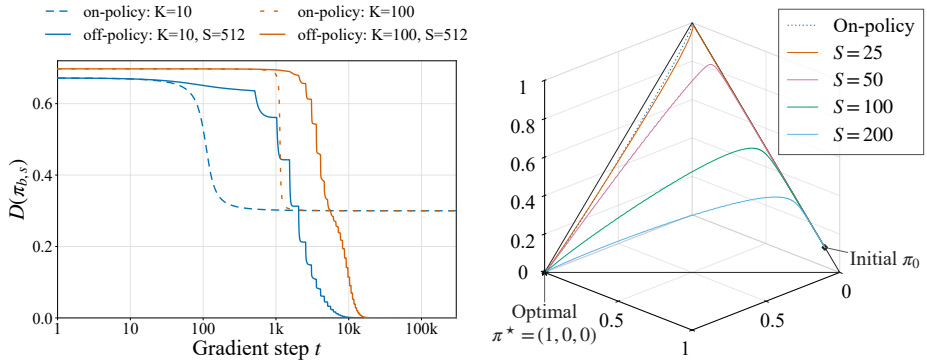

Figure 2: Empirical validation of the benefits of off-policyness. **Left:** $K \in \{10, 100\}$, the reward-mean vector is $\mu = [1, 0.7, 0.3, 0.3, \ldots]$, the initial policy $\pi_{0,0}$ satisfies $\pi_{0,0}(1) = 0.1/K$ and $\pi_{0,0}(a) \propto e^{-2\mu(a)}$ for $2 \leq a \leq K$. **Right:** a minimal setting for visualizing the optimization trajectories of RE(S), with $K = 3$, $\mu = [1, 0.9, 0.89]$, and $\pi_{0,0} = [0.0005, 0.1, 0.8995]$.

4 Related Work

Reward-guided self-training. Reward-guided self-training has been used to improve LLM reasoning and alignment by training on model-generated outputs that are weighted or selected based on task feedback. STaR learns from rationales that yield correct answers, RAFT selects outputs for training by reward ranking, and ReST-EM uses binary feedback in an expectation-maximization procedure (Zelikman et al., 2022; Dong et al., 2023; Singh et al., 2024). Ghosh et al. (2020) derive a policy-improvement bound for reward-weighted likelihood fitting that does not require exact optimization. In tabular settings, reward-weighted regression and zero-baseline AsymRE — both of which have a stage-wise structure akin to RE(S) — have been shown to converge globally from any full-support initial policy when the surrogate objective within each stage is fitted exactly or to convergence (Štrupl et al., 2022; Arnal et al., 2026), but their results do not cover finite-step fitting; in comparison, our convergence results in this work cover the full range $S \geq 1$. Methods with finite-step inner loops include iw-SFT, which periodically reweights a fixed curated dataset and yields a reward-weighted surrogate loss before clipping or smoothing operations (Qin & Springenberg, 2025). Russo (2026) formulates finite-step fitting of success-conditioned targets, but does not establish convergence of the resulting iterative procedure. All these works do not establish global convergence or rates measured in total gradient steps when each stage performs a fixed finite number S of updates.

Softmax policy gradient. Global convergence and $O(1/T)$ rates have been established for exact softmax policy gradient (Mei et al., 2020b; Liu et al., 2024; Lu et al., 2024); our results in Theorems 3.1, 3.3 and 3.4 extend these conclusions to RE(S) with $S \geq 1$. Almost-sure convergence has

also been established for stochastic REINFORCE in bandits and finite-horizon tabular MDPs (Mei et al., 2023; Robertson et al., 2025). Lower bounds nevertheless show slow escape from unfavorable initialization in bandits and exponential iteration complexity on particular discounted MDPs (Mei et al., 2020a; Li et al., 2023). Mei et al. (2020a) address this issue by changing the policy parameterization, while Mei & Osband (2026) use gradient gating to accelerate escape from suboptimal corners of the probability simplex. Our results in Theorem 3.6 show that an appropriate degree of off-policyness can also reduce this delay without further algorithmic modifications.

Off-policy algorithms. Global convergence guarantees for dedicated off-policy RL algorithms allow state-distribution mismatch (Laroche & Tachet des Combes, 2021; Zhang et al., 2022). They retain current-policy action weights, directly or through importance-sampling ratios, whereas RE(S) keeps the rollout distribution fixed within each stage. FMA-PG provides policy-improvement guarantees for finite-step surrogate optimization within each inner loop, while SPMA's convergence bound includes an additive term for errors due to approximate fitting (Vaswani et al., 2022; Asad et al., 2025). Dai et al. (2026) prove global linear convergence of multi-step PPO-Clip with KL regularization under a bound on the sum of inner step sizes. All these guarantees do not directly yield global convergence rates for the uncorrected, unregularized reward-weighted update rule in RE(S) with a fixed S value; our work fills in this gap of the literature.

5 Limitations and Future Work

This work has investigated the learning dynamics of RE(S)— probably the simplest possible algorithm for fine-tuning on self-generated and reward-weighted data — with in-depth convergence analysis for multi-arm bandits and softmax policies. Future work may extend the theoretical study to broader settings, such as contextual bandits, Markov decision processes, policies with different parameterization, or other learning algorithms. Moreover, our analysis focuses on convergence properties and considers the infinite-sample limit of gradient dynamics; future work may investigate sample complexities in finite-sample settings. In terms of empirical work, it remains open to verify which parts of our results hold true in realistic scenarios, see if our theoretical results can inspire better practice of reward-guided self-training, and identify gaps between theory and practice that request further research.

AI use statement

In this work, we used generative AI tools only for assistance in polishing the writing, finding related work, and writing code for numerical experiments. We have reviewed all AI-assisted work, by manually verifying the AI-generated content and reading the papers suggested by AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution Shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Remi Munos. Asymmetric reinforce for off-policy reinforcement learning: Balancing positive and negative rewards. *Advances in Neural Information Processing Systems*, 38:9640–9664, 2026.

Reza Asad, Reza Babanezhad Harikandeh, Issam H. Laradji, Nicolas Le Roux, and Sharan Vaswani. Fast convergence of softmax policy mirror ascent. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3943–3951. PMLR, 03–05 May 2025.

Qiming Dai, Yin Liu, Junyu Zhang, and Zaiwen Wen. Non-asymptotic global convergence of ppo-clip. *arXiv*, 2026.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv*, 2025.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. *Transactions on Machine Learning Research*, 2023.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1407–1416. PMLR, 10–15 Jul 2018.

Dibya Ghosh, Marlos C. Machado, and Nicolas Le Roux. An Operator View of Policy Gradient Methods. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling. *arXiv*, 2023.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. OpenRLHF: An easy-to-use, scalable and high-performance RLHF framework. *arXiv preprint arXiv:2405.11143*, 2024.

Kimi-Team. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *arXiv*, 2023.

Romain Laroche and Remi Tachet des Combes. Dr jekyll & mr hyde: the strange case of off-policy policy updates. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24442–24454, 2021.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. *Mathematical Programming*, 201(1):707–802, 2023.

Jiacai Liu, Wenye Li, and Ke Wei. Elementary analysis of policy gradient methods. *arXiv*, 2024.

Michael Lu, Matin Aghaei, Anant Raj, and Sharan Vaswani. Towards principled, practical policy gradient for bandits and tabular mdps. *arXiv preprint arXiv:2405.13136*, 2024.

Jincheng Mei and Ian Osband. Delightful Gradients Accelerate Corner Escape. *arXiv preprint arXiv:2605.11908*, 2026.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the Gravitational Pull of Softmax. In *Advances in Neural Information Processing Systems*, volume 33, 2020a.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the Global Convergence Rates of Softmax Policy Gradient Methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6820–6829, 2020b.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *International Conference on Machine Learning*, pp. 24325–24360. PMLR, 2023.

OpenAI. OpenAI o1 system card. *arXiv Preprint arXiv:2412.16720*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022.

Xuchen Pan, Yanxi Chen, Yushuo Chen, Yuchang Sun, Daoyuan Chen, Wenhao Zhang, Yuexiang Xie, Yilun Huang, Yilei Zhang, Dawei Gao, Weijie Shi, Yaliang Li, Bolin Ding, and Jingren Zhou. Trinity-RFT: A general-purpose and unified framework for reinforcement fine-tuning of large language models. *arXiv Preprint arXiv:2505.17826*, 2025.

