Stochastic Gradient Succeeds for Bandits

Jincheng Mei [*1] **Zixin Zhong** [*2] **Bo Dai** [1 3] **Alekh Agarwal** [4] **Csaba Szepesvari** [2] **Dale Schuurmans** [1 2]

Abstract

We show that the *stochastic gradient* bandit algorithm converges to a *globally optimal* policy at an $O(1/t)$ rate, even with a *constant* step size. Remarkably, global convergence of the stochastic gradient bandit algorithm has not been previously established, even though it is an old algorithm known to be applicable to bandits. The new result is achieved by establishing two novel technical findings: first, the noise of the stochastic updates in the gradient bandit algorithm satisfies a strong "growth condition" property, where the variance diminishes whenever progress becomes small, implying that additional noise control via diminishing step sizes is unnecessary; second, a form of "weak exploration" is automatically achieved through the stochastic gradient updates, since they prevent the action probabilities from decaying faster than $O(1/t)$, thus ensuring that every action is sampled infinitely often with probability 1. These two findings can be used to show that the stochastic gradient update is already "sufficient" for bandits in the sense that exploration versus exploitation is automatically balanced in a manner that ensures almost sure convergence to a global optimum. These novel theoretical findings are further verified by experimental results.

1. Introduction

Algorithms for multi-armed bandits (MABs) need to balance exploration and exploitation to achieve desirable performance properties (Lattimore & Szepesvári, 2020). Well known bandit algorithms generally introduce auxiliary mechanisms to control the exploration-exploitation trade-off. For example, the upper confidence bound strategies (UCB, Lai et al. (1985); Auer et al. (2002a)), manage exploration explicitly by designing auxiliary bonuses that induce optimism under uncertainty; Thompson sampling (Thompson, 1933; Agrawal & Goyal, 2012) maintains a posterior over rewards that has to be updated and sampled from tractably. The theoretical analysis of such algorithms typically focuses on bounding their regret, i.e., showing that the average reward obtained by the algorithm approaches that of the optimal action in hindsight, at a rate that is statistically optimal. However, managing exploration bonuses via uncertainty quantification is difficult in all but simple environments (Gawlikowski et al., 2021), while posterior updating and sampling can also be computationally difficult in practice, which make the UCB and Thompson sampling challenging to apply in real world scenarios.

Meanwhile, stochastic gradient-based techniques have witnessed widespread use across the breadth of machine learning. For bandit and reinforcement learning problems, stochastic gradient yields a particularly simple algorithm—the stochastic gradient bandit algorithm (Sutton & Barto, 2018, Section 2.8)—that omits any *explicit* control mechanism over *exploration*. This algorithm is naturally compatible with deep neural networks, in stark contrast to UCB and Thompson sampling, and has seen significant practical success (Schulman et al., 2015; 2017; Ouyang et al., 2022). Surprisingly, the theoretical understanding of this algorithm remains under-developed: the global convergence and regret properties of the stochastic gradient bandit algorithm is still open, which naturally raises the question:

Is the stochastic gradient bandit algorithm able to balance exploration vs. exploitation to identify an optimal action?

Such an understanding is paramount to further improving the underlying approach. In this paper, we take a step in this direction by studying the convergence properties of the canonical stochastic gradient bandit algorithm in the simplest setting of multi-armed bandits, and answer the question affirmatively: the distribution over the arms maintained by this algorithm almost surely concentrates asymptotically on a globally optimal action.

Of course, the broader literature on the use of stochastic gradient techniques in reinforcement learning has a long tradition, dating back to stochastic approximation (Robbins & Monro, 1951) with likelihood ratios ("log trick") (Glynn, 1990) and REINFORCE policy gradient estima-

[*]Equal contribution [1]Google Research, Brain Team [2]University of Alberta [3]Georgia Tech [4]Google Research. Correspondence to: Jincheng Mei <jcmei@google.com>, Zixin Zhong <zzhong10@ualberta.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Reference	Convergence rate	Learning rate η	Remarks
Zhang et al. (2020a)	$\tilde{O}(1/\sqrt{t})$	$\tilde{O}(1/\sqrt{t})$	Log-barrier regularization
Ding et al. (2021)	$O(1/\sqrt{t})$	$O(1/t)$	Entropy regularization
Zhang et al. (2021)	$O(1/\sqrt{t})$	$O(1/\log t)$	Gradient truncation + variance reduction
Yuan et al. (2022)	$\tilde{O}(1/t^{1/3})$	$O(1/t)$	ABC assumptions
Mei et al. (2022)	$O(C_\Delta/t)$	$O(1/t)$	Oracle baseline, C_Δ is initialization and problem dependent
This paper	$\boldsymbol{O(C_\Delta/t)}$	$\boldsymbol{O(1)}$	C_Δ **is initialization and problem dependent**

Table 1. Global convergence results for gradient based methods. To simplify comparison, regret or sample complexity bounds have been converted to convergence rates through the usual translations. (This comparison does not capture differences in other metrics like regret.) We also compare the learning rates used in the underlying algorithm, since the ability to use a constant learning rate is a key insight of the analysis presented and is generally preferred in practice. Note that Zhang et al. (2021) claimed a constant learning rate; however, their learning rate must follow a decaying "truncation threshold", which means the actual learning rate in (Zhang et al., 2021) is decaying.

tion (Williams, 1992). It is well known that REINFORCE (Williams, 1992) with on-line Monte Carlo sampling provides an unbiased gradient estimator of bounded variance, which is sufficient to guarantee convergence to a stationary point if the learning rates are decayed appropriately (Robbins & Monro, 1951; Zhang et al., 2020b). However, convergence to a stationary point is an extremely weak guarantee for bandits, since any deterministic policy, whether optimal or sub-optimal, has a zero gradient and is therefore a stationary point. Convergence to a stationary point, on its own, is insufficient to ensure convergence to a globally optimal policy or even to establish regret bounds.

Recently, it has been shown that if true gradients are used, softmax policy gradient (PG) methods converge to a globally optimal policy asymptotically (Agarwal et al., 2021), with an $O(1/t)$ rate (Mei et al., 2020b), albeit with initialization and problem dependent constants (Mei et al., 2020a; Li et al., 2021). However, these results do not apply to the gradient bandit algorithm as it uses *stochastic* gradients, and a key theoretical challenge is to account for the effects of stochasticity from on-policy sampling and reward noise.

More recent results on the global convergence of PG methods with *stochastic* updates have been established, as summarized in Table 1. In particular, Zhang et al. (2020b) showed that REINFORCE (Williams, 1992) with $O(1/\sqrt{t})$ learning rates and log-barrier regularization has $\tilde{O}(1/\sqrt{t})$ average regret. Ding et al. (2021) proved that softmax PG with $O(1/t)$ learning rates and entropy regularization has $\tilde{O}(\epsilon^{-2})$ sample complexity. Zhang et al. (2021) showed that with gradient truncation softmax PG gives $O(\epsilon^{-2})$ sample complexity. Under extra assumptions, Yuan et al. (2022) obtained $\tilde{O}(\epsilon^{-3})$ sample complexity with $O(1/t)$ learning rates. Mei et al. (2022) analyzed on-policy natural PG with value baselines and $O(1/t)$ learning rates and proved $O(1/t)$ convergence rate. The results in these works are expressed in different metrics, such as average regret, sample complexity, or convergence rate, which can sometimes make comparisons difficult. However, for the bandit case, where only one example is used in each iteration, these metrics become comparable, such that $O(\epsilon^{-n})$ sample complexity is equal to $O(t^{-1/n})$ convergence rate, which is stronger than $O(t^{-1/n})$ average regret, but not necessarily vice versa.[1]

The two key shortcomings in these existing results are, **first**, they introduce decaying learning rates (or regularization and/or variance reduction) to explicitly control noise, and **second**, such auxiliary techniques generally incur additional computation and decelerate convergence to an $O(1/\sqrt{t})$ or slower rate. The only exception to the latter is Mei et al. (2022), which considers an aggressive $O(1/t)$ learning rate decay and still establishes $O(1/t)$ convergence, but leverages an unrealistic baseline to achieve this.

In this paper, we provide a new global convergence analysis of stochastic gradient bandit algorithms *with constant learning rates* by establishing novel properties and techniques. There are two main benefits to the results presented.

- By considering only constant learning rates, we show that auxiliary forms of noise control such as learning rate decay, regularization and variance reduction are *unnecessary* to achieve global convergence, which justifies the use of far simpler algorithms in practice.
- Unlike previous work, we show that a *practical* and general algorithm can achieve an optimal $O(1/t)$ convergence rate and $O(\log T)$ regret asymptotically.

The remaining paper is organized as follows. Sections 2 and 3 introduce the gradient bandit algorithms and standard stochastic gradient analysis respectively. Section 4 presents our novel technical findings, characterizing the automatic noise cancellation effect and global landscape properties, which is then leveraged in Section 5 to establish novel global convergence results. Section 6 discusses the effect of using

[1] It is worth noting that these works also contain results for general Markov decision processes (MDPs). We express their results for bandits here by treating this case as a single state MDP.

baselines. Section 7 presents a simulation study to verify the theoretical findings, and Section 8 provides further discussions. Section 9 briefly concludes this work.

2. Gradient Bandit Algorithms

A multi-armed bandit (MAB) problem is specified by an action set $[K] \coloneqq \{1, 2, \cdots, K\}$ and random rewards with mean vector $r \in \mathbb{R}^K$. For each action $a \in [K]$, the mean reward $r(a)$ is the expectation of a bounded reward distribution,

$$r(a) = \int_{-R_{\max}}^{R_{\max}} x \cdot P_a(x)\mu(dx), \tag{1}$$

where μ is a finite measure over $[-R_{\max}, R_{\max}]$, $P_a(x) \geq 0$ is a probability density function with respect to μ, and $R_{\max} > 0$ is the reward range. Since the sampled reward is bounded, we also have $r \in [-R_{\max}, R_{\max}]^K$. We make the following assumption on r.

Assumption 2.1 (True mean reward has no ties)**.** For all $i, j \in [K]$, if $i \neq j$, then $r(i) \neq r(j)$.

Remark 2.2. Assumption 2.1 is used in the proofs for Theorem 5.1. In particular, "convergence toward strict one-hot policies" above Theorem 5.1 is needed as a result of Assumption 2.1. We discuss intuition later for establishing the same result without Assumption 2.1.

According to Sutton & Barto (2018, Section 2.8), the gradient bandit algorithm maintains a softmax distribution over actions $\Pr(a_t = a) = \pi_{\theta_t}(a)$ such that $\pi_{\theta_t} = \mathrm{softmax}(\theta_t)$, where

$$\pi_{\theta_t}(a) = \frac{\exp\{\theta_t(a)\}}{\sum_{a' \in [K]} \exp\{\theta_t(a')\}}, \quad \text{for all } a \in [K], \tag{2}$$

and $\theta_t \in \mathbb{R}^K$ is the parameter vector to be updated.

Algorithm 1 Gradient bandit algorithm (without baselines)

Input: initial parameters $\theta_1 \in \mathbb{R}^K$, learning rate $\eta > 0$.
Output: policies $\pi_{\theta_t} = \mathrm{softmax}(\theta_t)$.
while $t \geq 1$ **do**
  Sample one action $a_t \sim \pi_{\theta_t}(\cdot)$.
  Observe one reward sample $R_t(a_t) \sim P_{a_t}$.
  for all $a \in [K]$ **do**
    if $a = a_t$ **then**
      $\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot (1 - \pi_{\theta_t}(a)) \cdot R_t(a_t)$.
    else
      $\theta_{t+1}(a) \leftarrow \theta_t(a) - \eta \cdot \pi_{\theta_t}(a) \cdot R_t(a_t)$.
    end if
  end for
end while

It is obvious that Algorithm 1 is an instance of stochastic gradient ascent with an unbiased gradient estimator (Nemirovski et al., 2009), as shown below for completeness.

Proposition 2.3. *Algorithm 1 is equivalent to the following stochastic gradient ascent update on $\pi_\theta^\top r$,*

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \tag{3}$$

$$= \theta_t + \eta \cdot \left(diag(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right) \hat{r}_t, \tag{4}$$

where $\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right] = \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$, and $\left(\frac{d\pi_\theta}{d\theta}\right)^\top = diag(\pi_\theta) - \pi_\theta\pi_\theta^\top$ is the Jacobian of $\theta \mapsto \pi_\theta \coloneqq \mathrm{softmax}(\theta)$, $\hat{r}_t(a) \coloneqq \frac{\mathbb{I}\{a_t = a\}}{\pi_{\theta_t}(a)} \cdot R_t(a)$ for all $a \in [K]$ is the importance sampling (IS) estimator, and we set $R_t(a) = 0$ for all $a \neq a_t$.

Based on Proposition 2.3, Sutton & Barto (2018, Section 2.8) assert that Algorithm 1 has "robust convergence properties" toward stationary points without rigorous justification. However, as mentioned in Section 1, convergence to stationary points is a very weak assertion in a MAB, since this does not guarantee sub-optimal local maxima are avoided. Hence this claim does not assure global convergence or sub-linear regret for the gradient bandit algorithm.

3. Preliminary Stochastic Gradient Analysis

In this section, we start with local convergence of stochastic gradient bandit algorithm. The understanding of the behavior of the algorithm involves assessing whether optimization progress is able to overcome the effects of the sampling noise. This trade-off reveals inability of the vanilla analysis and inspires our refined analysis.

To illustrate the basic ideas, we first recall some known results about the form of $\pi_\theta^\top r$ and the behavior of Algorithm 1 and make a preliminary attempt to establish convergence to a stationary point. **First**, $\pi_\theta^\top r$ is a $5/2$-smooth function of $\theta \in \mathbb{R}^K$ (Mei et al., 2020b, Lemma 2), which implies that,

$$\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r \leq -\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{5}{4} \cdot \|\theta_{t+1} - \theta_t\|_2^2$$

$$= -\eta \cdot \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\rangle + \frac{5}{4} \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2^2,$$

where the last equation follows from Eq. (3). **Second**, as is well known, the on-policy stochastic gradient is unbiased, and its variance / scale is uniformly bounded over all π_θ.