Chongli Qin and Jost Tobias Springenberg. Supervised Fine Tuning on Curated Data is Reinforcement Learning (and can be improved). *arXiv preprint arXiv:2507.12856*, 2025.

Samuel Robertson, Thang Chu, Bo Dai, Dale Schuurmans, Csaba Szepesvari, and Jincheng Mei. Reinforce converges to optimal policies with any learning rate. In *Advances in Neural Information Processing Systems*, 2025.

Daniel Russo. Success Conditioning as Policy Improvement: The Optimization Problem Solved by Imitating Success. *arXiv preprint arXiv:2601.18175*, 2026.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv*, 2024.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv*, 2024.

Miroslav Štrupl, Francesco Faccio, Dylan R Ashley, Rupesh Kumar Srivastava, and Jürgen Schmidhuber. Reward-Weighted Regression Converges to a Global Optimum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8361–8369, 2022.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen

Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023.

Sharan Vaswani, Olivier Bachem, Simone Totaro, Robert Müller, Shivam Garg, Matthieu Geist, Marlos C. Machado, Pablo Samuel Castro, and Nicolas Le Roux. A general class of surrogate functions for stable and efficient reinforcement learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8619–8649. PMLR, 28–30 Mar 2022.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126.

Shangtong Zhang, Remi Tachet des Combes, and Romain Laroche. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *Journal of Machine Learning Research*, 23(343):1–91, 2022.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

A PROOFS FOR SECTION 3.1

Lemma A.1. *Consider the surrogate objective function*

$$L(\pi_{b,k};\pi_{b,0}) = \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)\log \pi_{b,k}(a).$$

With $\pi_{b,0}$ fixed, the gradient with respect to the logit parameter θ, evaluated at $\theta = \theta_{b,k}$, satisfies

$$g_{b,k} \coloneqq \nabla_\theta L(\pi_\theta;\pi_{b,0})|_{\theta=\theta_{b,k}} = \pi_{b,0}\odot\mu - J(\pi_{b,0})\pi_{b,k}.$$

Proof. Note that

$$\begin{aligned}
\nabla_\theta L(\pi_\theta;\pi_{b,0})|_{\theta=\theta_{b,k}} &= \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)\frac{\nabla_\theta \pi_\theta(a)|_{\theta=\theta_{b,k}}}{\pi_{b,k}(a)} \\
&= \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)\frac{\pi_{b,k}(a)e_a - \pi_{b,k}(a)\pi_{b,k}}{\pi_{b,k}(a)} \\
&= \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)e_a - \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)\pi_{b,k} \\
&= \pi_{b,0}\odot\mu - J(\pi_{b,0})\pi_{b,k} = g_{b,k}.
\end{aligned}$$

□

Lemma A.2. *$\nabla_\theta L(\pi_\theta;\pi_{b,0})$ has a Lipschitz constant no greater than $J(\pi_{b,0})/2$.*

Proof. Define

$$H(\pi) = \mathrm{diag}(\pi) - \pi\pi^\top.$$

This matrix is positive semidefinite. By Gershgorin's theorem, its eigenvalues are bounded above by $\max_a 2\pi(a)(1-\pi(a)) \le 1/2$, and hence

$$0 \preceq H(p) \preceq I/2. \tag{12}$$

We also have

$$\nabla^2_\theta L(\pi_\theta;\pi_{b,0}) = -J(\pi_{b,0})H(\pi_\theta). \tag{13}$$

Thus $\nabla_\theta L(\pi_\theta;\pi_{b,0})$ has a Lipschitz constant no greater than $J(\pi_{b,0})/2$ □

Lemma A.3. *Suppose $\eta\mu_{max} < 4$. Consider how the logit of each action evolves under this assumption. We have*

$$\|g_{b,k}\|_2 \le \|g_{b,0}\|_2.$$

Proof. Fix b and note that the concave function

$$L(\pi_\theta;\pi_{b,0}) = \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)\theta(a) - J(\pi_{b,0})\log\sum_{a\in\mathcal{A}} e^{\theta(a)}.$$

The Hessian bound from Eq. (12) and Eq. (13) shows that $\nabla L(\pi_\theta;\pi_{b,0})$ has a Lipschitz constant of at most $J(\pi_{b,0})/2$. For $0 \le k < S$, by the standard co-coercivity inequality,

$$\begin{aligned}
&\left\langle \nabla L(\pi_{\theta_{b,k}};\pi_{b,0}) - \nabla L(\pi_{\theta_{b,k+1}};\pi_{b,0}), \theta_{b,k+1} - \theta_{b,k}\right\rangle \\
&\quad\ge \frac{2}{J(\pi_{b,0})}\left\|\nabla L(\pi_{\theta_{b,k}};\pi_{b,0}) - \nabla L(\pi_{\theta_{b,k+1}}\right\|_2^2.
\end{aligned}$$

Substituting $\nabla L(\pi_{\theta_{b,k}};\pi_{b,0}) = g_{b,k}$ and $\theta_{b,k+1} - \theta_{b,k} = \eta g_{b,k}$ gives

$$\eta\langle g_{b,k} - g_{b,k+1}, g_{b,k}\rangle \ge \frac{2}{J(\pi_{b,0})}\|g_{b,k+1} - g_{b,k}\|_2^2.$$

Therefore, since $\eta J(\pi_{b,0}) < 4$,

$$\begin{aligned}\|g_{b,k+1}\|_2^2 &= \|g_{b,k}\|_2^2 + 2\langle g_{b,k+1} - g_{b,k}, g_{b,k}\rangle + \|g_{b,k+1} - g_{b,k}\|_2^2 \\ &\le \|g_{b,k}\|_2^2 - \left(\frac{4}{\eta J(\pi_{b,0})} - 1\right)\|g_{b,k+1} - g_{b,k}\|_2^2 \\ &\le \|g_{b,k}\|_2^2.\end{aligned}$$

It follows that

$$\|g_{b,k}\|_2 \le \|g_{b,0}\|_2,$$

which completes our proof. □

Proof of Theorem 3.1. Using lemma A.2, we have

$$L(\pi_{b,k+1};\pi_{b,0}) - L(\pi_{b,k};\pi_{b,0}) \ge \eta\left(1 - \frac{\eta J(\pi_{b,0})}{4}\right)\|g_{b,k}\|_2^2. \tag{14}$$

Action 1 is uniquely optimal and rewards are nonnegative, so $\mu_{\max} > 0$. Every policy coordinate remains strictly positive after finitely many updates, and hence $J(\pi_{b,0}) \ge \mu_{\max}\pi_{b,0}(1) > 0$. The right side is nonnegative because $0 < J(\pi_{b,0}) \le \mu_{\max}$ and $\eta\mu_{\max} < 4$. For $0 \le k \le S$, the inequality $\log x \le x - 1$ yields

$$\begin{aligned}L(\pi_{b,k};\pi_{b,0}) - L(\pi_{b,0};\pi_{b,0}) &= \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)\log\frac{\pi_{b,k}(a)}{\pi_{b,0}(a)} \\ &\le \sum_{a\in\mathcal{A}} \pi_{b,0}(a)\mu(a)\left(\frac{\pi_{b,k}(a)}{\pi_{b,0}(a)} - 1\right) \\ &= J(\pi_{b,k}) - J(\pi_{b,0}).\end{aligned} \tag{15}$$

Hence $J(\pi_{b,k}) \ge J(\pi_{b,0})$. Taking $k = S$ gives $J(\pi_{b+1,0}) \ge J(\pi_{b,0})$. Since $J(\pi_{b,0}) \le \mu_{\max}$, there exists $J_\infty \le \mu_{\max}$ such that $J(\pi_{b,0}) \to J_\infty$. Summing Eq. (14) over the updates in the inner loop and using Eq. (15) gives

$$J(\pi_{b+1,0}) - J(\pi_{b,0}) \ge \eta\left(1 - \frac{\eta\mu_{\max}}{4}\right)\sum_{k=0}^{S-1}\|g_{b,k}\|_2^2. \tag{16}$$

Summing over b then yields

$$\sum_{b=0}^{\infty}\sum_{k=0}^{S-1}\|g_{b,k}\|_2^2 \le \frac{\mu_{\max} - J_0}{\eta(1 - \eta\mu_{\max}/4)} < \infty.$$