Proposition 3.1 (Unbiased stochastic gradient with bounded variance / scale)**.** *Using Algorithm 1, we have, for all $t \geq 1$,*

$$\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right] = \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \text{ and } \mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \leq 2\, R_{\max}^2,$$

where $\mathbb{E}_t[\cdot]$ is on randomness from the on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$ and reward sampling $R_t(a_t) \sim P_{a_t}$.

Therefore, according to Proposition 3.1, we have,

$$\pi_{\theta_t}^\top r - \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] \le -\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 + \frac{5\, R_{\max}^2}{2} \cdot \eta^2. \quad (5)$$

Since the goal is to maximize $\pi_\theta^\top r$, we want the first term on the r.h.s. of Eq. (5) ("progress") to overcome the second term ("noise") to ensure that $\mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] > \pi_{\theta_t}^\top r$. Unfortunately, this is not achievable using a constant learning rate $\eta \in \Theta(1)$, since the "progress" contains a vanishing term of $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \to 0$ as $t \to \infty$ while the "noise" term will remain at constant level. Therefore, based on this bound, it seems necessary to use a diminishing learning rate $\eta \to 0$ to control the effect of noise for local convergence. In fact, with appropriate learning rate control (Robbins & Monro, 1951; Ghadimi & Lan, 2013; Zhang et al., 2020b), it can be shown that minimum gradient norm converge to zero. From Eq. (5), by algebra and telescoping, we have,

$$\min_{1\le t\le T} \mathbb{E}\left[\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \right] \le \frac{\mathbb{E}[\pi_{\theta_{T+1}}^\top r] - \mathbb{E}[\pi_{\theta_1}^\top r]}{\sum_{t=1}^T \eta_t} + \frac{5\, R_{\max}^2}{2} \frac{\sum_{t=1}^T \eta_t^2}{\sum_{t=1}^T \eta_t}.$$

Choosing $\eta_t \in \Theta(1/\sqrt{t})$, the r.h.s. of the above inequality is in $\tilde{O}(1/\sqrt{T})$, i.e., the minimum gradient norm approaches zero as $T \to \infty$ (Ghadimi & Lan, 2013; Zhang et al., 2020b). However, the decaying learning rate will slow down the convergence as is seen. Next, we will present our novel technical characterization of the noise in stochastic gradient bandit that can allow us to avoid this choice.

4. New Analysis: Noise Vanishes Automatically

As discussed in Section 3, noise control is at the heart of standard stochastic gradient analysis, and different techniques (entropy or log-barrier regularization, learning rate schemes, momentum) are used to explicitly combat noise in stochastic updates. Here we take a different perspective by asking whether the sampling noise will automatically diminish in a way that there is no need to explicitly control it. In particular, we investigate whether the constant order of the second term in Eq. (5) ("noise") is accurately characterizing the sampling noise, or whether this bound can be improved.

Note that the "noise" constant $5\, R_{\max}^2$ in Eq. (5) arises from two quantities: the standard smoothness constant of $5/2$, and the variance upper bound of $2\, R_{\max}^2$ in Proposition 3.1. It turns out that both of these quantities can be improved.

4.1. Non-uniform Smoothness: Landscape Properties

The first key observation is a landscape property originally derived in (Mei et al., 2021b) for true policy gradient settings, which is also applicable for stochastic gradient update.

Lemma 4.1 (Non-uniform smoothness (NS), Lemma 2 in (Mei et al., 2021b))**.** *For all $\theta \in \mathbb{R}^K$, and for all $r \in \mathbb{R}^K$, the spectral radius of the Hessian matrix $\frac{d^2\{\pi_\theta^\top r\}}{d\theta^2} \in \mathbb{R}^{K\times K}$ is upper bounded by $3 \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2$, i.e., for all $y \in \mathbb{R}^K$,*

$$\left| y^\top \frac{d^2\{\pi_\theta^\top r\}}{d\theta^2}\, y \right| \le 3 \cdot \left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \cdot \|y\|_2^2. \quad (6)$$

It is useful to understand the intuition behind this lemma. Note that when the PG norm $\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2$ is small, the policy π_θ is close to a one-hot policy, and the objective $\pi_\theta^\top r$ has a flat local landscape; ultimately implying that the the Hessian magnitude is upper bounded by the gradient.

However, directly using Lemma 4.1 remains challenging: Consider two iterates θ_t and θ_{t+1}. Then in a Taylor expansion the PG norm of an intermediate point $\theta_\zeta := \theta_t + \zeta \cdot (\theta_{t+1} - \theta_t)$, $\zeta \in [0,1]$, will appear, which is undesirable since ζ is unknown. Therefore, we require an additional lemma to assert that for a sufficiently small learning rate, the PG norm of θ_ζ will be controlled by that of θ_t.

Lemma 4.2 (NS between iterates)**.** *Using Algorithm 1 with $\eta \in \left(0, \frac{2}{9\cdot R_{\max}}\right)$, we have, for all $t \ge 1$,*

$$D(\theta_{t+1}, \theta_t) \le \frac{\beta(\theta_t)}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \quad (7)$$

where $D(\theta_{t+1}, \theta_t)$ is Bregman divergence defined as,

$$D(\theta_{t+1}, \theta_t) := \left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right|,$$

and $\beta(\theta_t) = \frac{6}{2 - 9\cdot R_{\max}\cdot\eta} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2$.

With the learning rate requirement, Lemma 4.2 is no longer only a landscape property, but also depends on updates. Using Lemma 4.2 rather than standard smoothness, $5\, R_{\max}^2$ in Eq. (5) can be replaced with $\beta(\theta_t)$, which implies that the "noise" is also vanishing since $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \to 0$ as $t \to \infty$. However, with simple constant upper bound of the variance of noise in Proposition 3.1, the progress term in Eq. (5) still decays faster than the noise term since $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \in o\left(\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \right)$. Unfortunately, this suggests that a decaying learning rate is still necessary to control the noise. Therefore, a further refined analysis of the noise variance is required.

4.2. Growth Conditions: Softmax Jacobian Behavior

Since only Lemmas 4.1 and 4.2 are still insufficient to guarantee progress without learning rate control, we need to develop a more refined variance bound of the noise that was previously unknown for gradient bandit algorithms. We first consider an example to intuitively explain why a tighter bound on the noise variance might be possible.

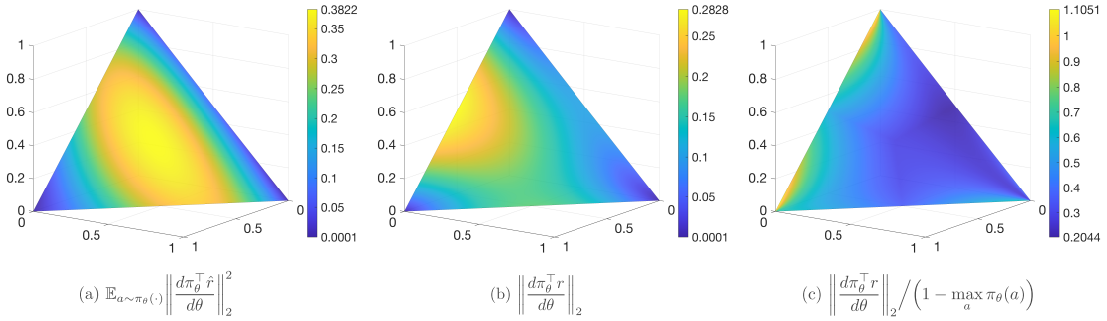


Figure 1. Visualization and intuition for Lemma 4.3. **(a)** Stochastic gradient scale. **(b)** True gradient norm. **(c)** The ratio of true gradient norm over 1 minus largest action probability. Color bars contain minimum and maximum values of corresponding quantities.

Illustration. Consider Figure 1(a), which depicts the probability simplex containing all policies π_θ over $K = 3$ actions. Figure 1(a) shows the scale of the stochastic gradient $\mathbb{E}_{a\sim\pi_\theta(\cdot)}\left\|\frac{d\pi_\theta^\top \hat{r}}{d\theta}\right\|_2^2$, illustrating that when π_θ is close to any corner of the simplex, the stochastic gradient scale becomes close to 0. This suggests that the $2\,R_{\max}^2$ in Proposition 3.1 is quite loose and improvable. Figure 1(b) presents a similar visualization for the true gradient norm $\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2$, demonstrating a similar behavior to the stochastic gradient scale.

We formalize this observation by proving that the stochastic gradient scale is controlled by the true gradient norm, significantly improving Proposition 3.1.

Lemma 4.3 (Strong growth condition; self-bounding noise property)**.** *Using Algorithm 1, we have, for all $t \geq 1$,*

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \leq \frac{8\cdot R_{\max}^3\cdot K^{3/2}}{\Delta^2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2, \tag{8}$$

where $\Delta \coloneqq \min_{i\neq j}|r(i)-r(j)|$.

The proof sketch of Lemma 4.3 is as follows. For any $t \geq 1$, let k_t denote the action with the largest probability, i.e.,

$$k_t \coloneqq \arg\max_{a\in[K]} \pi_{\theta_t}(a). \tag{9}$$

Note that $1-\pi_{\theta_t}(k_t)$ characterizes how close π_{θ_t} is to any corner of the probability simplex. We first prove that,

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \leq 4\cdot R_{\max}^2\cdot\left(1-\pi_{\theta_t}(k_t)\right), \tag{10}$$

which formalizes the observation in Figure 1(a). Additionally, Figure 1(c) shows that $\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \Big/ \left(1-\pi_{\theta_t}(k_t)\right)$ is larger than about 0.2044, which suggests that the true gradient norm also characterizes a distance of π_θ from any corner of probability simplex since it has a "variance-like" structure (Eq. (14)), which is formalized by proving that,

$$1-\pi_{\theta_t}(k_t) \leq \frac{2\cdot R_{\max}\cdot K^{3/2}}{\Delta^2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2. \tag{11}$$

Combining Eqs. (10) and (11) allows one to establish Lemma 4.3, verifying the intuitive observation in Figure 1.

From this explanation, whenever the true PG norm $\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2$ is small and π_θ is close to a one-hot policy, the stochastic PG norm $\left\|\frac{d\pi_\theta^\top \hat{r}}{d\theta}\right\|_2$ in Eq. (3) will also be small. A deeper explanation is that the softmax Jacobian $\mathrm{diag}(\pi_\theta)-\pi_\theta\pi_\theta^\top$ is involved in the stochastic PG in Eq. (4), which cancels and annihilates the unbounded noise in reward estimator $\hat{r}$ arising from the use of importance sampling.

Remark 4.4. The "strong growth condition" was first proposed by Schmidt & Roux (2013) and later found to be satisfied in supervised learning with over-parameterized neural networks (NNs) (Allen-Zhu et al., 2019). There, given a dataset $\mathcal{D} \coloneqq \{(x_1,y_1),(x_2,y_2),\cdots(x_N,y_N)\}$, the goal is to minimize a composite loss function,

$$\min_w f(w) = \min_w \sum_{i\in[N]} f_i(w). \tag{12}$$

Since over-parameterized NNs fit all the data points, i.e., $f(w)=0$ for some w, each individual loss is also $f_i(w)=0$ since $f_i(w)\geq 0$ for all $i\in[N]$ (e.g., squared loss and cross entropy). This guarantees that when the true gradient $\nabla f(w)=\mathbf{0}$, the stochastic gradient $\nabla f_i(w)=\mathbf{0}$ for all $i\in[N]$. Hence the strong growth conditions are satisfied, and stochastic gradient descent (SGD) attains the same convergence speed as gradient descent (GD) with an over-parameterized NN (Allen-Zhu et al., 2019).

Remark 4.5. In probability and statistics, the property of a variance bounded by the expectation is also called a self-bounding property, which can be used to get strong variance bounds (McDiarmid & Reed, 2006; Boucheron et al., 2009).

Lemma 4.3 proves that such strong growth conditions are also satisfied in the stochastic gradient bandit algorithm, but

for a different reason. For over-parameterized NNs, since the model fits every data point, zero gradient implies that the stochastic gradient is also zero. Here, in Lemma 4.3, the strong growth condition alternatively arises because of the landscape; that is, the presence of the softmax Jacobian in the stochastic gradient update annihilates the sampling noise and leads to the strong growth condition being satisfied.

4.3. No Learning Rate Decay

Finally, from Lemmas 4.2 and 4.3 we reach the result that expected progress can be guaranteed with a *constant* learning rate for the stochastic gradient bandit update.

Lemma 4.6 (Constant learning rates)**.** *Using Algorithm 1 with* $\eta = \frac{\Delta^2}{40 \cdot K^{3/2} \cdot R_{\max}^3}$*, we have, for all* $t \geq 1$*,*

$$\pi_{\theta_t}^\top r - \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] \leq -\frac{\Delta^2}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2.$$

Lemma 4.2 indicates that $\{\pi_{\theta_t}^\top r\}_{t \geq 1}$ is a sub-martingale. Since the reward is bounded $r \in [-R_{\max}, R_{\max}]$, by Doob's super-martingale convergence theorem we have that, almost surely, $\pi_{\theta_t}^\top r \to c \in [-R_{\max}, R_{\max}]$ as $t \to \infty$.

Corollary 4.7. *Using Algorithm 1, we have, the sequence* $\{\pi_{\theta_t}^\top r\}_{t \geq 1}$ *converges w.p.* 1.

Therefore, from Lemma 4.6 and Corollary 4.7 it follows that $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \to 0$ as $t \to 0$ almost surely, which implies that convergence to a stationary point is achieved without a decaying learning rate (Robbins & Monro, 1951). From Lemma 4.6, using telescoping we immediately have,

$$\frac{1}{T} \cdot \sum_{1 \leq t \leq T} \mathbb{E}\left[\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \right] \leq \frac{\mathbb{E}[\pi_{\theta_{T+1}}^\top r] - \mathbb{E}[\pi_{\theta_1}^\top r]}{C \cdot T}, \quad (13)$$

where $C \coloneqq \frac{\Delta^2}{80 \cdot K^{3/2} \cdot R_{\max}^3}$. Comparing to Section 3, Eq. (13) is a stronger result of averaged gradient norm approaches zero, in terms of faster $O(1/T)$ rate and constant learning rate. An interesting observation is that the average gradient convergence rate is independent with the initialization, which is different with the global convergence results in later sections. The key reason behind this outcome is that Lemmas 4.1 and 4.3 establish that the "noise" in Eq. (5) decays on the same order as the "progress" $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2$, so that a constant learning rate is sufficient for the "progress" term to overcome the "noise" term (Lemma 4.6).

5. New Global Convergence Results

Given the refined stochastic analysis from Section 4 we are now ready to establish new global convergence results for the stochastic gradient bandit in Algorithm 1.

5.1. Asymptotic Global Convergence

First, note that the true gradient norm takes the following "variance-like" expression,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2^2 = \sum_{a \in [K]} \pi_\theta(a)^2 \cdot \left(r(a) - \pi_\theta^\top r \right)^2. \quad (14)$$

According to $\left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \to 0$ as $t \to 0$, we have that π_{θ_t} approaches a one-hot policy, i.e., $\pi_{\theta_t}(i) \to 1$ for some $i \in [K]$ as $t \to \infty$. Asymptotic global convergence is then proved by constructing contradictions against the assumption that the algorithm converges to a sub-optimal one-hot policy.

Theorem 5.1 (Asymptotic global convergence)**.** *Using Algorithm 1, we have, almost surely,*

$$\pi_{\theta_t}(a^*) \to 1, \text{ as } t \to \infty, \quad (15)$$

which implies that $\inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$*.*

It is highly challenging to prove almost surely global convergence because **(i)** the iteration $\{\theta_t\}_{t \geq 1}$ is a stochastic process, which it different with the true gradient settings (Agarwal et al., 2021), and **(ii)** the iteration $\theta_t \in \mathbb{R}^K$ is unbounded, which makes Doob's super-martingale convergence results not applicable, unlike Corollary 4.7. The strategy and insights of Theorem 5.1 are as follows. According to Proposition 3.1, we have, for all $a \in [K]$,

$$\mathbb{E}_t[\theta_{t+1}(a)] = \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \left(r(a) - \pi_{\theta_t}^\top r \right). \quad (16)$$

Now we suppose that using Algorithm 1, there exists a sub-optimal action $i \in [K]$, $r(i) < r(a^*)$, such that,

$$\pi_{\theta_t}(i) \to 1, \text{ as } t \to \infty, \quad (17)$$

which implies that,

$$\pi_{\theta_t}^\top r \to r(i), \text{ as } t \to \infty. \quad (18)$$

Since $r(i) < r(a^*)$, there exists a "good" action set,

$$\mathcal{A}^+(i) \coloneqq \left\{ a^+ \in [K] : r(a^+) > r(i) \right\}. \quad (19)$$

By Eqs. (16), (18) and (19), for all large enough $t \geq 1$,

$$\mathbb{E}_t[\theta_{t+1}(a^+)] \geq \theta_t(a^+), \quad (20)$$

which means a "good" action's parameter $\{\theta_t(a^+)\}_{t \geq 1}$ is a sub-martingale. The major part of the proofs are devoted to the following key results. We have, almost surely,

$$\inf_{t \geq 1} \theta_t(a^+) \geq c_1 > -\infty, \text{ and} \quad (21)$$

$$\sup_{t \geq 1} \theta_t(i) \leq c_2 < \infty. \quad (22)$$

Eq. (21) is non-trivial since an unbounded sub-martingale is not necessarily lower bounded and could have positive probability of approaching negative infinity[2], while Eq. (22) is non-trivial since the behavior of $\theta_t(i)$ depends on different cases of how many times "good" actions are sampled as $t \to \infty$. With the above results, we have,

$$\begin{aligned}\pi_{\theta_t}(i) &< \frac{e^{\theta_t(i)}}{e^{\theta_t(i)} + \sum_{a^+ \in \mathcal{A}^+(i)} e^{\theta_t(a^+)}} \qquad \left(e^{\theta_t(a^-)} > 0\right) \\ &\le \frac{e^{\theta_t(i)}}{e^{\theta_t(i)} + e^{c_1}} \qquad \text{(by Eq. (21))} \\ &\le \frac{e^{c_2}}{e^{c_2} + e^{c_1}} \qquad \text{(by Eq. (22))} \\ &\not\to 1,\end{aligned}$$

which is a contradiction with the assumption of Eq. (17). Therefore, the asymptotically convergent one-hot policy has to satisfy $r(i) = r(a^*)$, proving Theorem 5.1. The detailed proof is provided in Appendix A.1.

Remark 5.2. As mentioned in Remark 2.2, the arguments above Theorem 5.1, i.e., π_{θ_t} approaches a one-hot policy, is based on Assumption 2.1. With this result, Theorem 5.1 proves asymptotic global convergence by contradiction with the assumption of Eq. (17). In general, without Assumption 2.1, Eq. (14) approaches zero can only imply π_{θ_t} approaches a "generalized one-hot policy" (rather than a strict one-hot policy). The definition of generalized one-hot policies can be found in Eq. (141).

Remark 5.3. It is true that Eq. (14) approaches zero is not enough for showing π_{θ_t} approaches a one-hot policy. However, Algorithm 1 is special that it is always making one action's probability dominate others' when there are ties. Consider $r \in \mathbb{R}^K$ with $r(1) = r(2)$. If $\pi_{\theta_t}(1) > \pi_{\theta_t}(2)$, then using the expected softmax PG update $\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)$, we have,

$$\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(2)} = \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(2)} \cdot e^{\eta \cdot (\pi_{\theta_t}(1) - \pi_{\theta_t}(2)) \cdot (r(1) - \pi_{\theta_t}^\top r)} > \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(2)},$$

which means that after one update $\pi_{\theta_{t+1}}(1)$ will be even larger than $\pi_{\theta_{t+1}}(2)$. Therefore, we have,

$$\theta_{t+1}(1) - \theta_{t+1}(2) > \theta_t(1) - \theta_t(2) + C/t, \tag{23}$$

for some $C > 0$, which implies that,

$$\lim_{t\to\infty} \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(2)} = \lim_{t\to\infty} e^{\theta_t(1) - \theta_t(2)} = \infty, \tag{24}$$

i.e., $\pi_{\theta_t}(1) \to 1$ as $t \to \infty$. The above arguments illustrate the "self-reinforcing" nature of Algorithm 1, such that whenever two (or more) actions have the same mean reward, the update will make only one one of their probabilities larger and larger, until one eventually dominates the others as $t \to \infty$. Generalizing the arguments to stochastic updates will remove Assumption 2.1 in the proofs for Theorem 5.1.

[2] Consider a random walk, $Y_{t+1} = Y_t + Z_t$, where $Z_t = \pm 1$ with equal chance of $1/2$. We have $\mathbb{E}_t[Y_{t+1}] \ge Y_t$. However, we also have $\liminf_{t\to\infty} Y_t = -\infty$, and with positive probability, $\inf_{t\ge 1} Y_t$ is not lower bounded.

5.2. Convergence Rate

Given Theorem 5.1, a convergence rate result can then be proved using the following inequality (Mei et al., 2020b).

Lemma 5.4 (Non-uniform Łojasiewicz (NŁ), Lemma 3 of Mei et al. (2020b))**.** *Assume r has a unique maximizing action a^*. Let $\pi^* = \arg\max_{\pi\in\Delta} \pi^\top r$. Then,*

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \ge \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r\,. \tag{25}$$

Theorem 5.5 (Convergence rate and regret)**.** *Using Algorithm 1 with $\eta = \frac{\Delta^2}{40 \cdot K^{3/2} \cdot R_{\max}^3}$, we have, for all $t \ge 1$,*

$$\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \le \frac{C}{t}, \quad \textit{and}$$

$$\mathbb{E}\left[\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r\right] \le \min\{\sqrt{2\, R_{\max}\, C\, T}, C\, \log T + 1\},$$

where $C := \frac{80 \cdot K^{3/2} \cdot R_{\max}^3}{\Delta^2 \cdot \mathbb{E}[c^2]}$, and $c := \inf_{t\ge 1} \pi_{\theta_t}(a^*) > 0$ is from Theorem 5.1.

In Theorem 5.5, the dependence on t is optimal (Lai et al., 1985). However, the constant can be large, especially for a bad initialization. In short, the stochastic gradient algorithm inherits the initialization sensitivity and sub-optimal plateaus from the true policy gradient algorithm with softmax parameterization (Mei et al., 2020a). The detailed proof of Theorem 5.5 is elaborated in Appendix A.2.

Theorem 5.6. *There exists a problem, initialization $\theta_1 \in \mathbb{R}^K$, and a positive constant $C > 0$, such that, for all $t \le t_0 := \frac{C}{\delta \cdot \pi_{\theta_1}(a^*)}$, we have*

$$\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \ge 0.9 \cdot \Delta, \tag{26}$$

where $\Delta := r(a^*) - \max_{a\ne a^*} r(a)$ is the reward gap of r.

6. The Effect of Baselines

The original gradient bandit algorithm (Sutton & Barto, 2018) uses a baseline, which is a slightly modification of Algorithm 1. The difference is that $R_t(a_t)$ in Algorithm 1 is replaced with $R_t(a_t) - B_t$, where $B_t \in \mathbb{R}$ is an action independent baseline, as shown in Algorithm 2 in Appendix B.

It is well known that action independent baselines do not introduce bias in the gradient estimate (Sutton & Barto, 2018). The utility of adding a baseline has typically been considered to be reducing the variance of the gradient estimates (Greensmith et al., 2004; Bhatnagar et al., 2007; Tucker

et al., 2018; Mao et al., 2018; Wu et al., 2018). Here we show that a similar effect manifests itself through improvements in the strong growth condition.

Lemma 6.1 (Strong growth condition, Self-bounding noise property). *Using Algorithm 2, we have, for all $t \geq 1$,*

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top(\hat{r}_t - \hat{b}_t)}{d\theta_t}\right\|_2^2\right] \leq \frac{8\,\bar{R}_{\max}^2\,R_{\max}\,K^{3/2}}{\Delta^2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2,$$

where $\Delta := \min_{i\neq j}|r(i)-r(j)|$, $\hat{b}_t(a) := \frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)}\cdot B_t$ for all $a \in [K]$, and $R_t(a_t) - B_t \in [-\bar{R}_{\max}, \bar{R}_{\max}]$.

Note that $\bar{R}_{\max}$ denotes the range of $R_t(a_t) - B_t$ after minus a baseline from sampled reward. Comparing Lemma 6.1 to Lemma 4.3 shows that the only difference is that a constant factor of $R_{\max}^2$ is changed to $\bar{R}_{\max}^2$. This indicates that a deeper reason for the variance reduction effect of adding a baseline is to reduce the effective reward range. The same improved constant will carry over to all the similar results, including larger constant learning rates, larger progress, and better constants in the global convergence results.

It is worth noting that the effect of baseline differs between algorithms. Here we see that without any baseline Algorithm 1 already achieves global convergence, while adding a baseline provides constant improvements. For a different natural policy gradient (NPG) method (Kakade, 2002; Agarwal et al., 2021), it is known that without using baselines, on-policy NPG can fail by converging to a sub-optimal deterministic policy (Chung et al., 2020; Mei et al., 2021a), while adding a value baseline $\pi_{\theta_t}^\top r$ restores a guarantee of global convergence by reducing the update aggressiveness.

7. Simulation Results

In this section, we conduct several simulations to empirically verify the theoretical findings of asymptotic global convergence and convergence rate.

7.1. Asymptotic Global Convergence

We first design experiments to justify the asymptotic global convergence. We run Algorithm 1 on stochastic bandit problems with $K = 10$ actions. The mean reward r is random generated in $(0,1)^K$. For each sampled action $a_t \sim \pi_{\theta_t}(\cdot)$, the observed reward is generated as $R_t(a_t) = r(a_t) + Z_t$, where $Z_t \sim \mathcal{N}(0,1)$ is Gaussian noise. For the baseline in Algorithm 2, we use average reward as suggested in (Sutton & Barto, 2018), i.e., $B_t := \sum_{s=1}^{t-1} R_s(a_s)/(t-1)$ for all $t > 1$. The learning rate is $\eta = 0.01$. We use adversarial initialization, such that $\pi_{\theta_1}(a^*) < 1/K$.

As shown in Figure 2, $\pi_{\theta_t}(a^*) \to 1$ eventually, even if its initial value $\pi_{\theta_1}(a^*)$ is very small, verifying the asymptotic global convergence in Theorem 5.1. On the other hand, the long plateaus observed in Figure 2 verify Theorem 5.6.

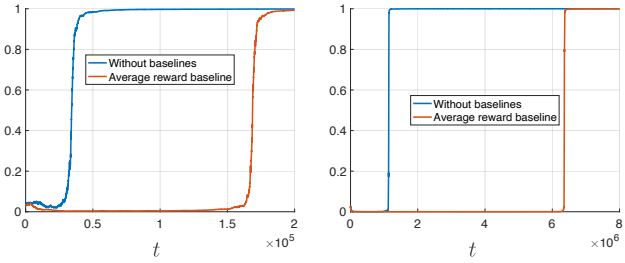
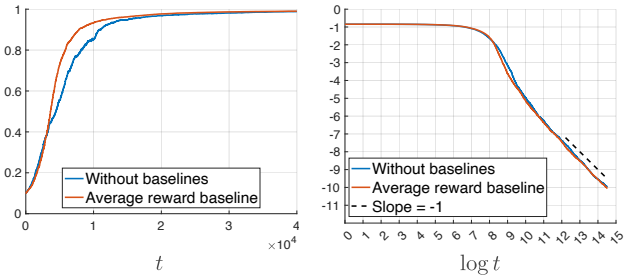


Figure 2. Both subfigures show results for $\pi_{\theta_t}(a^*)$. The left is for $\pi_{\theta_1}(a^*) = 0.03$, and the right is for $\pi_{\theta_1}(a^*) = 0.02$.