In particular, $g_{b,0} \to 0$. We have

$$g_{b,0}(a) = \pi_{b,0}(a)(\mu(a) - J(\pi_{b,0})),$$

Thus, if $\mu(a) \ne J_\infty$, then $\pi_{b,0}(a) \to 0$. Define

$$\mathcal{M} = \{a \in \mathcal{A} : \mu(a) = J_\infty\}.$$

Since the probabilities sum to 1, we have $\mathcal{M} \ne \varnothing$ and

$$\sum_{a\in\mathcal{M}} \pi_{b,0}(a) \longrightarrow 1. \tag{17}$$

We now prove $J_\infty = \mu(1)$. Suppose for contradiction that $J_\infty < \mu(1)$, and define

$$\mathcal{M}_+ = \{a : \mu(a) > J_\infty\}, \qquad \mathcal{M}_- = \{a : \mu(a) < J_\infty\}.$$

Then $1 \in \mathcal{M}_+$. Since $J(\pi_{b,0})$ increases monotonically to J_∞, we have $J(\pi_{b,0}) \le J_\infty$ throughout. By lemma A.3, we have

$$\max_{0\le k\le S}\|\theta_{b,k} - \theta_{b,0}\|_2 \le \eta S\|g_{b,0}\|_2 \longrightarrow 0. \tag{18}$$

By the definition of π_θ,

$$\frac{\pi_{b,k}(a)}{\pi_{b,0}(a)} = \frac{\exp(\theta_{b,k}(a) - \theta_{b,0}(a))}{\sum_{c\in\mathcal{A}} \pi_{b,0}(c)\exp(\theta_{b,k}(c) - \theta_{b,0}(c))} \longrightarrow 1, \tag{19}$$

Here the convergence is uniform over the finitely many actions a and steps $0 \le k \le S$. Thus

$$\frac{g_{b,k}(a)}{\pi_{b,0}(a)} = \mu(a) - J(\pi_{b,0})\frac{\pi_{b,k}(a)}{\pi_{b,0}(a)} \longrightarrow \mu(a) - J_\infty.$$

Since the action set is finite, there exists B_0 such that, whenever $b \ge B_0$, the logit of each action in $\mathcal{M}_+$ strictly increases at every update of the inner loop, while the logit of each action in $\mathcal{M}_-$ strictly decreases. More precisely,

$$\begin{aligned} \theta_{b+1,0}(a) - \theta_{b,0}(a) &\ge \frac{\eta S}{2}(\mu(a) - J_\infty)\pi_{b,0}(a), \qquad a \in \mathcal{M}_+, \\ \theta_{b+1,0}(a) - \theta_{b,0}(a) &\le -\frac{\eta S}{2}(J_\infty - \mu(a))\pi_{b,0}(a), \quad a \in \mathcal{M}_-. \end{aligned} \tag{20}$$

The logit of action 1 is nondecreasing from this point onward, so $\theta_{b,0}(1)$ is bounded below. At the same time, Eq. (17) and $\pi_{b,0}(1) \to 0$ give

$$\sum_{a\in\mathcal{M}} \exp(\theta_{b,0}(a) - \theta_{b,0}(1)) = \frac{\sum_{a\in\mathcal{M}} \pi_{b,0}(a)}{\pi_{b,0}(1)} \longrightarrow +\infty.$$

Hence

$$\max_{a\in\mathcal{M}} \theta_{b,0}(a) \longrightarrow +\infty. \tag{21}$$

When $\mathcal{M} = \{a_0\}$, the assumed convergence to a_0 therefore gives $\theta_{b,0}(a_0) \to +\infty$. Using the set $\mathcal{M}$ also covers the case where several suboptimal actions have the same reward. We now bound the entire gradient vector directly by the suboptimality gap. For notational simplicity, write

$$h_b := \|g_{b,0}\|_2.$$

After the rollout policy is refreshed for the next stage, its initial gradient satisfies

$$\begin{aligned} g_{b+1,0}(a) &= \pi_{b+1,0}(a)(\mu(a) - J(\pi_{b+1,0})) \\ &= \frac{\pi_{b+1,0}(a)}{\pi_{b,0}(a)} g_{b,0}(a) - (J(\pi_{b+1,0}) - J(\pi_{b,0}))\pi_{b+1,0}(a). \end{aligned}$$

The gradient norm does not increase within the inner loop, so $\|\theta_{b+1,0} - \theta_{b,0}\|_2 \le \eta S h_b$. The difference between the largest and smallest coordinates of a vector is at most $\sqrt{2}$ times its Euclidean norm. The ratio formula Eq. (19) for π_θ therefore gives

$$\frac{\pi_{b+1,0}(a)}{\pi_{b,0}(a)} \ge e^{-\sqrt{2}\eta S h_b}.$$

Combining the reverse triangle inequality with $\|\pi_{b+1,0}\|_2 \le 1$ and $J(\pi_{b+1,0}) - J(\pi_{b,0}) \ge 0$ gives

$$\begin{aligned} h_{b+1} &\ge \left(\sum_{a\in\mathcal{A}} \left[\frac{\pi_{b+1,0}(a)}{\pi_{b,0}(a)} g_{b,0}(a)\right]^2\right)^{1/2} - (J(\pi_{b+1,0}) - J(\pi_{b,0}))\|\pi_{b+1,0}\|_2 \\ &\ge e^{-\sqrt{2}\eta S h_b}\left(\sum_{a\in\mathcal{A}} g_{b,0}(a)^2\right)^{1/2} - (J(\pi_{b+1,0}) - J(\pi_{b,0})) \\ &= e^{-\sqrt{2}\eta S h_b} h_b - (J(\pi_{b+1,0}) - J(\pi_{b,0})). \end{aligned}$$

Using $1 - e^{-x} \le x$, we obtain

$$\begin{aligned} h_b - h_{b+1} &\le (1 - e^{-\sqrt{2}\eta S h_b})h_b + (J(\pi_{b+1,0}) - J(\pi_{b,0})) \\ &\le \sqrt{2}\eta S h_b^2 + (J(\pi_{b+1,0}) - J(\pi_{b,0})). \end{aligned}$$

The lower bound Eq. (16) on the increase in the objective gives

$$\eta\left(1-\frac{\eta\mu_{\max}}{4}\right)h_b^2 \le J(\pi_{b+1,0}) - J(\pi_{b,0}),$$

Then $h_b - h_{b+1} \le (1+\sqrt{2}S/(1-\eta\mu_{\max}/4))(J(\pi_{B+1,0}) - J(\pi_{b,0}))$. Summing from round b to round B gives

$$h_b - h_{B+1} \le (1+\sqrt{2}S/(1-\eta\mu_{\max}/4))(J(\pi_{B+1,0}) - J(\pi_{b,0})).$$

Since $h_{B+1} \to 0$ and $J(\pi_{B+1,0}) \to J_\infty$, letting $B \to \infty$ yields

$$\|g_{b,k}\|_2 \le \|g_{b,0}\|_2 \le (1+\sqrt{2}S/(1-\eta\mu_{\max}/4))(J_\infty - J(\pi_{b,0})), \qquad 0 \le k < S.$$

If $\sum_b (J_\infty - J(\pi_{b,0})) < \infty$, then

$$\begin{aligned}\sum_{b=0}^{\infty}\|\theta_{b+1,0}-\theta_{b,0}\|_2 &\le \eta\sum_{b=0}^{\infty}\sum_{k=0}^{S-1}\|g_{b,k}\|_2 \\ &\le \eta S(1+\sqrt{2}S/(1-\eta\mu_{\max}/4))\sum_{b=0}^{\infty}(J_\infty - J(\pi_{b,0})) < \infty.\end{aligned}$$

Every logit coordinate would then converge to a finite value, contradicting Eq. (21). Therefore

$$\sum_{b=0}^{\infty}(J_\infty - J(\pi_{b,0})) = +\infty. \tag{22}$$

We now use the increasing logits of actions with higher rewards and the decreasing logits of actions with lower rewards to derive a contradiction. We have

$$\begin{aligned}0 \le J_\infty - J(\pi_{b,0}) &= \sum_{a\in\mathcal{M}_-}(J_\infty - \mu(a))\pi_{b,0}(a) - \sum_{a\in\mathcal{M}_+}(\mu(a) - J_\infty)\pi_{b,0}(a) \\ &\le \sum_{a\in\mathcal{M}_-}(J_\infty - \mu(a))\pi_{b,0}(a)\end{aligned}$$