One unexpected observation in Figure 2 is that average reward baselines have worse performances, which is different with Sutton & Barto (2018). After checking numerical values, we found that since the initialization is bad, a sub-optimal action with $r(i) < r(a^*)$ will be pulled for most of the time, which results in $B_t \approx r(i)$. This implies that when a^* is pulled, $\theta_t(a^*)$ is increased less than without baselines, since the reward gap is also relatively small. Therefore, the average reward baseline might not be a baseline that is universally beneficial, which raises the question to design adaptive baseline, which is out of the scope of this paper, and we leave as our future work.

7.2. Convergence Rate

We further check the convergence rate empirically in the same problem settings. We use uniform initialization i.e., $\pi_{\theta_1}(a) = 1/K$ for all $a \in [K]$ and the results are shown in Figure 3. Each curve is an average from 10 independent runs, and the total iteration number is $T = 2 \times 10^6$. As shown in Figure 3(b), the slope in log scale is close to -1, which implies that $\log\left(\pi^* - \pi_{\theta_t}\right)^\top r \approx -\log t + C$. Equivalently, we have $\left(\pi^* - \pi_{\theta_t}\right)^\top r \approx C'/t$, verifying the $O(1/t)$ convergence rate in Theorem 5.5.



(a) $\pi_{\theta_t}(a^*)$ (b) $\log\left(\pi^* - \pi_{\theta_t}\right)^\top r$

Figure 3. Figure (a) shows the optimal action's probability and (b) shows log sub-optimal gap, which justifies our global convergence rate in Theorem 5.5.

7.3. Average Gradient Norm Convergence

In this section, we empirically verify the finite-step convergence rate in terms of average gradient norm. We follow exactly the same experimental settings in Section 7.2, but

evaluate the average gradient norm along the algorithm iterations. We illustrate the results in log-scale in Figure 4. It obviously aligned well with our convergence rate in terms of average gradient norm in Eq. (13).

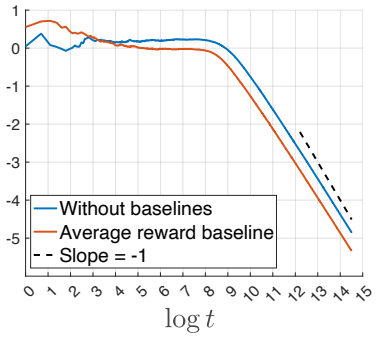


Figure 4. Average squared gradient norm $\frac{1}{t}\cdot\sum_{1\le s\le t}\mathbb{E}\left[\left\|\frac{d\pi_{\theta_s}^\top r}{d\theta_s}\right\|_2^2\right]$ in log scale (l.h.s. of Eq. (13)).

8. Boltzmann Exploration

The softmax parameterization used in gradient bandit algorithms is also called Boltzmann distribution (Sutton & Barto, 2018), based on which a classic algorithm EXP3 uses $O(1/\sqrt{t})$ learning rate and achieves a $O(1/\sqrt{t})$ rate (Auer et al., 2002b). The Botlzmann distribution has also been used in other policy gradient based algorithms. For example, Lan et al. (2022) show that mirror decent (MD) or NPG with strongly convex regularizers, increasing *batch sizes* and $O(1/t)$ learning rates achieves a $O(\log t/t)$ rate. The convergence in (Lan et al., 2022) heavily relies on batch observation for an accurate estimation of full gradient approximation, which is impossible in stochastic bandit setting, and thus not applicable.

There are also existing results revealing the weakness of Boltzmann distribution. Cesa-Bianchi et al. (2017) show that without count based bonuses, "Boltzmann exploration done wrong". In particular, there exists a 2-armed stochastic bandit problem with rewards bounded in $[0,1]$, when using $\Pr(a_t=a)\propto\exp\{\eta_t\cdot\hat{\mu}_{t,a}\}$, where

$$\hat{\mu}_{t,a}:=\frac{\sum_{s=1}^t\mathbb{I}\{a_s=a\}\cdot R_s(a)}{\sum_{s=1}^t\mathbb{I}\{a_s=a\}}$$

is the empirical mean estimator for $r(a)$, with $\eta_t>2\log t$ for all $t\ge 1$, would incur linear regret $\Omega(T)$. Instead of the aggressive update of parameters in softmax policy in (Cesa-Bianchi et al., 2017), the stochastic gradient bandit can be understood as a better way for parameter updates (with weights diminishing if the action is not selected) to ensure global convergence.

Cesa-Bianchi et al. (2017) also claim that the Boltzmann exploration is equivalent to the rule selecting

$$a_{t+1}=\arg\max_{a\in[K]}(\hat{\mu}_{t,a}+Z_{t,a}),$$

which is the widely used "Gumbel-Softmax" trick (Jang et al., 2016), where $Z_{t,a}$ is a Gumbel random variable independent for all $a\in[K]$. Cesa-Bianchi et al. (2017) show a $O(\log T)$ regret when replacing $Z_{t,a}$ by $\beta_{t,a}\cdot Z_{t,a}$, where $\beta_{t,a}$ is determined by count information $\sum_{s=1}^t\mathbb{I}\{a_t=a\}$. This inspires us that we may incorporate other techniques into gradient bandit algorithm and further improve it, especially for the poor initialization and problem dependent constant in Theorem 5.1 as also observed in Figure 2.

9. Conclusions

This work provides the first global convergence result for the gradient bandit algorithms (Sutton & Barto, 2018) using constant learning rates. The main technical finding is that the noise in stochastic gradient updates automatically vanishes such that noise control is unnecessary for global convergence. This work uncover a new understanding of stochastic gradient itself manages to achieve "weak exploraton" in the sense that the distribution over the arms almost surely concentrates asymptotically on a globally optimal action. One important future direction is to improve stochastic gradient to achieve "strong" exploration with finite-time optimal rates. Another direction of interest is to generalize the ideas and techniques to reinforcement learning.

Acknowledgements

The authors would like to thank anonymous reviewers for their valuable comments. Jincheng Mei would like to thank Ramki Gummadi for providing feedback on a draft of this manuscript. Csaba Szepesvári, Dale Schuurmans and Zixin Zhong gratefully acknowledge funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Mach. Learn. Res.*, 22(98):1–76, 2021.

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1. JMLR Workshop and Conference Proceedings, 2012.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International*

Conference on Machine Learning, pp. 242–252. PMLR, 2019.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Bhatnagar, S., Ghavamzadeh, M., Lee, M., and Sutton, R. S. Incremental natural actor-critic algorithms. *Advances in neural information processing systems*, 20, 2007.

Boucheron, S., Lugosi, G., and Massart, P. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14:1884–1899, 2009.

Breiman, L. *Probability*. SIAM, 1992.

Cesa-Bianchi, N., Gentile, C., Lugosi, G., and Neu, G. Boltzmann exploration done right. *Advances in neural information processing systems*, 30, 2017.

Chung, W., Thomas, V., Machado, M. C., and Roux, N. L. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. *arXiv preprint arXiv:2008.13773*, 2020.

Ding, Y., Zhang, J., and Lavaei, J. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

Doob, J. L. *Measure theory*, volume 143. Springer Science & Business Media, 2012.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Glynn, P. W. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33 (10):75–84, 1990.

Greensmith, E., Bartlett, P. L., and Baxter, J. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9), 2004.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Kakade, S. M. A natural policy gradient. In *Advances in neural information processing systems*, pp. 1531–1538, 2002.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Lan, G., Li, Y., and Zhao, T. Block policy mirror descent. *arXiv preprint arXiv:2201.05756*, 2022.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pp. 3107–3110. PMLR, 2021.

Mao, H., Venkatakrishnan, S. B., Schwarzkopf, M., and Alizadeh, M. Variance reduction for reinforcement learning in input-driven environments. *arXiv preprint arXiv:1807.02264*, 2018.

McDiarmid, C. and Reed, B. Concentration for self-bounding functions and an inequality of talagrand. *Random Structures & Algorithms*, 29(4):549–557, 2006.

Mei, J., Xiao, C., Dai, B., Li, L., Szepesvári, C., and Schuurmans, D. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33: 21130–21140, 2020a.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pp. 6820–6829. PMLR, 2020b.

Mei, J., Dai, B., Xiao, C., Szepesvari, C., and Schuurmans, D. Understanding the effect of stochasticity in policy optimization. *Advances in Neural Information Processing Systems*, 34:19339–19351, 2021a.

Mei, J., Gao, Y., Dai, B., Szepesvari, C., and Schuurmans, D. Leveraging non-uniformity in first-order non-convex optimization. In *International Conference on Machine Learning*, pp. 7555–7564. PMLR, 2021b.

Mei, J., Chung, W., Thomas, V., Dai, B., Szepesvari, C., and Schuurmans, D. The role of baselines in policy gradient optimization. *Advances in Neural Information Processing Systems*, 2022.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Robbins, H. and Monro, S. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. *arXiv preprint arXiv:1308.6370*, 2013.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

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

Tucker, G., Bhupatiraju, S., Gu, S., Turner, R., Ghahramani, Z., and Levine, S. The mirage of action-dependent baselines in reinforcement learning. In *International conference on machine learning*, pp. 5015–5024. PMLR, 2018.

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

Wu, C., Rajeswaran, A., Duan, Y., Kumar, V., Bayen, A. M., Kakade, S., Mordatch, I., and Abbeel, P. Variance reduction for policy gradient with action-dependent factorized baselines. *arXiv preprint arXiv:1803.07246*, 2018.

Yuan, R., Gower, R. M., and Lazaric, A. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 3332–3380. PMLR, 2022.

Zhang, J., Kim, J., O'Donoghue, B., and Boyd, S. Sample efficient reinforcement learning with reinforce. *arXiv preprint arXiv:2010.11364*, 2020a.

Zhang, J., Ni, C., Szepesvari, C., Wang, M., et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.

Zhang, K., Koppel, A., Zhu, H., and Basar, T. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020b.

A. Proofs for Algorithm 1

Proposition 2.3. Algorithm 1 is equivalent to the following stochastic gradient ascent update,

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \tag{27}$$

$$= \theta_t + \eta \cdot \left(\mathrm{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right) \hat{r}_t, \tag{28}$$

where $\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right] = \frac{d\pi_{\theta_t}^\top r}{d\theta_t}$, and $\left(\frac{d\pi_\theta}{d\theta}\right)^\top = \mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top$ is the Jacobian of $\theta \mapsto \pi_\theta := \mathrm{softmax}(\theta)$, and $\hat{r}_t(a) := \frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)} \cdot R_t(a)$ for all $a \in [K]$ is the importance sampling (IS) estimator, and we set $R_t(a) = 0$ for all $a \neq a_t$.

Proof. Using the definition of softmax Jacobian and $\hat{r}_t$, we have, for all $a \in [K]$,

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \left(\hat{r}_t(a) - \pi_{\theta_t}^\top \hat{r}_t\right) \tag{29}$$

$$= \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot (\hat{r}_t(a) - R_t(a_t)) \tag{30}$$

$$= \theta_t(a) + \begin{cases} \eta \cdot (1 - \pi_{\theta_t}(a)) \cdot R_t(a), & \text{if } a_t = a\,, \\ -\eta \cdot \pi_{\theta_t}(a) \cdot R_t(a_t), & \text{otherwise}\,. \end{cases}$$ □

Proposition 3.1 (Unbiased stochastic gradient with bounded variance / scale)**.** Using Algorithm 1, we have, for all $t \geq 1$,

$$\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right] = \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \tag{31}$$

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \leq 2\, R_{\max}^2, \tag{32}$$

where $\mathbb{E}_t[\cdot]$ is on randomness from the on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$ and reward sampling $R_t(a_t) \sim P_{a_t}$.

Proof. **First part, Eq. (31).** For all action $a \in [K]$, the true softmax PG is,

$$\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} = \pi_{\theta_t}(a) \cdot \left(r(a) - \pi_{\theta_t}^\top r\right). \tag{33}$$

For all $a \in [K]$, the stochastic softmax PG is,

$$\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)} = \pi_{\theta_t}(a) \cdot \left(\hat{r}_t(a) - \pi_{\theta_t}^\top \hat{r}_t\right) \tag{34}$$

$$= \pi_{\theta_t}(a) \cdot (\hat{r}_t(a) - R_t(a_t)) \tag{35}$$

$$= (\mathbb{I}\{a_t = a\} - \pi_{\theta_t}(a)) \cdot R_t(a_t). \tag{36}$$

For the sampled action a_t, we have,

$$\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a_t)}\right] = \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[(1 - \pi_{\theta_t}(a_t)) \cdot R_t(a_t)\right] \tag{37}$$

$$= (1 - \pi_{\theta_t}(a_t)) \cdot \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[R_t(a_t)\right] \tag{38}$$

$$= (1 - \pi_{\theta_t}(a_t)) \cdot r(a_t). \tag{39}$$

For any other not sampled action $a \neq a_t$, we have,

$$\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)}\right] = \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[-\pi_{\theta_t}(a) \cdot R_t(a_t)\right] \tag{40}$$

$$= -\pi_{\theta_t}(a) \cdot \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[R_t(a_t)\right] \tag{41}$$

$$= -\pi_{\theta_t}(a) \cdot r(a_t). \tag{42}$$

Combing Eqs. (37) and (40), we have, for all $a \in [K]$,

$$\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)}\right] = (\mathbb{I}\{a_t = a\} - \pi_{\theta_t}(a)) \cdot r(a_t). \tag{43}$$

Taking expectation over $a_t \sim \pi_{\theta_t}(\cdot)$, we have,

$$\begin{aligned}
\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)}\right] &= \Pr(a_t = a) \cdot \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)} \,\middle|\, a_t = a\right] + \Pr(a_t \neq a) \cdot \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)} \,\middle|\, a_t \neq a\right] && (44)\\
&= \pi_{\theta_t}(a) \cdot (1 - \pi_{\theta_t}(a)) \cdot r(a) + \sum_{a' \neq a} \pi_{\theta_t}(a') \cdot (-\pi_{\theta_t}(a)) \cdot r(a') && (45)\\
&= \pi_{\theta_t}(a) \cdot \sum_{a' \neq a} \pi_{\theta_t}(a') \cdot (r(a) - r(a')) && (46)\\
&= \pi_{\theta_t}(a) \cdot \left(r(a) - \pi_{\theta_t}^\top r\right). && (47)
\end{aligned}$$

Combining Eqs. (33) and (44), we have, for all $a \in [K]$,

$$\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)}\right] = \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}, \tag{48}$$

which implies Eq. (31) since $a \in [K]$ is arbitrary.