Together with Eq. (22), this shows that $\mathcal{M}_- \neq \varnothing$ and that there exists $\ell \in \mathcal{M}_-$ such that

$$\sum_{b=0}^{\infty}\pi_{b,0}(\ell) = +\infty.$$

Summing the inequality in Eq. (20) for action ℓ gives $\theta_{b,0}(\ell) \to -\infty$. Since $\theta_{b,0}(1)$ is bounded below,

$$\frac{\pi_{b,0}(1)}{\pi_{b,0}(\ell)} = \exp\big(\theta_{b,0}(1) - \theta_{b,0}(\ell)\big) \longrightarrow +\infty.$$

Thus, for all sufficiently large b, we have $\pi_{b,0}(1) \ge \pi_{b,0}(\ell)$, and hence

$$\sum_{b=0}^{\infty}\pi_{b,0}(1) = +\infty.$$

Applying Eq. (20) again gives $\theta_{b,0}(1) \to +\infty$. For every $a \in \mathcal{M}_-$, $\theta_{b,0}(a)$ is nonincreasing from B_0 onward and is therefore bounded above. Hence

$$\frac{\pi_{b,0}(a)}{\pi_{b,0}(1)} = \exp\big(\theta_{b,0}(a) - \theta_{b,0}(1)\big) \longrightarrow 0, \qquad a \in \mathcal{M}_-.$$

Consequently,

$$\begin{aligned}J(\pi_{b,0}) - J_\infty &= \sum_{a\in\mathcal{M}_+}(\mu(a) - J_\infty)\pi_{b,0}(a) - \sum_{a\in\mathcal{M}_-}(J_\infty - \mu(a))\pi_{b,0}(a) \\ &\ge \pi_{b,0}(1)\left[\mu(1) - J_\infty - \sum_{a\in\mathcal{M}_-}(J_\infty - \mu(a))\frac{\pi_{b,0}(a)}{\pi_{b,0}(1)}\right] > 0\end{aligned}$$

This holds for all sufficiently large b, contradicting $J(\pi_{b,0}) \le J_\infty$. Hence $J_\infty = \mu(1)$. Since action 1 is uniquely optimal, $\mathcal{M} = \{1\}$, and Eq. (17) gives

$$\pi_{b,0} \longrightarrow e_1, \qquad J(\pi_{b,0}) \longrightarrow \mu(1).$$

This completes the proof.

B Proofs for Section 3.2

B.1 Proof of Lemma 3.2

Constants. The constants in Lemma 3.2 are

$$A = 4 + 6\eta\mu_{\max}, \qquad \rho = \frac{\mu_{\min}\Delta}{\mu_{\max}^2}\exp\left(-\frac{\mu_{\max}^2}{\mu_{\min}\Delta}\right).$$

$$\lambda(\pi_{b,0}(1)) = \left(1 - \frac{\eta\mu_{\max}}{4}\right)\frac{\pi_{b,0}(1)^2}{8\sqrt{2}}\log\left(1 + \frac{\Delta}{4\mu_{\max}}\right).$$

Here $0 < \rho < 1$ and $0 < \lambda(\pi_{b,0}(1)) < 1$.

Focusing on the b-th outer iteration, we write

$$q \coloneqq \pi_{b,0}, \quad d \coloneqq D(q) > 0, \quad p_k \coloneqq \pi_{b,k}, \quad g_k \coloneqq g_{b,k}$$

for notational simplicity. Finite logits and the unique optimum ensure $d > 0$. We keep the inner loop index on logits only when needed.

For this fixed q, Eq. (14) ensures that $L(p_{k+1}; q) \geq L(p_k; q)$ for every $0 \leq k < S$. By Lemma A.3,

$$\|g_k\|_2 \leq \|g_0\|_2 \leq \|g_0\|_1 = \sum_a q(a)|d - (\mu_{\max} - \mu(a))| \leq \sum_a q(a)(d + \mu_{\max} - \mu(a)) = 2d. \tag{23}$$

Lower bound. Let

$$r_k = \max_a(\theta_{b,k}(a) - \theta_{b,0}(a)) - \min_a(\theta_{b,k}(a) - \theta_{b,0}(a)).$$

Note that

$$\frac{p_k(a)}{q(a)} = \frac{\exp(\theta_{b,k}(a) - \theta_{b,0}(a))}{\sum_j q(j)\exp(\theta_{b,k}(j) - \theta_{b,0}(j))}. \tag{24}$$

Eq. (24) gives $e^{-r_k} \leq p_k(a)/q(a) \leq e^{r_k}$. Then we have

$$p_k(1) \leq \frac{q(1)e^{r_k}}{1 - q(1) + q(1)e^{r_k}}, \qquad p_k(1) - q(1) \leq q(1)(1 - q(1))(e^{r_k} - 1). \tag{25}$$

Also, $d = \sum_{a>1}(\mu_{\max} - \mu(a))q(a) \geq \Delta(1 - q(1))$. For any $1 \leq n \leq S$, choose

$$N = \min\left\{n, \left\lfloor \frac{\log(1 + \Delta/(4\mu_{\max}))}{2\sqrt{2}\eta d} \right\rfloor + 1\right\}.$$

By Eq. (23), for every $k < N$,

$$r_k \leq \sqrt{2}\eta\sum_{j<k}\|g_j\|_2 \leq 2\sqrt{2}\eta k d \leq \log\left(1 + \frac{\Delta}{4\mu_{\max}}\right).$$

Substituting into Eq. (25) yields

$$p_k(1) - q(1) \leq \frac{q(1)(1 - q(1))\Delta}{4\mu_{\max}} \leq \frac{q(1)d}{4\mu_{\max}}.$$

Since $0 < J(\pi_{b,0}) \leq \mu_{\max}$, this implies

$$g_k(1) = q(1)d - J(\pi_{b,0})(p_k(1) - q(1)) \geq \frac{1}{2}q(1)d$$

Sum Eq. (14) over $k = 0, \ldots, N-1$ and use monotonicity between steps N and n to obtain

$$\begin{aligned} L(p_n; q) - L(q; q) &\geq L(p_N; q) - L(q; q) \\ &\geq \left(1 - \frac{\eta\mu_{\max}}{4}\right)\frac{q(1)^2\eta N d^2}{4} \\ &\geq \left(1 - \frac{\eta\mu_{\max}}{4}\right)\frac{q(1)^2\log(1 + \Delta/(4\mu_{\max}))}{8\sqrt{2}}\min\{\eta n d^2, d\}. \end{aligned}$$

For the last inequality, set $\alpha = \log(1 + \Delta/(4\mu_{\max}))/(2\sqrt{2}) \in (0, 1)$. Then $N \geq \min\{n, \alpha/(\eta d)\}$, so $\eta N d^2 \geq \min\{\eta n d^2, \alpha d\} \geq \alpha\min\{\eta n d^2, d\}$.

Upper bound. We have

$$\nabla J(\pi(\theta))(a) = \pi(a)(\mu(a) - J(\pi)), \qquad \|\nabla J(\pi(\theta))\|_2 \le \|\nabla J(\pi(\theta))\|_1 \le 2D(\pi).$$

Equations (14) and (15) imply $D(p_k) \le d$, hence $\|\nabla J(\pi(\theta_{b,k}))\|_2 \le 2d$. Differentiating once more yields

$$\nabla^2 J(\pi(\theta)) = \operatorname{diag}(\nabla J(\pi(\theta))) - \nabla J(\pi(\theta))\pi^\top - \pi \nabla J(\pi(\theta))^\top.$$

Because $|\mu(a) - J(\pi)| \le \mu_{\max}$, we have $\|\nabla J(\pi(\theta))\|_2 \le \|\nabla J(\pi(\theta))\|_1 \le \mu_{\max}$, $\|\operatorname{diag}(\nabla J(\pi(\theta)))\|_2 \le \mu_{\max}$, and $\|\pi\|_2 \le 1$. Therefore $\|\nabla^2 \nabla J(\pi(\theta))\|_2 \le 3\mu_{\max}$. From Taylor's theorem and Eq. (23) we have

$$\begin{aligned} J(p_{k+1}) - J(p_k) &\le \eta \nabla J(\pi(\theta_{b,k}))^\top g_k + \frac{3\mu_{\max}\eta^2}{2}\|g_k\|_2^2 \\ &\le 4\eta d^2 + 6\mu_{\max}\eta^2 d^2 = (4 + 6\eta\mu_{\max})\eta d^2. \end{aligned}$$

Summing over $k < n$ and using $0 \le J(p_n) - J(\pi_{b,0}) \le d$ proves the upper bound in Eq. (2). The middle inequality is Eq. (15).

A lower bound on the suboptimality gap. We prove a bound for any positive policy p satisfying $L(p;q) \ge L(q;q)$, and then apply it to p_n. The gap and reward bounds imply

$$\Delta \sum_{a>1} q(a) \le d \le \mu_{\max} \sum_{a>1} q(a), \qquad H_q := \sum_{a>1} q(a)\mu(a) \ge \mu_{\min} \sum_{a>1} q(a) \ge \frac{\mu_{\min} d}{\mu_{\max}} > 0.$$

The optimal action contributes at most

$$\mu_{\max} q(1) \log \frac{p(1)}{q(1)} \le \mu_{\max}(p(1) - q(1)) \le \mu_{\max}(1 - q(1)) \le \frac{\mu_{\max} d}{\Delta}.$$

Jensen's inequality for the suboptimal terms gives

$$\begin{aligned} \sum_{a>1} q(a)\mu(a) \log \frac{p(a)}{q(a)} &= H_q \sum_{a>1} \frac{q(a)\mu(a)}{H_q} \log \frac{p(a)}{q(a)} \\ &\le H_q \log \left(\sum_{a>1} \frac{q(a)\mu(a)}{H_q} \frac{p(a)}{q(a)} \right) = H_q \log \frac{\sum_{a>1} \mu(a)p(a)}{H_q}. \end{aligned}$$

Since $D(p) \ge \Delta \sum_{a>1} p(a)$, we have

$$\frac{\sum_{a>1} \mu(a)p(a)}{H_q} \le \frac{\mu_{\max} D(p)}{\Delta H_q} \le \frac{\mu_{\max}^2}{\mu_{\min}\Delta} \frac{D(p)}{d}.$$

Consequently,

$$0 \le L(p;q) - L(q;q) \le \frac{\mu_{\max} d}{\Delta} + H_q \log \left(\frac{\mu_{\max}^2}{\mu_{\min}\Delta} \frac{D(p)}{d} \right).$$

Rearranging and using $H_q \ge \mu_{\min} d / \mu_{\max}$ yields

$$\log \left(\frac{\mu_{\max}^2}{\mu_{\min}\Delta} \frac{D(p)}{d} \right) \ge -\frac{\mu_{\max} d}{\Delta H_q} \ge -\frac{\mu_{\max}^2}{\mu_{\min}\Delta}, \qquad D(p) \ge \rho d.$$

Taking $p = p_n$ and combining with $D(p_n) \le d$ proves Eq. (3) in the lemma.

B.2 Proofs of Theorems 3.3 and 3.4

Lemma B.1. *Under the assumptions of Lemma 3.2,* $p_{\min} := \inf_{b \ge 0} \pi_{b,0}(1) > 0$. *Set* $\lambda_* = \lambda(p_{\min})$. *For every* $b \ge 0$ *and* $0 \le s \le S$,

$$\frac{1}{D(\pi_{b,s})} - \frac{1}{d_b} \ge \lambda_* \min \left\{ \eta s, \frac{1}{d_b} \right\}, \tag{26}$$

$$\frac{1}{D(\pi_{b,s})} - \frac{1}{d_b} \le \min \left\{ \frac{A}{\rho}\eta s, \frac{1-\rho}{\rho}\frac{1}{d_b} \right\}. \tag{27}$$

Proof. Using Lemma 3.2 and $D(\pi_{b,s}) \le d_b$ give

$$\frac{1}{D(\pi_{b,s})} - \frac{1}{d_b} = \frac{J(\pi_{b,s}) - J(\pi_{b,0})}{D(\pi_{b,s}) d_b} \ge \lambda_* \min\left\{\eta s, \frac{1}{d_b}\right\}.$$

For the upper bound, let $\xi = (d_b - D(\pi_{b,s}))/d_b$. Then

$$0 \le \xi \le A\eta s d_b, \qquad \xi \le 1-\rho, \qquad \frac{1}{D(\pi_{b,s})} - \frac{1}{d_b} = \frac{\xi}{(1-\xi)d_b}.$$

Bounding the last expression with these two inequalities proves Eq. (27). At $s = 0$, both sides are zero. □

For $x, y > 0$, define the comparison sequence

$$a_{0,0}^{x,y} = d_0^{-1}, \qquad a_{b,s}^{x,y} = a_{b,0}^{x,y} + \min\{x\eta s, y a_{b,0}^{x,y}\}, \qquad a_{b+1,0}^{x,y} = a_{b,S}^{x,y}. \tag{28}$$

Set

$$b_{x,y} = \begin{cases} 0, & \eta S d_0 \le y/x, \\ \left\lceil \dfrac{\log(x\eta S d_0/y)}{\log(1+y)} \right\rceil, & \eta S d_0 > y/x. \end{cases} \tag{29}$$

Define the following function:

$$\mathcal{C}_{x,y}(t,S,d_0) = \begin{cases} d_0^{-1} + x\eta t, & \eta S d_0 \le y/x, \\ \dfrac{(1+y)^b}{d_0} + \min\left\{x\eta s, \dfrac{y(1+y)^b}{d_0}\right\}, & \eta S d_0 > y/x,\ b < b_{x,y}, \\ \dfrac{(1+y)^{b_{x,y}}}{d_0} + x\eta(t - S b_{x,y}), & \eta S d_0 > y/x,\ b \ge b_{x,y}. \end{cases} \tag{30}$$

Then we have:

Lemma B.2. *The map $a \mapsto a + \min\{x\eta s, ya\}$ is increasing for every $s \ge 0$. For $t = bS + s$ with $0 \le s < S$, we have $a_{b,s}^{x,y} = C_{x,y}(t, S, d_0)$.*

Proof. The map is increasing because a is strictly increasing and $\min\{x\eta s, ya\}$ is nondecreasing in a. To compute the sequence, first consider the complete-stage recurrence

$$a_{b+1,0}^{x,y} = \begin{cases} (1+y)a_{b,0}^{x,y}, & a_{b,0}^{x,y} < x\eta S/y, \\ a_{b,0}^{x,y} + x\eta S, & a_{b,0}^{x,y} \ge x\eta S/y. \end{cases}$$

Since the sequence is increasing, once it reaches $x\eta S/y$, every subsequent stage adds the same amount $x\eta S$.

If $\eta S d_0 \le y/x$, then $a_{0,0}^{x,y} = d_0^{-1} \ge x\eta S/y$. The additive update therefore applies from the first stage, giving $a_{b,0}^{x,y} = d_0^{-1} + bx\eta S$.

If $\eta S d_0 > y/x$, each complete stage initially multiplies the current value by $1+y$. The first stage index at which the threshold is reached is the smallest integer b satisfying

$$\frac{(1+y)^b}{d_0} \ge \frac{x\eta S}{y},$$

which is exactly $b_{x,y}$ in Eq. (29). Consequently,

$$a_{b,0}^{x,y} = \begin{cases} \dfrac{(1+y)^b}{d_0}, & 0 \le b < b_{x,y}, \\ \dfrac{(1+y)^{b_{x,y}}}{d_0} + x\eta S(b - b_{x,y}), & b \ge b_{x,y}. \end{cases}$$

Finally, the definition of $a_{b,s}^{x,y}$ adds $\min\{x\eta s, y a_{b,0}^{x,y}\}$ to the stage-start value. Before the threshold is reached, this minimum remains as written. Once $a_{b,0}^{x,y} \ge x\eta S/y$, it equals $x\eta s$ because $s < S$. Substituting the stage-start values above and using $t = bS + s$ gives the three branches of Eq. (30). □

Proof of main theorems. By Lemmas B.1 and B.2, induction over complete stages, starting from $1/d_0$, gives

$$a_{b,s}^{\lambda_*,\lambda_*} \le \frac{1}{D(\pi_{b,s})} \le a_{b,s}^{A/\rho,(1-\rho)/\rho}.$$

At the final incomplete stage, apply Eqs. (26)–(27) with its actual number s of completed steps. Then we have the two-sided bound

$$\frac{1}{\mathcal{C}_{A/\rho,(1-\rho)/\rho}(t,S,d_0)} \le D(\pi_{b,s}) \le \frac{1}{\mathcal{C}_{\lambda_*,\lambda_*}(t,S,d_0)}.$$