Second part, Eq. (32). The squared stochastic PG norm is,

$$\begin{aligned}
\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2 &= \sum_{a\in[K]} \left(\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t(a)}\right)^2 && (49)\\
&= \sum_{a\in[K]} (\mathbb{I}\{a_t = a\} - \pi_{\theta_t}(a))^2 \cdot R_t(a_t)^2 \qquad \text{(by Eq. (34))} && (50)\\
&\le R_{\max}^2 \cdot \sum_{a\in[K]} (\mathbb{I}\{a_t = a\} - \pi_{\theta_t}(a))^2 \qquad \text{(by Eq. (1))} && (51)\\
&= R_{\max}^2 \cdot \left[(1 - \pi_{\theta_t}(a_t))^2 + \sum_{a\neq a_t} \pi_{\theta_t}(a)^2\right] && (52)\\
&\le R_{\max}^2 \cdot \left[(1 - \pi_{\theta_t}(a_t))^2 + \Big(\sum_{a\neq a_t} \pi_{\theta_t}(a)\Big)^2\right] \qquad (\|x\|_2 \le \|x\|_1) && (53)\\
&= 2 \cdot R_{\max}^2 \cdot (1 - \pi_{\theta_t}(a_t))^2. && (54)
\end{aligned}$$

Therefore, we have, for all $a \in [K]$, conditioning on $a_t = a$,

$$\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2 \,\middle|\, a_t = a\right] \le 2 \cdot R_{\max}^2 \cdot (1 - \pi_{\theta_t}(a))^2. \tag{55}$$

Taking expectation over $a_t \sim \pi_{\theta_t}(\cdot)$, we have,

$$\begin{aligned}
\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] &= \sum_{a\in[K]} \Pr(a_t = a) \cdot \left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2 \,\middle|\, a_t = a\right] && (56)\\
&\le \sum_{a\in[K]} \pi_{\theta_t}(a) \cdot 2 \cdot R_{\max}^2 \cdot (1 - \pi_{\theta_t}(a))^2 && (57)\\
&\le 2 \cdot R_{\max}^2 \cdot \sum_{a\in[K]} \pi_{\theta_t}(a) \qquad (\pi_{\theta_t}(a) \in (0,1) \text{ for all } a \in [K]) && (58)\\
&= 2\, R_{\max}^2. && \square
\end{aligned}$$

Lemma 4.1 (Non-uniform smoothness (NS), Mei et al. (2021b, Lemma 2))**.** For all $\theta \in \mathbb{R}^K$, the spectral radius of Hessian matrix $\frac{d^2\{\pi_\theta^\top r\}}{d\theta^2} \in \mathbb{R}^{K\times K}$ is upper bounded by $3 \cdot \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2$, i.e., for all $y \in \mathbb{R}^K$,

$$\left| y^\top \frac{d^2\{\pi_\theta^\top r\}}{d\theta^2} y \right| \le 3 \cdot \left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2 \cdot \|y\|_2^2. \tag{59}$$

Proof. See the proof in Mei et al. (2021b, Lemma 2). We include a proof for completeness.

Let $S \coloneqq S(r,\theta) \in \mathbb{R}^{K\times K}$ be the second derivative of the map $\theta \mapsto \pi_\theta^\top r$. Denote $H(\pi_\theta) \coloneqq \mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top$ as the softmax Jacobian. By definition we have,

$$\begin{aligned} S &= \frac{d}{d\theta}\left\{\frac{d\pi_\theta^\top r}{d\theta}\right\} && (60)\\ &= \frac{d}{d\theta}\left\{H(\pi_\theta) r\right\} && (61)\\ &= \frac{d}{d\theta}\left\{(\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top) r\right\}. && (62)\end{aligned}$$

Continuing with our calculation fix $i, j \in [K]$. Then,

$$\begin{aligned} S_{(i,j)} &= \frac{d\{\pi_\theta(i)\cdot(r(i) - \pi_\theta^\top r)\}}{d\theta(j)} && (63)\\ &= \frac{d\pi_\theta(i)}{d\theta(j)}\cdot(r(i) - \pi_\theta^\top r) + \pi_\theta(i)\cdot\frac{d\{r(i) - \pi_\theta^\top r\}}{d\theta(j)} && (64)\\ &= (\delta_{ij}\cdot\pi_\theta(j) - \pi_\theta(i)\cdot\pi_\theta(j))\cdot(r(i) - \pi_\theta^\top r) - \pi_\theta(i)\cdot(\pi_\theta(j)\cdot r(j) - \pi_\theta(j)\cdot\pi_\theta^\top r) && (65)\\ &= \delta_{ij}\cdot\pi_\theta(j)\cdot(r(i) - \pi_\theta^\top r) - \pi_\theta(i)\cdot\pi_\theta(j)\cdot(r(i) - \pi_\theta^\top r) - \pi_\theta(i)\cdot\pi_\theta(j)\cdot(r(j) - \pi_\theta^\top r), && (66)\end{aligned}$$

where δ_{ij} is the Kronecker's δ-function defined as,

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j,\\ 0, & \text{otherwise.}\end{cases} \tag{67}$$

To show the bound on the spectral radius of S, pick $y \in \mathbb{R}^K$. Then,

$$\begin{aligned} |y^\top S y| &= \left|\sum_{i=1}^K\sum_{j=1}^K S_{(i,j)}\cdot y(i)\cdot y(j)\right| && (68)\\ &= \left|\sum_i \pi_\theta(i)\cdot(r(i) - \pi_\theta^\top r)\cdot y(i)^2 - 2\cdot\sum_i \pi_\theta(i)\cdot(r(i) - \pi_\theta^\top r)\cdot y(i)\cdot\sum_j \pi_\theta(j)\cdot y(j)\right| && (69)\\ &= \left|(H(\pi_\theta)r)^\top (y\odot y) - 2\cdot(H(\pi_\theta)r)^\top y\cdot\left(\pi_\theta^\top y\right)\right| && (70)\\ &\le \left|(H(\pi_\theta)r)^\top (y\odot y)\right| + 2\cdot\left|(H(\pi_\theta)r)^\top y\right|\cdot\left|\pi_\theta^\top y\right| \qquad \text{(triangle inequality)} && (71)\\ &\le \|H(\pi_\theta)r\|_\infty\cdot\|y\odot y\|_1 + 2\cdot\|H(\pi_\theta)r\|_2\cdot\|y\|_2\cdot\|\pi_\theta\|_1\cdot\|y\|_\infty \qquad \text{(Hölder's inequality)} && (72)\\ &\le 3\cdot\|H(\pi_\theta)r\|_2\cdot\|y\|_2^2, && (73)\end{aligned}$$

where $\odot$ is Hadamard (component-wise) product, and the last inequality uses $\|H(\pi_\theta)r\|_\infty \le \|H(\pi_\theta)r\|_2$, $\|y\odot y\|_1 = \|y\|_2^2$, $\|\pi_\theta\|_1 = 1$, and $\|y\|_\infty \le \|y\|_2$. Therefore, we have,

$$\begin{aligned} |y^\top S y| &\le 3\cdot\|H(\pi_\theta)r\|_2\cdot\|y\|_2^2 \qquad \text{(by Eq. (68))} && (74)\\ &= 3\cdot\left\|\left(\mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top\right) r\right\|_2\cdot\|y\|_2^2 \qquad \left(H(\pi_\theta) \coloneqq \mathrm{diag}(\pi_\theta) - \pi_\theta\pi_\theta^\top\right) && (75)\\ &= 3\cdot\left\|\frac{d\pi_\theta^\top r}{d\theta}\right\|_2\cdot\|y\|_2^2. \qquad \text{(by Eq. (3))} \end{aligned}$$

□

Lemma 4.2 (NS between iterates)**.** Using Algorithm 1 with $\eta \in \left(0, \frac{2}{9 \cdot R_{\max}}\right)$, we have, for all $t \geq 1$,

$$D(\theta_{t+1}, \theta_t) \coloneqq \left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \leq \frac{\beta(\theta_t)}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{76}$$

where

$$\beta(\theta_t) = \frac{6}{2 - 9 \cdot R_{\max} \cdot \eta} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2. \tag{77}$$

Proof. Denote $\theta_\zeta \coloneqq \theta_t + \zeta \cdot (\theta_{t+1} - \theta_t)$ with some $\zeta \in [0, 1]$. According to Taylor's theorem, we have,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| = \frac{1}{2} \cdot \left| (\theta_{t+1} - \theta_t)^\top \frac{d^2 \pi_{\theta_\zeta}^\top r}{d\theta_\zeta^2} (\theta_{t+1} - \theta_t) \right| \tag{78}$$

$$\leq \frac{3}{2} \cdot \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2. \qquad \text{(by Lemma 4.1)} \tag{79}$$

Denote $\theta_{\zeta_1} \coloneqq \theta_t + \zeta_1 \cdot (\theta_\zeta - \theta_t)$ with some $\zeta_1 \in [0, 1]$. We have,

$$\left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} - \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 = \left\| \int_0^1 \left\langle \frac{d^2 \{\pi_{\theta_{\zeta_1}}^\top r\}}{d\theta_{\zeta_1}^2}, \theta_\zeta - \theta_t \right\rangle d\zeta_1 \right\|_2 \qquad \text{(Fundamental theorem of calculus)} \tag{80}$$

$$\leq \int_0^1 \left\| \frac{d^2 \{\pi_{\theta_{\zeta_1}}^\top r\}}{d\theta_{\zeta_1}^2} \right\|_2 \cdot \|\theta_\zeta - \theta_t\|_2 \, d\zeta_1 \qquad \text{(by Cauchy–Schwarz)} \tag{81}$$

$$\leq \int_0^1 3 \cdot \left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 \cdot \|\theta_\zeta - \theta_t\|_2 \, d\zeta_1 \qquad \text{(by Lemma 4.1)} \tag{82}$$

$$= \int_0^1 3 \cdot \left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 \cdot \zeta \cdot \|\theta_{t+1} - \theta_t\|_2 \, d\zeta_1 \qquad (\theta_\zeta \coloneqq \theta_t + \zeta \cdot (\theta_{t+1} - \theta_t)) \tag{83}$$

$$\leq \int_0^1 3 \cdot \left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 \cdot \eta \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2 d\zeta_1, \qquad \left(\zeta \in [0, 1], \text{ and } \theta_{t+1} = \theta_t + \eta \cdot \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right) \tag{84}$$

where the second inequality is because of the Hessian is symmetric, and its operator norm is equal to its spectral radius. Therefore, we have,

$$\left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2 \leq \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 + \left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} - \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \qquad \text{(by triangle inequality)} \tag{85}$$

$$\leq \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 + 3\,\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2 \cdot \int_0^1 \left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 d\zeta_1. \qquad \text{(by Eq. (80))} \tag{86}$$

Denote $\theta_{\zeta_2} \coloneqq \theta_t + \zeta_2 \cdot (\theta_{\zeta_1} - \theta_t)$ with $\zeta_2 \in [0, 1]$. Using similar calculation in Eq. (80), we have,

$$\left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} \right\|_2 \leq \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 + \left\| \frac{d\pi_{\theta_{\zeta_1}}^\top r}{d\theta_{\zeta_1}} - \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \tag{87}$$

$$\leq \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 + 3\,\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2 \cdot \int_0^1 \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} \right\|_2 d\zeta_2. \tag{88}$$

Combining Eqs. (85) and (87), we have,

$$\left\| \frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta} \right\|_2 \leq \left(1 + 3\,\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2 \right) \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 + \left(3\,\eta \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2 \right)^2 \cdot \int_0^1 \int_0^1 \left\| \frac{d\pi_{\theta_{\zeta_2}}^\top r}{d\theta_{\zeta_2}} \right\|_2 d\zeta_2 d\zeta_1, \tag{89}$$

which, by recurring the above arguments, implies that,

$$\left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2 \leq \sum_{i=0}^{\infty}\left(3\,\eta\cdot\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2\right)^i\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2. \tag{90}$$

Next, we have,

$$\begin{aligned}
3\,\eta\cdot\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2 &\leq 3\,\eta\cdot\sqrt{2\cdot R_{\max}^2\cdot\left(1-\pi_{\theta_t}(a_t)\right)^2} && \text{(by Eq. (49))} && (91)\\
&< \frac{3\cdot 2}{9\cdot R_{\max}}\cdot\sqrt{2}\cdot R_{\max} && \left(\pi_{\theta_t}(a_t)\in(0,1),\ \text{and}\ \eta<\frac{2}{9\cdot R_{\max}}\right) && (92)\\
&< 1. && && (93)
\end{aligned}$$

Combining Eqs. (90) and (91), we have,

$$\begin{aligned}
\left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2 &\leq \frac{1}{1-3\,\eta\cdot\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 && \left(3\,\eta\cdot\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2\in(0,1)\ \text{from Eq. (91)}\right) && (94)\\
&\leq \frac{1}{1-3\,\eta\cdot\sqrt{2}\cdot R_{\max}}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 && \text{(by Eq. (49))} && (95)\\
&< \frac{1}{1-\frac{9}{2}\cdot R_{\max}\cdot\eta}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2. && && (96)
\end{aligned}$$

Combining Eqs. (78) and (94), we have,

$$\begin{aligned}
\left|(\pi_{\theta_{t+1}}-\pi_{\theta_t})^\top r-\left\langle\frac{d\pi_{\theta_t}^\top r}{d\theta_t},\theta_{t+1}-\theta_t\right\rangle\right| &\leq \frac{3}{2}\cdot\left\|\frac{d\pi_{\theta_\zeta}^\top r}{d\theta_\zeta}\right\|_2\cdot\|\theta_{t+1}-\theta_t\|_2^2 && (97)\\
&\leq \frac{3}{2-9\cdot R_{\max}\cdot\eta}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\cdot\|\theta_{t+1}-\theta_t\|_2^2. && \square
\end{aligned}$$

Lemma 4.3 (Strong growth conditions / Self-bounding noise property)**.** Using Algorithm 1, we have, for all $t\geq 1$,

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right]\leq\frac{8\cdot R_{\max}^3\cdot K^{3/2}}{\Delta^2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2, \tag{98}$$

where $\Delta\coloneqq\min_{i\neq j}|r(i)-r(j)|$.

Proof. Given $t\geq 1$, denote k_t as the action with largest probability, i.e., $k_t\coloneqq\arg\max_{a\in[K]}\pi_{\theta_t}(a)$. We have,

$$\pi_{\theta_t}(k_t)\geq\frac{1}{K}. \tag{99}$$

According to Eq. (56), we have,

$$\begin{aligned}
\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] &= \sum_{a\in[K]}\Pr\left(a_t=a\right)\cdot\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\,\middle|\,a_t=a\right] && (100)\\
&\leq 2\cdot R_{\max}^2\cdot\sum_{a\in[K]}\pi_{\theta_t}(a)\cdot\left(1-\pi_{\theta_t}(a)\right)^2 && (101)\\
&= 2\cdot R_{\max}^2\cdot\left[\pi_{\theta_t}(k_t)\cdot\left(1-\pi_{\theta_t}(k_t)\right)^2+\sum_{a\neq k_t}\pi_{\theta_t}(a)\cdot\left(1-\pi_{\theta_t}(a)\right)^2\right] && (102)\\
&\leq 2\cdot R_{\max}^2\cdot\left[1-\pi_{\theta_t}(k_t)+\sum_{a\neq k_t}\pi_{\theta_t}(a)\right] && (\pi_{\theta_t}(a)\in(0,1)\ \text{for all}\ a\in[K]) && (103)\\
&= 4\cdot R_{\max}^2\cdot\left(1-\pi_{\theta_t}(k_t)\right). && && (104)
\end{aligned}$$

On the other hand, we have,

$$\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 = \sum_{a\in[K]} \pi_{\theta_t}(a)^2 \cdot (r(a) - \pi_{\theta_t}^\top r)^2 \tag{105}$$

$$= \sum_{a'\in[K]} (r(a') - \pi_{\theta_t}^\top r)^2 \cdot \sum_{a\in[K]} \pi_{\theta_t}(a)^2 \cdot \frac{(r(a) - \pi_{\theta_t}^\top r)^2}{\sum_{a'\in[K]} (r(a') - \pi_{\theta_t}^\top r)^2} \tag{106}$$

$$\geq \sum_{a'\in[K]} (r(a') - \pi_{\theta_t}^\top r)^2 \cdot \left[\sum_{a\in[K]} \pi_{\theta_t}(a) \cdot \frac{(r(a) - \pi_{\theta_t}^\top r)^2}{\sum_{a'\in[K]} (r(a') - \pi_{\theta_t}^\top r)^2}\right]^2 \qquad \text{(Jensen's inequality)} \tag{107}$$

$$= \frac{1}{\sum_{a'\in[K]} (r(a') - \pi_{\theta_t}^\top r)^2} \cdot \left[\sum_{a\in[K]} \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)^2\right]^2 \tag{108}$$

$$\geq \frac{1}{4 \cdot K \cdot R_{\max}^2} \cdot \left[\sum_{a\in[K]} \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)^2\right]^2, \qquad \left(r \in [-R_{\max}, R_{\max}]^K\right) \tag{109}$$

which implies that,

$$\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \geq \frac{1}{2 \cdot \sqrt{K} \cdot R_{\max}} \cdot \sum_{a\in[K]} \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)^2. \tag{110}$$

Using similar calculations in the proofs for Mei et al. (2021a, Lemma 2), we have,

$$\sum_{a\in[K]} \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)^2 = \sum_{i=1}^K \pi_{\theta_t}(i) \cdot r(i)^2 - \left[\sum_{i=1}^K \pi_{\theta_t}(i) \cdot r(i)\right]^2 \tag{111}$$

$$= \sum_{i=1}^K \pi_{\theta_t}(i) \cdot r(i)^2 - \sum_{i=1}^K \pi_{\theta_t}(i)^2 \cdot r(i)^2 - 2 \cdot \sum_{i=1}^{K-1} \pi_{\theta_t}(i) \cdot r(i) \cdot \sum_{j=i+1}^K \pi_{\theta_t}(j) \cdot r(j) \tag{112}$$

$$= \sum_{i=1}^K \pi_{\theta_t}(i) \cdot r(i)^2 \cdot (1 - \pi_{\theta_t}(i)) - 2 \cdot \sum_{i=1}^{K-1} \pi_{\theta_t}(i) \cdot r(i) \cdot \sum_{j=i+1}^K \pi_{\theta_t}(j) \cdot r(j) \tag{113}$$

$$= \sum_{i=1}^K \pi_{\theta_t}(i) \cdot r(i)^2 \cdot \sum_{j\neq i} \pi_{\theta_t}(j) - 2 \cdot \sum_{i=1}^{K-1} \pi_{\theta_t}(i) \cdot r(i) \cdot \sum_{j=i+1}^K \pi_{\theta_t}(j) \cdot r(j) \tag{114}$$

$$= \sum_{i=1}^{K-1} \pi_{\theta_t}(i) \cdot \sum_{j=i+1}^K \pi_{\theta_t}(j) \cdot \left(r(i)^2 + r(j)^2\right) - 2 \cdot \sum_{i=1}^{K-1} \pi_{\theta_t}(i) \cdot r(i) \cdot \sum_{j=i+1}^K \pi_{\theta_t}(j) \cdot r(j) \tag{115}$$

$$= \sum_{i=1}^{K-1} \pi_{\theta_t}(i) \cdot \sum_{j=i+1}^K \pi_{\theta_t}(j) \cdot (r(i) - r(j))^2, \tag{116}$$

which implies that,

$$\sum_{a\in[K]} \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)^2 \geq \sum_{i=1}^{k_t} \pi_{\theta_t}(i) \cdot \sum_{j=i+1}^K \pi_{\theta_t}(j) \cdot (r(i) - r(j))^2 \qquad \text{(fewer terms)} \tag{117}$$

$$\geq \sum_{i=1}^{k_t-1} \pi_{\theta_t}(i) \cdot \pi_{\theta_t}(k_t) \cdot (r(i) - r(k_t))^2 + \pi_{\theta_t}(k_t) \cdot \sum_{j=k_t+1}^K \pi_{\theta_t}(j) \cdot (r(k_t) - r(j))^2 \qquad \text{(fewer terms)} \tag{118}$$

$$= \pi_{\theta_t}(k_t) \cdot \sum_{a\neq k_t} \pi_{\theta_t}(a) \cdot (r(a) - r(k_t))^2 \tag{119}$$

$$\geq \frac{\Delta^2}{K} \cdot (1 - \pi_{\theta_t}(k_t)), \qquad \text{(by Eq. (99))} \tag{120}$$

where $\Delta \coloneqq \min_{i \neq j} |r(i) - r(j)|$. Therefore, we have,

$$\mathbb{E}_t \left[\left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2^2 \right] \leq 4 \cdot R_{\max}^2 \cdot (1 - \pi_{\theta_t}(k_t)) \qquad \text{(by Eq. (100))} \tag{121}$$

$$\leq \frac{4 \cdot R_{\max}^2 \cdot K}{\Delta^2} \cdot \sum_{a \in [K]} \pi_{\theta_t}(a) \cdot (r(a) - \pi_{\theta_t}^\top r)^2 \qquad \text{(by Eq. (117))} \tag{122}$$

$$\leq \frac{4 \cdot R_{\max}^2 \cdot K}{\Delta^2} \cdot 2 \cdot \sqrt{K} \cdot R_{\max} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \qquad \text{(by Eq. (110))} \tag{123}$$

$$= \frac{8 \cdot R_{\max}^3 \cdot K^{3/2}}{\Delta^2} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2. \qquad \square$$

Lemma 4.6 (Constant learning rates)**.** Using Algorithm 1 with $\eta = \frac{\Delta^2}{40 \cdot K^{3/2} \cdot R_{\max}^3}$, we have, for all $t \geq 1$,

$$\pi_{\theta_t}^\top r - \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] \leq -\frac{\Delta^2}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2. \tag{124}$$

Proof. Using the learning rate,

$$\eta = \frac{\Delta^2}{40 \cdot K^{3/2} \cdot R_{\max}^3} \tag{125}$$

$$= \frac{4}{45 \cdot R_{\max}} \cdot \frac{\Delta^2}{R_{\max}^2} \cdot \frac{1}{K^{3/2}} \cdot \frac{45}{4} \cdot \frac{1}{40} \tag{126}$$

$$\leq \frac{4}{45 \cdot R_{\max}} \cdot 4 \cdot \frac{1}{2 \cdot \sqrt{2}} \cdot \frac{45}{4} \cdot \frac{1}{40}, \qquad (\Delta \leq 2 \cdot R_{\max}, \text{ and } K \geq 2) \tag{127}$$

$$< \frac{4}{45 \cdot R_{\max}}, \tag{128}$$

we have $\eta \in \left(0, \frac{2}{9 \cdot R_{\max}}\right)$. According to Lemma 4.2, we have,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle \right| \leq \frac{3}{2 - 9 \cdot R_{\max} \cdot \eta} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{129}$$

$$\leq \frac{3}{2 - 9 \cdot R_{\max} \cdot \frac{4}{45 \cdot R_{\max}}} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \qquad \text{(by Eq. (128))} \tag{130}$$

$$= \frac{5}{2} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{131}$$

which implies that,

$$\pi_{\theta_t}^\top r - \pi_{\theta_{t+1}}^\top r \leq -\left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle + \frac{5}{2} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{132}$$

$$= -\eta \cdot \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\rangle + \frac{5}{2} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \cdot \eta^2 \cdot \left\| \frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t} \right\|_2^2, \tag{133}$$

where the last equation uses Algorithm 1. Taking expectation over $a_t \sim \pi_{\theta_t}(\cdot)$ and $R_t(a_t) \sim P_{a_t}$, we have,

$$\pi_{\theta_t}^\top r - \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] \le -\eta \cdot \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right]\right\rangle + \frac{5}{2}\cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \eta^2 \cdot \mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \tag{134}$$

$$= -\eta \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 + \frac{5}{2}\cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \eta^2 \cdot \mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \quad \text{(by Proposition 3.1)} \tag{135}$$

$$\le -\eta \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 + \frac{5}{2}\cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \eta^2 \cdot \frac{8\cdot R_{\max}^3 \cdot K^{3/2}}{\Delta^2} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \quad \text{(by Lemma 4.3)} \tag{136}$$

$$= \left(-\eta + \eta^2 \cdot \frac{20\cdot R_{\max}^3 \cdot K^{3/2}}{\Delta^2}\right) \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 \tag{137}$$

$$= -\frac{\Delta^2}{80\cdot K^{3/2}\cdot R_{\max}^3} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2. \quad \text{(by Eq. (125))} \tag{138}$$

□

Corollary 4.7. Using Algorithm 1, we have, the sequence $\{\pi_{\theta_t}^\top r\}_{t\ge 1}$ converges w. p. 1.

Proof. Setting $Y_t = r(a^*) - \pi_{\theta_t}^\top r$, we have $Y_t \in [-R_{\max}, R_{\max}]$ by Eq. (1). Define $\mathcal{F}_t$ as the σ-algebra generated by $\{a_1, R_1(a_1), a_2, R_2(a_2), \ldots, a_{t-1}, R_{t-1}(a_{t-1})\}$. Note that Y_t is $\mathcal{F}_t$-measurable since θ_t is a deterministic function of $a_1, R_1(a_1), \ldots, a_{t-1}, R_{t-1}(a_{t-1})$. According to Lemma 4.6, using Algorithm 1, we have, for all $t \ge 1$, $\pi_{\theta_t}^\top r - \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] \le 0$, which indicates that $\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \le Y_t$. Hence, the conditions of Doob's super-martingale theorem (Theorem C.1) are satisfied and the result follows. □

A.1. Proof of Theorem 5.1

Theorem 5.1 (Asymptotic global convergence)**.** Using Algorithm 1, we have, almost surely,

$$\pi_{\theta_t}(a^*) \to 1, \text{ as } t \to \infty, \tag{139}$$

which implies that $\inf_{t\ge 1} \pi_{\theta_t}(a^*) > 0$.

Proof. According to Algorithm 1, for each $a \in [K]$, the update is,

$$\theta_{t+1}(a) = \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \left(\frac{\mathbb{I}\{a_t = a\}}{\pi_{\theta_t}(a)} \cdot R_t(a) - R_t(a_t)\right). \tag{140}$$

Given $i \in [K]$, define the following set $\mathcal{P}(i)$ of "generalized one-hot policy",

$$\mathcal{A}(i) \coloneqq \{j \in [K] : r(j) = r(i)\}, \tag{141}$$

$$\mathcal{P}(i) \coloneqq \left\{\pi \in \Delta(K) : \sum_{j\in\mathcal{A}(i)} \pi(j) = 1\right\}. \tag{142}$$

We make the following two claims.

Claim 1. Almost surely, π_{θ_t} approaches one "generalized one-hot policy", i.e., there exists (a possibly random) $i \in [K]$, such that $\sum_{j\in\mathcal{A}(i)} \pi_{\theta_t}(j) \to 1$ almost surely as $t \to \infty$.

Claim 2. Almost surely, π_{θ_t} cannot approach any "sub-optimal generalized one-hot policies", i.e., i in the previous claim must be an optimal action.

From Claim 2, it follows that $\sum_{j\in\mathcal{A}(a^*)} \pi_{\theta_t}(j) \to 1$ almost surely, as $t \to \infty$ and thus the policy sequence obtained almost surely convergences to a globally optimal policy π^*.