Upper bound. Set $c = \lambda_*$ and let $\bar{b} = b_{c,c}$. For $0 \le b \le \bar{b}$, Eq. (30) gives

$$D(\pi_{b,s}) \le d_0(1+c)^{-b}, \qquad 0 \le s \le S.$$

At $b = \bar{b}$, it also gives $d_{\bar{b}} \le 1/(\eta S)$. By monotonicity, we have $1/d_b \ge \eta S$ for every $b \ge \bar{b}$. For these stages, the minimum in Eq. (26) equals ηs. Summing over complete stages and then the final s steps yields

$$D(\pi_{b,s}) \le \frac{1}{d_{\bar{b}}^{-1} + c\eta(t - \bar{b}S)}, \qquad b \ge \bar{b}, \quad t = bS + s.$$

Lower bound. For the upper comparison sequence, define

$$\bar{b}' = b_{A/\rho,(1-\rho)/\rho} = \begin{cases} 0, & \eta S d_0 \le (1-\rho)/A, \\ \left\lceil \dfrac{\log(A\eta S d_0/(1-\rho))}{\log(1/\rho)} \right\rceil, & \eta S d_0 > (1-\rho)/A. \end{cases} \tag{31}$$

Here $1 + y = 1/\rho$. Eq. (30) gives the bound

$$D(\pi_{b,s}) \ge \left[\frac{\rho^{-b}}{d_0} + \min\left\{ \frac{A}{\rho}\eta s, \frac{(1-\rho)\rho^{-b}}{\rho d_0} \right\} \right]^{-1}, \qquad b < \bar{b}'.$$

In particular, for the start and end of a stage, we have

$$D(\pi_{b,0}) \ge d_0\rho^b, \qquad D(\pi_{b,S}) \ge d_0\rho^{b+1}. \tag{32}$$

For $b \ge \bar{b}'$, the Eq. (30) gives

$$D(\pi_{b,s}) \ge \frac{1}{d_0^{-1}\rho^{-\bar{b}'} + (A/\rho)\eta(t - \bar{b}'S)}.$$

□

B.3 Explicit constants for the convergence rates

Lemma B.3. *If $0 < \eta\mu_{max} \le 2$ and $\pi_{0,0}(1) \ge \pi_{0,0}(a)$ for every $a > 1$, then $\pi_{b,s}(1) \ge \pi_{b,s}(a)$ for every $b \ge 0$ and $0 \le s \le S$. Consequently, $p_{\min} \ge 1/K$.*

Proof. Let $q = \pi_{b,0}$ and we have $q(1) \ge q(a)$ for all $a > 1$. Fix $a > 1$ and write $\xi_k = \theta_{b,k}(1) - \theta_{b,k}(a)$. Then $\xi_0 = \log(q(1)/q(a)) \ge 0$, and $q(1)\mu_{\max} - q(a)\mu(a) > 0$ because action 1 is uniquely optimal and $q(a) > 0$. Whenever $\xi_k \ge 0$,

$$0 \le \pi_{b,k}(1) - \pi_{b,k}(a) = \frac{e^{\theta_{b,k}(a)}(e^{\xi_k} - 1)}{\sum_{j=1}^K e^{\theta_{b,k}(j)}} \le \frac{e^{\xi_k} - 1}{e^{\xi_k} + 1} = \tanh(\xi_k/2) \le \frac{\xi_k}{2}.$$

The update therefore gives

$$\begin{aligned} \xi_{k+1} &= \xi_k + \eta\big[q(1)\mu_{\max} - q(a)\mu(a) - J(q)(\pi_{b,k}(1) - \pi_{b,k}(a))\big] \\ &\ge \left(1 - \frac{\eta J(q)}{2}\right)\xi_k + \eta(q(1)\mu_{\max} - q(a)\mu(a)) \ge 0. \end{aligned}$$

The last inequality uses $\eta J(q) \le \eta\mu_{\max} \le 2$. Induction over k preserves the ordering within the stage. Taking $k = S$ transfers it to the next stage, so induction over b proves the ordering at every iterate. In particular, $\pi_{b,0}(1) \ge 1/K$ for every b. □

Explicit rate constants. By Lemma B.3, $p_{\min} \geq 1/K$. Set $C_2 = \lambda(1/K)$ and $C_1 = A/\rho$. Since $A \geq 4$ and $0 < \rho < 1$, $(1-\rho)/\rho \leq C_1$. Equations (26) and (27) now imply

$$C_2 \min\left\{\eta s, \frac{1}{d_b}\right\} \leq \frac{1}{D(\pi_{b,s})} - \frac{1}{d_b} \leq C_1 \min\left\{\eta s, \frac{1}{d_b}\right\}.$$

Applying the same comparison as in Appendix B.2 gives

$$\frac{1}{\mathcal{C}_{C_1,C_1}(t,S,d_0)} \leq D(\pi_{b,s}) \leq \frac{1}{\mathcal{C}_{C_2,C_2}(t,S,d_0)}, \qquad t = bS + s.$$

For completeness, writing $b_\beta = b_{\beta,\beta}$, the symmetric comparison function is

$$\mathcal{C}_{\beta,\beta}(t,S,d_0) = \begin{cases} d_0^{-1} + \beta\eta t, & \eta S d_0 \leq 1, \\ \dfrac{(1+\beta)^b}{d_0} + \beta \min\left\{\eta s, \dfrac{(1+\beta)^b}{d_0}\right\}, & \eta S d_0 > 1,\ b < b_\beta, \\ \dfrac{(1+\beta)^{b_\beta}}{d_0} + \beta\eta(t - S b_\beta), & \eta S d_0 > 1,\ b \geq b_\beta. \end{cases}$$

If $\eta S d_0 > 1$, then $b_\beta = \lceil \log(\eta S d_0)/\log(1+\beta) \rceil$, otherwise $b_\beta = 0$.

C Proofs for Section 3.3

Throughout this section, we consider the problem settings as described in Section 3.3, with an action space of size $K \geq 3$. For a full-support policy q and positive reward-mean vector μ, write

$$\widehat{\pi}_q(a) := \frac{q(a)\mu(a)}{J(q)}, \qquad 1 \leq a \leq K.$$

When $q = \pi_{b,0}$, this is the reward-weighted target $\tilde{\pi}_b^\mu$ from Section 2. All constants denoted by $c, C > 0$ below, whose values can change across different parts of our analysis, are independent of the initial optimal-action probability x and parameters like B, S, T that account for the total number of iterations in the RE(S) algorithm.

C.1 Part 1 of Theorem 3.6: lower bound for on-policy RE(1)

For $S = 1$, each stage consists of one gradient step, so $\pi_{b,0}$ is reached after exactly b steps. We will show that reaching the target accuracy requires passing through a policy that assigns probability $1 - O(x^{K-1})$ to action 2. The following estimate then bounds how quickly the remaining probability can grow.

Lemma C.1. *Fix an action j and let $m_b = 1 - \pi_{b,0}(j)$. Then we have*

$$\frac{1}{m_{b+1}} \geq \frac{1}{m_b} - 2\eta\mu_{max}. \tag{33}$$

Therefore, for integers $b, n \geq 0$ such that $2\eta\mu_{max} n m_b < 1$,

$$\pi_{b+n,0}(a) \leq m_{b+n} \leq \frac{m_b}{1 - 2\eta\mu_{max} n m_b}, \qquad \forall a \neq j. \tag{34}$$

Proof. For $a \neq j$, we have $|g_{b,0}(a)| \leq \mu_{\max}\pi_{b,0}(a) \leq \mu_{\max} m_b$. Since $\sum_a g_{b,0}(a) = 0$, also $|g_{b,0}(j)| \leq \sum_{a \neq j} |g_{b,0}(a)| \leq \mu_{\max} m_b$. The update of each probability ratio gives

$$\begin{aligned} \frac{m_{b+1}}{1 - m_{b+1}} &= \sum_{a \neq j} \frac{\pi_{b,0}(a)}{\pi_{b,0}(j)} \exp\big(\eta(g_{b,0}(a) - g_{b,0}(j))\big) \\ &\leq \frac{m_b}{1 - m_b} \exp(2\eta\mu_{\max} m_b). \end{aligned}$$

Policy probability remain positive at finite iterates for each action. Using $e^{-v} \geq 1 - v$, we obtain

$$\begin{aligned}\frac{1}{m_{b+1}} - 1 &\geq \left(\frac{1}{m_b} - 1\right) e^{-2\eta\mu_{\max} m_b} \\ &\geq \left(\frac{1}{m_b} - 1\right)(1 - 2\eta\mu_{\max} m_b).\end{aligned}$$

Hence $1/m_{b+1} \geq 1/m_b - 2\eta\mu_{\max}(1 - m_b) \geq 1/m_b - 2\eta\mu_{\max}$. Summing this inequality over n updates and taking reciprocals when the resulting lower bound is positive proves Eq. (34). □

Proof of the first part of Theorem 3.6. For $S = 1$, Eqs. (14) and (15) imply $J(\pi_{b+1,0}) \geq J(\pi_{b,0})$. We first use this monotonicity to obtain a uniformly negative advantage for every action $a \geq 3$. Set

$$\overline{\mu} := \max_{3 \leq a \leq K} \mu(a), \qquad \kappa := \frac{\mu(2) - \overline{\mu}}{4} > 0.$$

Consider the separate bandit on actions $2, \ldots, K$, initialized at w. It has a unique optimal action 2, finite initial logits $\log w(a)$, and $\eta\mu(2) < 4$. Theorem 3.1, applied with $S = 1$, shows that its mean reward converges to $\mu(2)$. Choose a fixed integer $b_0 \geq 1$ at which this mean reward is at least $\overline{\mu} + 2\kappa$.

The full on-policy update has the probability representation

$$\pi_{b+1,0}(a) = \frac{\pi_{b,0}(a)\exp(\eta g_{b,0}(a))}{\sum_{j=1}^{K} \pi_{b,0}(j)\exp(\eta g_{b,0}(j))}. \tag{35}$$

When $x = 0$, action 1 has zero probability and the update reduces to that of the bandit on actions $2, \ldots, K$, initialized at w. By the choice of b_0, this reduced bandit has mean reward at least $\overline{\mu} + 2\kappa$ after b_0 updates. The update map is continuous in the policy, including at the boundary of the simplex. Thus, for sufficiently small $x > 0$, the mean reward after the same b_0 updates differs from that of the reduced bandit by less than κ, giving $J(\pi_{b_0,0}) \geq \overline{\mu} + \kappa$. Since the mean reward is nondecreasing and $\mu(a) \leq \overline{\mu}$ for every $a \geq 3$, it follows that

$$J(\pi_{b,0}) - \mu(a) \geq \kappa, \qquad b \geq b_0, \quad 3 \leq a \leq K. \tag{36}$$

Here b_0 is chosen from the reduced bandit and depends not on x.

Probability concentration on the suboptimal action. Define the cumulative increase of the optimal logit by

$$\ell_b := \eta \sum_{k=0}^{b-1} \pi_{k,0}(1) D(\pi_{k,0}) = \theta_{b,0}(1) - \theta_{0,0}(1). \tag{37}$$

This sequence is nondecreasing, with increments in $[0, \eta\mu_{\max}]$. The probability-ratio update and $\sum_a g_{b,0}(a) = 0$ give the exact identity

$$\prod_{a=2}^{K} \frac{\pi_{b,0}(a)}{\pi_{b,0}(1)} = \frac{(1-x)^{K-1} \prod_{a=2}^{K} w(a)}{x^{K-1}} e^{-K\ell_b}. \tag{38}$$

Indeed, the one-step change in the logarithm of the left-hand side is $\eta \sum_{a=2}^{K} (g_{b,0}(a) - g_{b,0}(1)) = -K\eta g_{b,0}(1)$.

Now we write $b_\epsilon := T_\epsilon(1)$. Since

$$D(\pi_{b_\epsilon,0}) \geq \Delta\big(1 - \pi_{b_\epsilon,0}(1)\big),$$

the condition $\epsilon < \Delta/2$ implies $\pi_{b_\epsilon,0}(1) > 1/2$. Every factor on the left-hand side of Eq. (38) is then less than one, so

$$\ell_{b_\epsilon} > \frac{1}{K} \log \frac{(1-x)^{K-1} \prod_{a=2}^{K} w(a)}{x^{K-1}} = \frac{K-1}{K} \log(1/x) + O(1).$$

On the other hand, $\ell_{b_0} \leq b_0 \eta\mu_{\max}$. For sufficiently small x, it follows that $b_\epsilon > b_0$ and $\ell_{b_\epsilon} - \ell_{b_0} > 1$. Thus the first crossing index

$$b_c := \min\{b \geq b_0 : \ell_b - \ell_{b_0} \geq 1\}$$

exists and satisfies $b_0 < b_c \le b_\epsilon$. The one-step bound on ℓ_b gives

$$1 \le \ell_{b_c} - \ell_{b_0} \le 1 + \eta\mu_{\max}, \qquad \ell_{b_c} \le 1 + (b_0 + 1)\eta\mu_{\max}. \tag{39}$$

For $a \ge 3$, write $r_b(a) = \pi_{b,0}(1)/\pi_{b,0}(a)$. When $b \ge b_0$, Eq. (36) and the nonnegative optimal advantage imply $g_{b,0}(1) - g_{b,0}(a) \ge \kappa\pi_{b,0}(a)$. Using $e^v - 1 \ge v$ in the exact ratio update, we obtain

$$\begin{aligned} r_{b+1}(a) - r_b(a) &= r_b(a)\left[e^{\eta(g_{b,0}(1) - g_{b,0}(a))} - 1\right] \\ &\ge \eta\kappa\pi_{b,0}(1) \ge \frac{\kappa}{\mu_{\max}}(\ell_{b+1} - \ell_b). \end{aligned} \tag{40}$$

Here the last inequality follows from $\ell_{b+1} - \ell_b = \eta\pi_{b,0}(1)D(\pi_{b,0}) \le \eta\mu_{\max}\pi_{b,0}(1)$. Summing from b_0 to $b_c - 1$ gives $r_{b_c}(a) \ge \kappa/\mu_{\max}$, and hence

$$\frac{\pi_{b_c,0}(a)}{\pi_{b_c,0}(1)} \le \frac{\mu_{\max}}{\kappa}, \qquad 3 \le a \le K. \tag{41}$$

By Eq. (39), the product in Eq. (38) at $b = b_c$ is bounded below by $cx^{-(K-1)}$. Its $K-2$ factors corresponding to actions $3, \dots, K$ are bounded above by Eq. (41). The remaining factor therefore satisfies

$$\frac{\pi_{b_c,0}(2)}{\pi_{b_c,0}(1)} \ge cx^{-(K-1)}.$$

Combining these ratio bounds gives

$$\pi_{b_c,0}(a) \le Cx^{K-1} \quad (a \ne 2), \qquad m_{b_c} := 1 - \pi_{b_c,0}(2) \le Cx^{K-1}. \tag{42}$$

The constants are independent of x, although b_c may depend on it. In particular, $\pi_{b_c,0}(1) < 1/2$ for sufficiently small x, which also shows that $b_c < b_\epsilon$.

The steps required to reach the target. Apply Lemma C.1 with $j = 2$, so $m_b = 1 - \pi_{b,0}(2)$. At the successful endpoint, $m_{b_\epsilon} \ge \pi_{b_\epsilon,0}(1) > 1/2$. Summing Eq. (33) from b_c to $b_\epsilon - 1$ and using Eq. (42), we obtain

$$T_\epsilon(1) = b_\epsilon \ge b_\epsilon - b_c \ge \frac{m_{b_c}^{-1} - 2}{2\eta\mu_{\max}} \ge cx^{-(K-1)}$$

for all sufficiently small x. This proves the first part of Theorem 3.6. □

C.2 Part 2 of Theorem 3.6: upper bound for off-policy RE(S)

We first bound the error in fitting a frozen target. We then use this bound to show that a sufficiently long inner loop preserves part of the target's improvement in the optimal-action log-odds.

Lemma C.2 (Fitting a frozen target)**.** *For a stage beginning at $q = \pi_{b,0}$ and any integer $S \ge 1$,*

$$D_{\mathrm{KL}}(\widehat{\pi}_q \| \pi_{b,S}) \le \frac{\|\log\mu\|_2^2}{2\eta\mu_{min}(1 - \eta\mu_{max}/4)S}, \tag{43}$$

where $\|\log\mu\|_2^2 = \sum_{a=1}^K (\log\mu(a))^2$.