Proof of Claim 1. According to Corollary 4.7, we have that for some (possibly random) $c \in [-R_{\max}, R_{\max}]$, almost surely,

$$\lim_{t\to\infty} \pi_{\theta_t}^\top r = c. \tag{143}$$

Thanks to $\pi_{\theta_t}^\top r \in [-R_{\max}, R_{\max}]$ and $\pi_{\theta_t}^\top r - \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] \le 0$ by Lemma 4.6, we have that $X_t = \pi_{\theta_t}^\top r$ ($t \ge 1$) satisfies the conditions of Corollary 3 in (Mei et al., 2022). Hence, by this result, almost surely,

$$\lim_{t\to\infty} \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] - \pi_{\theta_{t+1}}^\top r = 0\,, \tag{143}$$

which, combined with Eq. (142) also gives that $\lim_{t\to\infty} \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] = c$ almost surely. Hence,

$$\lim_{t\to\infty} \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] - \pi_{\theta_t}^\top r = c - c = 0, \qquad \text{a.s.} \tag{144}$$

According to Lemma 4.6, we have,

$$\mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] - \pi_{\theta_t}^\top r \ge \frac{\Delta^2}{80\cdot K^{3/2}\cdot R_{\max}^3} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \tag{145}$$

$$= \frac{\Delta^2}{80\cdot K^{3/2}\cdot R_{\max}^3} \cdot \sum_{i=1}^K \pi_{\theta_t}(i)^2 \cdot \left(r(i) - \pi_{\theta_t}^\top r\right)^2. \qquad \text{(by Eq. (14))} \tag{146}$$

Combining Eqs. (144) and (145), we have, with probability 1,

$$\lim_{t\to\infty} \sum_{i=1}^K \pi_{\theta_t}(i)^2 \cdot \left(r(i) - \pi_{\theta_t}^\top r\right)^2 = 0, \tag{147}$$

which implies that, for all $i \in [K]$, almost surely,

$$\lim_{t\to\infty} \pi_{\theta_t}(i)^2 \cdot \left(r(i) - \pi_{\theta_t}^\top r\right)^2 = 0. \tag{148}$$

We claim that c, the almost sure limit of $\pi_{\theta_t}^\top r$, is such that almost surely, for some (possibly random) $i \in [K]$, $c = r(i)$ almost surely. We prove this by contradiction. Let $\mathcal{E}_i = \{c = r(i)\}$. Hence, our goal is to show that $\mathbb{P}(\cup_i \mathcal{E}_i) = 1$. Clearly, this follows from $\mathbb{P}(\cap_i \mathcal{E}_i^c) = 0$, hence, we prove this. On $\mathcal{E}_i^c$, since $\lim_{t\to\infty} \pi_{\theta_t}^\top r \ne r(i)$, we also have

$$\lim_{t\to\infty} \left(r(i) - \pi_{\theta_t}^\top r\right)^2 > 0, \qquad \text{almost surely on } \mathcal{E}_i^c\,. \tag{149}$$

This, together with Eq. (148) gives that almost surely on $\mathcal{E}_i^c$,

$$\lim_{t\to\infty} \pi_{\theta_t}(i)^2 = 0. \tag{150}$$

Hence, on $\cap_i \mathcal{E}_i^c$, almost surely, for all $i \in [K]$, $\lim_{t\to\infty} \pi_{\theta_t}(i)^2 = 0$. This contradicts with that $\sum_i \pi_{\theta_t}(i) = 1$ holds for all $t \ge 1$, and hence we must have that $\mathbb{P}(\cap_i \mathcal{E}_i^c) = 0$, finishing the proof that $\mathbb{P}(\cup_i \mathcal{E}_i) = 1$.

Now, let $i \in [K]$ be the (possibly random) index of the action for which $c = r(i)$ almost surely. Recall that $\mathcal{A}(i)$ contains all actions j with $r(j) = r(i)$ (cf. Eq. (140)). Clearly, it holds that for all $j \in \mathcal{A}(i)$,

$$\lim_{t\to\infty} \pi_{\theta_t}^\top r = r(j), \qquad \text{a.s.}, \tag{151}$$

and we have, for all $k \notin \mathcal{A}(i)$,

$$\lim_{t\to\infty} \left(r(k) - \pi_{\theta_t}^\top r\right)^2 > 0, \qquad \text{a.s.}, \tag{152}$$

which implies that,

$$\lim_{t\to\infty} \sum_{k\notin\mathcal{A}(i)} \pi_{\theta_t}(k)^2 = 0, \qquad \text{a.s.} \tag{153}$$

Therefore, we have,

$$\lim_{t\to\infty} \sum_{j\in\mathcal{A}(i)} \pi_{\theta_t}(j) = 1, \qquad \text{a.s.}, \tag{154}$$

which means π_{θ_t} a.s. approaches the "generalized one-hot policy" $\mathcal{P}(i)$ in Eq. (141) as $t \to \infty$, finishing the proof of the first claim.

Proof of Claim 2. Recall that this claim stated that $\lim_{t\to\infty} \sum_{j\in\mathcal{A}(a^*)} \pi_{\theta_t}(j) = 1$. The brief sketch of the proof is as follows: By Claim 1, there exists a (possibly random) $i \in [K]$ such that $\sum_{j\in\mathcal{A}(i)} \pi_{\theta_t}(j) \to 1$ almost surely, as $t \to \infty$. If $i = a^*$ almost surely, Claim 2 follows. Hence, it suffices to consider the event that $\{i \neq a^*\}$ and show that this event has zero probability mass. Hence, in the rest of the proof we assume that we are on the event when $i \neq a^*$.

Since $i \neq a^*$, there exists at least one "good" action $a^+ \in [K]$ such that $r(a^+) > r(i)$. The two cases are as follows.

2a) All "good" actions are sampled finitely many times as $t \to \infty$.

2b) At least one "good" action is sampled infinitely many times as $t \to \infty$.

In both cases, we show that $\sum_{j\in\mathcal{A}(i)} \exp\{\theta_t(j)\} < \infty$ as $t \to \infty$ (but for different reasons), which is a contradiction with the assumption of $\sum_{j\in\mathcal{A}(i)} \pi_{\theta_t}(j) \to 1$ as $t \to \infty$, given that a "good" action's parameter is almost surely lower bounded. Hence, $i \neq a^*$ almost surely does not happen, which means that almost surely $i = a^*$. Let us now turn to the details of the proof. We start with some useful extra notation. For each action $a \in [K]$, for $t \geq 2$, we have the following decomposition,

$$\theta_t(a) = \underbrace{\theta_t(a) - \mathbb{E}_{t-1}[\theta_t(a)]}_{W_t(a)} + \underbrace{\mathbb{E}_{t-1}[\theta_t(a)] - \theta_{t-1}(a)}_{P_{t-1}(a)} + \theta_{t-1}(a), \tag{155}$$

while we also have,

$$\theta_1(a) = \underbrace{\theta_1(a) - \mathbb{E}[\theta_1(a)]}_{W_1(a)} + \mathbb{E}[\theta_1(a)], \tag{156}$$

where $\mathbb{E}[\theta_1(a)]$ accounts for possible randomness in initialization of θ_1.

Define the following notations,

$$Z_t(a) \coloneqq W_1(a) + \cdots + W_t(a), \qquad \text{("cumulative noise")} \tag{157}$$
$$W_t(a) \coloneqq \theta_t(a) - \mathbb{E}_{t-1}[\theta_t(a)], \qquad \text{("noise")} \tag{158}$$
$$P_t(a) \coloneqq \mathbb{E}_t[\theta_{t+1}(a)] - \theta_t(a). \qquad \text{("progress")} \tag{159}$$

Recursing Eq. (155) gives,

$$\theta_t(a) = \mathbb{E}[\theta_1(a)] + Z_t(a) + \underbrace{P_1(a) + \cdots + P_{t-1}(a)}_{\text{"cumulative progress"}}. \tag{160}$$

We have that $\mathbb{E}_t[W_{t+1}(a)] = 0$, for $t = 0, 1, \ldots$. Let

$$I_t(a) = \begin{cases} 1, & \text{if } a_t = a\,; \\ 0, & \text{otherwise}\,. \end{cases} \tag{161}$$

The update rule (cf. Algorithm 1) is,

$$\theta_{t+1}(a) = \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \left(\frac{\mathbb{I}\{a_t = a\}}{\pi_{\theta_t}(a)} \cdot R_t(a) - R_t(a_t) \right), \tag{162}$$

where $a_t \sim \pi_{\theta_t}(\cdot)$, and $x_t(a) \sim P_a$. Let $\mathcal{F}_t$ be the σ-algebra generated by $a_1, x_1(a_1), \cdots, a_{t-1}, x_{t-1}(a_{t-1})$:

$$\mathcal{F}_t = \sigma(\{a_1, R_1(a_1), \cdots, a_{t-1}, R_{t-1}(a_{t-1})\})\,. \tag{163}$$

Note that θ_t, I_t are $\mathcal{F}_t$-measurable and $\hat{x}_t$ is $\mathcal{F}_{t+1}$-measurable for all $t \geq 1$. Let $\mathbb{E}_t$ denote the conditional expectation with respect to $\mathcal{F}_t$: $\mathbb{E}_t[X] = \mathbb{E}[X|\mathcal{F}_t]$.

Using the above notations, we have,

$$W_{t+1}(a) = \theta_{t+1}(a) - \mathbb{E}_t[\theta_{t+1}(a)] \tag{164}$$
$$= \theta_t(a) + \eta \cdot [I_t(a) - \pi_{\theta_t}(a)] \cdot R_t(a_t) - \left[\theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \left(r(a) - \pi_{\theta_t}^\top r\right)\right] \tag{165}$$
$$= \eta \cdot [I_t(a) - \pi_{\theta_t}(a)] \cdot R_t(a_t) - \eta \cdot \pi_{\theta_t}(a) \cdot \left[r(a) - \pi_{\theta_t}^\top r\right], \tag{166}$$

which implies that,

$$Z_t(a) = W_1(a) + \cdots + W_t(a) \tag{167}$$
$$= \sum_{s=1}^{t} \eta \cdot [I_s(a) - \pi_{\theta_s}(a)] \cdot R_s(a_s) - \eta \cdot \pi_{\theta_s}(a) \cdot \left[r(a) - \pi_{\theta_s}^\top r\right]. \tag{168}$$

We also have,

$$P_t(a) = \mathbb{E}_t[\theta_{t+1}(a)] - \theta_t(a) = \eta \cdot \pi_{\theta_t}(a) \cdot \left[r(a) - \pi_{\theta_t}^\top r\right]. \tag{169}$$

Now we apply Theorem 1 in Abbasi-Yadkori et al. (2011) to bound $Z_t(a)$. Fix any a. Let

$$\eta_t = \eta, \quad X_t = \pi_{\theta_t}(a) \cdot \left[r(a) - \pi_{\theta_t}^\top r\right], \tag{170}$$
$$\bar{V}_t = 1 + \sum_{s=1}^{t} X_s^2, \quad S_t = \sum_{s=1}^{t} \eta_s X_s, \tag{171}$$

then $\|\eta_t\| = \eta$ and hence is $\frac{\eta}{2}$-Sub-Gaussian. Consequently, there exists event $\mathcal{E}_1$ such that $\mathbb{P}(\mathcal{E}_1) \geq 1 - \delta$, and when $\mathcal{E}_1$ holds,

$$\|S_t\|_{\bar{V}_t^{-1}}^2 \leq \frac{\eta^2}{2} \log\left(\frac{\det(\bar{V}_t)^{1/2}}{\sqrt{2}\delta}\right), \tag{172}$$
$$\left|\sum_{s=1}^{t} P_s(a)\right| \leq \eta \cdot \sqrt{\frac{1 + S_t^1(a)}{2} \log\left(\frac{(1 + S_t^1(a))^{1/2}}{\sqrt{2}\delta}\right)}, \tag{173}$$

where $S_t^1(a) = \sum_{s=1}^{t-1} \pi_{\theta_s}(a)^2 \cdot \left(r(a) - \pi_{\theta_s}^\top r\right)^2$. Noted that $|R_t(a_t)| \leq R_{\max}$ for all t. Similarly, there exists event $\mathcal{E}_2$ such that $\mathbb{P}(\mathcal{E}_2) \geq 1 - \delta$, and when $\mathcal{E}_2$ holds,

$$\left|\sum_{s=1}^{t-1} I_t(a) \cdot R_s(a_s)\right| \leq R_{\max} \cdot \sqrt{\frac{1 + S_t^2(a)}{2} \log\left(\frac{(1 + S_t^2(a))^{1/2}}{\sqrt{2}\delta}\right)}, \tag{174}$$

where $S_t^2(a) = \sum_{s=1}^{t-1} I_s(a)^2$; there exists event $\mathcal{E}_3$ such that $\mathbb{P}(\mathcal{E}_3) \geq 1 - \delta$, and when $\mathcal{E}_3$ holds,

$$\left|\sum_{s=1}^{t-1} \pi_{\theta_s}(a) \cdot R_s(a_s)\right| \leq R_{\max} \cdot \sqrt{\frac{1 + S_t^3(a)}{2} \log\left(\frac{(1 + S_t^3(a))^{1/2}}{\sqrt{2}\delta}\right)}, \tag{175}$$

where $S_t^3(a) = \sum_{s=1}^{t-1} \pi_{\theta_s}(a)^2$. Let

$$A = \sum_{s=1}^{t-1} I_t(a), \quad B = \sum_{s=1}^{t-1} \pi_{\theta_s}(a), \tag{176}$$

Since $0 \leq I_t(a) \leq 1$ and $0 \leq \pi_{\theta_t}(a) \leq 1$ for all t, we have

$$S_t^2(a) \leq \sum_{s=1}^{t-1} |I_s(a)| = A, \quad S_t^3(a) \leq \sum_{s=1}^{t-1} |\pi_{\theta_s}(a)| = B. \tag{177}$$

When $\mathcal{E}_2$ and $\mathcal{E}_3$ hold, Eqs. (174) and (175) indicate that

$$|A-B| \leq A+B \leq \sqrt{\frac{-\log(\sqrt{2}\delta)}{2}} \cdot \left(\sqrt{A\log(1+A)}+\sqrt{B\log(1+B)}\right). \tag{178}$$