Proof. The frozen target is represented by the logits

$$\theta^* := \theta_{b,0} + \log\mu, \qquad \|\theta^* - \theta_{b,0}\|_2^2 = \|\log\mu\|_2^2. \tag{44}$$

Indeed, $\pi_{\theta^*} = \widehat{\pi}_q$, so θ^* maximizes $L(\pi_\theta; q)$ and its gradient vanishes there. Let $e_s = \theta^* - \theta_{b,s}$ and denote the surrogate gap by

$$\delta_s := L(\widehat{\pi}_q; q) - L(\pi_{b,s}; q) = J(\pi_{b,0})D_{\mathrm{KL}}(\widehat{\pi}_q \| \pi_{b,s}).$$

The surrogate is concave with a $J(\pi_{b,0})/2$-Lipschitz gradient by Lemma A.2 and Eq. (13). Co-coercivity therefore gives $\|g_{b,s}\|_2^2 \le (J(\pi_{b,0})/2)\langle e_s, g_{b,s}\rangle$. Using $e_{s+1} = e_s - \eta g_{b,s}$ and concavity,

we obtain

$$\begin{aligned}\|e_s\|_2^2 - \|e_{s+1}\|_2^2 &= 2\eta\langle e_s, g_{b,s}\rangle - \eta^2\|g_{b,s}\|_2^2 \\ &\geq 2\eta\left(1 - \frac{\eta J(\pi_{b,0})}{4}\right)\langle e_s, g_{b,s}\rangle \\ &\geq 2\eta\left(1 - \frac{\eta J(\pi_{b,0})}{4}\right)\delta_s.\end{aligned}$$

By Eq. (14), δ_s is nonincreasing. Summing over the inner loop and using $\|e_S\|_2^2 \geq 0$ gives

$$S\delta_S \leq \sum_{s=0}^{S-1}\delta_s \leq \frac{\|\log\mu\|_2^2}{2\eta(1-\eta J(\pi_{b,0})/4)}.$$

Dividing by $J(\pi_{b,0})S$ and using $\mu_{\min} \leq J(\pi_{b,0}) \leq \mu_{\max}$ proves the claim. □

Lemma C.3 (From fitting error to log-odds error)**.** *Let $\widehat{\pi}, p$ be positive probability vectors, $\alpha = \widehat{\pi}(1)$, and $0 < v \leq 1/2$. Writing $\mathrm{logit}(u) := \log(u/(1-u))$, we have*

$$D_{\mathrm{KL}}(\widehat{\pi}\|p) \leq \frac{1}{8}\min\{\alpha, 1-\alpha\}v^2 \quad \Longrightarrow \quad \mathrm{logit}(p(1)) \geq \mathrm{logit}(\alpha) - v. \tag{45}$$

Proof. If $u := p(1) \geq \alpha$, the conclusion is immediate. Otherwise, set $d = \mathrm{logit}(\alpha) - \mathrm{logit}(u) > 0$. Combining all actions other than action 1 by the log-sum inequality, we obtain

$$D_{\mathrm{KL}}(\widehat{\pi}\|p) \geq \alpha\log\frac{\alpha}{u} + (1-\alpha)\log\frac{1-\alpha}{1-u} = f(d),$$

where $f(t) := \alpha t + \log(1-\alpha+\alpha e^{-t})$. Here $f(0) = f'(0) = 0$, and f is strictly increasing for $t > 0$. For $0 \leq t \leq v \leq 1/2$,

$$f''(t) = \frac{\alpha(1-\alpha)e^{-t}}{(1-\alpha+\alpha e^{-t})^2} \geq \frac{\alpha(1-\alpha)}{2} \geq \frac{1}{4}\min\{\alpha, 1-\alpha\}.$$

Integrating twice gives $f(v) \geq \min\{\alpha, 1-\alpha\}v^2/8$. If $d > v$, then $f(d) > f(v)$, contradicting the assumed KL bound. Thus $d \leq v$, as required. □

Proof of the second part of Theorem 3.6. Set $h := 1 - \epsilon/(\mu_{\max} - \mu_{\min}) > 1/2$. Reaching $\pi_{b,0}(1) \geq h$ is sufficient for $D(\pi_{b,0}) \leq \epsilon$. We work with

$$0 < x \leq \frac{\mu_{\min}(1-h)}{\mu_{\max}} = \frac{\mu_{\min}\epsilon}{\mu_{\max}(\mu_{\max} - \mu_{\min})}. \tag{46}$$

This is a fixed positive upper bound and implies $x < h$.

Progress in one stage. Consider a stage beginning at $q = \pi_{b,0}$ with $x \leq q(1) < h$, and let $\alpha = \widehat{\pi}_q(1)$. Both parts of the target distribution have mass at least x:

$$\alpha = \frac{\mu_{\max}q(1)}{J(\pi_{b,0})} \geq q(1) \geq x, \qquad 1 - \alpha \geq \frac{\mu_{\min}(1-q(1))}{\mu_{\max}} > \frac{\mu_{\min}(1-h)}{\mu_{\max}} \geq x.$$

The exact target improves the optimal-action log-odds by

$$\begin{aligned}\mathrm{logit}(\alpha) - \mathrm{logit}(q(1)) &= \log\frac{\mu_{\max}(1-q(1))}{\sum_{a>1}q(a)\mu(a)} \\ &\geq \log\frac{\mu_{\max}}{\mu(2)} \geq \frac{\Delta}{\mu_{\max}}.\end{aligned} \tag{47}$$

To preserve at least half of this last lower bound after fitting, choose a fixed coefficient

$$c_S = \frac{16\mu_{\max}^2\|\log\mu\|_2^2}{\eta\mu_{\min}(1-\eta\mu_{\max}/4)\Delta^2}, \qquad S_x := \left\lceil\frac{c_S}{x}\right\rceil. \tag{48}$$

Use this same stage length throughout the run. For $v := \Delta/(2\mu_{\max}) \in (0, 1/2)$, Lemma C.2 gives

$$D_{\mathrm{KL}}(\widehat{\pi}_q \| \pi_{b,S_x}) \le \frac{x\Delta^2}{32\mu_{\max}^2} = \frac{xv^2}{8} \le \frac{1}{8}\min\{\alpha, 1-\alpha\}v^2.$$

Lemma C.3 and Eq. (47) now yield

$$\mathrm{logit}(\pi_{b+1,0}(1)) \ge \mathrm{logit}(\alpha) - \frac{\Delta}{2\mu_{\max}} \ge \mathrm{logit}(\pi_{b,0}(1)) + \frac{\Delta}{2\mu_{\max}}, \tag{49}$$

where $\pi_{b+1,0} = \pi_{b,S_x}$.

The number of stages. Let $b_* := \inf\{b \ge 0 : \pi_{b,0}(1) \ge h\}$. Until this threshold is reached, Eq. (49) shows that the optimal-action probability increases across stage endpoints and stays at least x. Induction therefore gives

$$\mathrm{logit}(\pi_{b,0}(1)) \ge \mathrm{logit}(x) + \frac{b\Delta}{2\mu_{\max}}$$

up to and including the first crossing, or for all b if no crossing occurs. Since Eq. (46) implies $x \le 1 - h$,

$$\mathrm{logit}(h) - \mathrm{logit}(x) = \log\frac{h(1-x)}{(1-h)x} \le 2\log(1/x).$$

Consequently, the threshold must be reached within

$$B_x := \left\lceil \frac{4\mu_{\max}}{\Delta}\log(1/x) \right\rceil \tag{50}$$

stages, otherwise the preceding lower bound at $b = B_x$ would already imply $\pi_{B_x,0}(1) \ge h$. At the first such endpoint,

$$D(\pi_{b_*,0}) \le (\mu_{\max} - \mu_{\min})(1 - \pi_{b_*,0}(1)) \le \epsilon.$$

Thus

$$T_\epsilon(S_x) \le b_* S_x \le B_x S_x = O\big(x^{-1}\log(1/x)\big).$$

Here $c_S > 0$ is fixed independently of x, so $S_x = \Theta(x^{-1})$. Combining this bound with the first part of the theorem gives

$$\frac{T_\epsilon(S_x)}{T_\epsilon(1)} \le Cx^{K-2}\log(1/x) \longrightarrow 0 \qquad \text{as } x \downarrow 0,$$

since $K \ge 3$. □