Let $C_1 = \sqrt{-\log(\sqrt{2}\delta)/2}$, we have

$$A \leq B + C_1 \cdot \sqrt{A\log(1+A)} + B + C_1 \cdot \sqrt{B\log(1+B)} \tag{179}$$

$$\left(\sqrt{A} - C_1 \cdot \frac{\log(1+A)}{2}\right)^2 \leq B + C_1 \cdot \sqrt{B\log(1+B)} + \frac{C_1^2 \cdot (\log(1+A))^2}{4}. \tag{180}$$

There exists $C_2 \geq C_1$, $A' > 10$ such that when $A \geq A'$, $C_1 \cdot \sqrt{\log(1+A)} < C_2 \cdot A^{1/4}$ and

$$\sqrt{A} \leq \sqrt{B + C_1 \cdot \sqrt{B\log(1+B)} + \frac{C_1^2 \cdot (\log(1+A))^2}{4}} + C_1 \cdot \frac{\log(1+A)}{2} \tag{181}$$

$$\sqrt{A} \leq \sqrt{(1+C_1)\cdot B + \frac{C_2^2 \cdot \sqrt{A}}{4}} + \frac{C_2 \cdot A^{1/4}}{2} \leq \sqrt{(1+C_1)\cdot B} + \frac{C_2 \cdot A^{1/4}}{2} + \frac{C_2 \cdot A^{1/4}}{2} \tag{182}$$

$$\sqrt{A} - C_2 \cdot A^{1/4} \leq \sqrt{(1+C_1)\cdot B} \tag{183}$$

$$A^{1/4} \leq \sqrt{\sqrt{(1+C_1)\cdot B} + \frac{C_2^2}{4}} + \frac{C_2}{4} \tag{184}$$

$$\sqrt{A} \leq 2\left(\sqrt{(1+C_1)\cdot B} + \frac{C_2}{4} + \frac{C_2}{4}\right) \tag{185}$$

$$A \leq 4\left((1+C_1)\cdot B + \frac{C_2}{4}\right) \leq 4C_1 \cdot B + 4 + C_2. \tag{186}$$

When $A \leq A'$,

$$A \leq C_1 \cdot \sqrt{A'\log(1+A')} + B + C_1 \cdot \sqrt{B\log(1+B)}. \tag{187}$$

Hence, whether $A \leq A'$ or not, we have

$$A \leq \max\{4\cdot C_1, 1+C_1\} \cdot B + \max\left\{4 + C_2, C_1 \cdot \sqrt{A'\log(1+A')}\right\}. \tag{188}$$

Moreover, there exists C_3 such that

$$\left|\sum_{s=1}^{t-1}(I_t(a) - \pi_{\theta_t}(a))\right| \leq C_3 \cdot R_{\max} \cdot \sqrt{\frac{1+\sum_{s=1}^{t-1}\pi_{\theta_s}\cdot R_s(a_s)}{2}\log\left(\frac{(1+\sum_{s=1}^{t-1}\pi_{\theta_s}\cdot R_s(a_s))^{1/2}}{\sqrt{2}\delta}\right)} \tag{189}$$

$$\leq C_3 \cdot R_{\max} \cdot \sqrt{\frac{1+S_t^4(a)}{2}\log\left(\frac{(1+S_t^4(a))^{1/2}}{\sqrt{2}\delta}\right)}, \tag{190}$$

where $S_t^4(a) = \sum_{s=1}^{t-1}\pi_{\theta_s}(a)$. Since C_1 and C_2 only depends on δ, C_3 only depends on δ. In other words, C_3 is a constant when δ is fixed. Since $I_t(a) - \pi_{\theta_t}(a) = (1-\pi_{\theta_t}(a)) - (1-I_t(a))$ and $0 \leq 1-\pi_{\theta_t}(a) \leq 1$, $0 \leq 1-I_t(a) \leq 1$, using the similar calculation to that of deriving Eq. (190), we can show that, there exists event $\mathcal{E}_4$ such that $\mathbb{P}(\mathcal{E}_4) \geq 1-2\delta$, and when $\mathcal{E}_4$ holds,

$$\left|\sum_{s=1}^{t-1}(I_t(a) - \pi_{\theta_s}(a))\right| \leq C_3 \cdot R_{\max} \cdot \sqrt{\frac{1+\sum_{s=1}^{t-1}\pi_{\theta_s}\cdot R_s(a_s)}{2}\log\left(\frac{(1+\sum_{s=1}^{t-1}\pi_{\theta_s}\cdot R_s(a_s))^{1/2}}{\sqrt{2}\delta}\right)} \tag{191}$$

$$\leq C_3 \cdot R_{\max} \cdot \sqrt{\frac{1+S_t^5(a)}{2}\log\left(\frac{(1+S_t^5(a))^{1/2}}{\sqrt{2}\delta}\right)}, \tag{192}$$

where $S_t^5(a) = \sum_{s=1}^{t-1}(1 - \pi_{\theta_s}(a))$.

Recall that i is the index of the (random) action $I \in [K]$ with

$$\lim_{t\to\infty} \sum_{j\in\mathcal{A}(I)} \pi_{\theta_t}(j) = 1, \qquad \text{a.s.} \tag{193}$$

As noted earlier we consider the event $\{I \neq a^*\}$, where a^* is the index of an optimal action and we will show that this event has zero probability. Since $\{I \neq a^*\} = \cup_{i\in[K]}\{I = i, i \neq a^*\}$, it suffices to show that for any fixed $i \in [K]$ index with $r(i) < r(a^*)$, $\{I = i, i \neq a^*\}$ has zero probability. Hence, in what follows we fix such a suboptimal action's index $i \in [K]$ and consider the event $\{I = i, i \neq a^*\}$.

Partition the action set $[K]$ into three parts using $r(i)$ as follows,

$$\mathcal{A}(i) \coloneqq \{j \in [K] : r(j) = r(i)\}, \qquad \text{(from Eq. (140))} \tag{194}$$
$$\mathcal{A}^+(i) \coloneqq \left\{a^+ \in [K] : r(a^+) > r(i)\right\}, \tag{195}$$
$$\mathcal{A}^-(i) \coloneqq \left\{a^- \in [K] : r(a^-) < r(i)\right\}. \tag{196}$$

Because i was the index of a sub-optimal action, we have $\mathcal{A}^+(i) \neq \emptyset$. According to Eq. (193), on $\{I = i\} \supset \{I = i, i \neq a^*\}$, we have $\pi_{\theta_t}^\top r \to r(i)$ as $t \to \infty$ because

$$\left|r(i) - \pi_{\theta_t}^\top r\right| = \left|\sum_{k\notin\mathcal{A}(i)} \pi_{\theta_t}(k) \cdot (r(i) - r(k))\right| \tag{197}$$
$$\leq \sum_{k\notin\mathcal{A}(i)} \pi_{\theta_t}(k) \cdot |r(i) - r(k)| \tag{198}$$
$$\leq 1 - \sum_{j\in\mathcal{A}(i)} \pi_{\theta_t}(j). \qquad \left(r \in [0,1]^K\right) \tag{199}$$

Therefore, there exists $\tau \geq 1$ such that almost surely on $\{I = i, i \neq a^*\}$ $\tau < \infty$ while we also have

$$r(a^+) - c' \geq \pi_{\theta_t}^\top r \geq r(a^-) + c', \qquad \text{for all } t \geq \tau, \tag{200}$$

for all $a^+ \in \mathcal{A}^+(i)$, $a^- \in \mathcal{A}^-(i)$, where $c' > 0$. Hence, for all $t \geq \tau$, $a^+ \in \mathcal{A}^+(i)$, $a^- \in \mathcal{A}^-(i)$, $P_t(a^+) > 0 > P_t(a^-)$. For all $a^+ \in \mathcal{A}^+(i)$, when $t > \tau$,

$$S_t^1(a^+) = \sum_{s=1}^{t-1} \pi_{\theta_s}(a^+)^2 \cdot (r(a^+) - \pi_{\theta_s}^\top r)^2 \leq R_{\max}^2 \cdot \sum_{s=1}^{t} \pi_{\theta_s}(a^+)^2 \leq \sum_{s=1}^{t-1} \pi_{\theta_s}(a^+) = R_{\max}^2 \cdot S_t^4(a^+), \tag{201}$$
$$\sum_{s=\tau}^{t} P_s(a^+) = \sum_{s=\tau}^{t} \eta \cdot \pi_{\theta_s}(a^+) \cdot (r(a^+) - \pi_{\theta_s}^\top r) \geq \eta \cdot c' \cdot \sum_{s=\tau}^{t} \pi_{\theta_s}(a^+). \tag{202}$$

Hence, when $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ hold, we have

$$\theta_t(a^+) = \mathbb{E}[\theta_1(a^+)] + Z_t(a^+) + P_1(a^+) + \cdots + P_{\tau-1}(a^+) + P_\tau(a^+) + \cdots + P_{t-1}(a^+) \qquad \big(\text{by Eq. (160)}\big) \tag{203}$$

$$\geq \mathbb{E}[\theta_1(a^+)] - \eta \cdot \sqrt{\frac{1+S_t^1(a^+)}{2} \log\left(\frac{(1+S_t^1(a^+))^{1/2}}{\sqrt{2}\delta}\right)} \tag{204}$$

$$- \eta \cdot C_3 \cdot R_{\max} \cdot \sqrt{\frac{1+S_t^4(a^+)}{2} \log\left(\frac{(1+S_t^4(a^+))^{1/2}}{\sqrt{2}\delta}\right)} \tag{205}$$

$$+ P_1(a^+) + \cdots + P_{\tau-1}(a^+) + P_\tau(a^+) + \cdots + P_{t-1}(a^+) \qquad \text{(by Eqs. (173) and (190))} \tag{206}$$

$$\geq \mathbb{E}[\theta_1(a^+)] - \underbrace{\left\{\eta \cdot R_{\max}^2 \cdot (1+C_3) \cdot \sqrt{\frac{1+\sum_{s=1}^{t-1}\pi_{\theta_s}(a^+)}{2} \log\left(\frac{(1+\sum_{s=1}^{t-1}\pi_{\theta_s}(a^+))^{1/2}}{\sqrt{2}\delta}\right)}\right\}}_{(\spadesuit)} \tag{207}$$

$$+ P_1(a^+) + \cdots + P_{\tau-1}(a^+) + \underbrace{\eta \cdot c' \cdot \sum_{s=\tau}^{t} \pi_{\theta_s}(a^+)}_{(\heartsuit)}. \tag{208}$$

If $\sum_{s=1}^{\infty}\pi_{\theta_s}(a^+) < \infty$, $\theta_t(a^+)$ is always finite and $\inf_{t\geq 1}\theta_t(a^+) > -\infty$; if $\sum_{s=1}^{\infty}\pi_{\theta_s}(a^+) = \infty$, we have $(\heartsuit)$ goes to ∞ faster than $(\spadesuit)$, and $\inf_{t\geq 1}\theta_t(a^+) > -\infty$.

Now take any $\omega \in \mathcal{E} := \{I = i, i \neq a^*\}$. Because $\mathbb{P}(\mathcal{E} \setminus (\mathcal{E}\cap\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3)) \leq \mathbb{P}(\Omega \setminus (\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3)) \leq 3\delta \to 0$ as $\delta \to 0$, we have that $\mathbb{P}$-almost surely for all $\omega \in \mathcal{E}'$ there exists $\delta > 0$ such that $\omega \in \mathcal{E}\cap\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$ while Eq. (208) also holds for this δ. Take such a δ. By Eq. (208),

$$\inf_{t\geq 1}\theta_t(a^+)(\omega) > -\infty. \tag{209}$$

Hence, almost surely on $\mathcal{E}$,

$$c_1(a^+) := \inf_{t\geq 1}\theta_t(a^+) > -\infty. \tag{210}$$

Furthermore,

$$c_1 := \min_{a^+\in\mathcal{A}^+}\inf_{t\geq 1}\theta_t(a^+) = \min_{a^+\in\mathcal{A}^+} c_1(a^+) > -\infty. \tag{211}$$

Similarly, we can show that, almost surely on $\mathcal{E}$,

$$c_2 := \max_{a^-\in\mathcal{A}^-}\sup_{t\geq 1}\theta_t(a^-) < \infty. \tag{212}$$

First case. 2a). Consider the event,

$$\mathcal{E}_0 := \bigcap_{a^+\in\mathcal{A}^+(i)} \underbrace{\{N_\infty(a^+) < \infty\}}_{\mathcal{E}_0(a^+)}, \tag{213}$$

i.e., any "good" action $a^+ \in \mathcal{A}^+(i)$ has finitely many updates as $t \to \infty$. Pick $a^+ \in \mathcal{A}^+(i)$, such that $\mathbb{P}(N_\infty(a^+) < \infty) > 0$. According to the extended Borel-Cantelli lemma (Lemma C.2), we have, almost surely,

$$\Big\{\sum_{t\geq 1}\pi_{\theta_t}(a^+) = \infty\Big\} = \big\{N_\infty(a^+) = \infty\big\}. \tag{214}$$

Hence, taking complements, we have,

$$\Big\{\sum_{t\geq 1}\pi_{\theta_t}(a^+) < \infty\Big\} = \big\{N_\infty(a^+) < \infty\big\} \tag{215}$$

also holds almost surely. Note that, for all "good" action a^+,

$$\theta_{t+1}(a^+) \leftarrow \theta_t(a^+) + \begin{cases} \eta \cdot (1 - \pi_{\theta_t}(a^+)) \cdot R_t(a^+), & \text{if } a_t = a^+ \,, \\ -\eta \cdot \pi_{\theta_t}(a^+) \cdot R_t(a_t), & \text{otherwise} \,. \end{cases} \tag{216}$$

Since the first update will be conducted finitely many times, and the second update will be conducted for infinitely many times, we have

$$c_3 \coloneqq \sup_{t \geq 1} \theta_t(a^+) < \infty. \tag{217}$$

Next,

$$1 - \sum_{j \in \mathcal{A}(i)} \pi_{\theta_t}(j) = \frac{\sum_{a^+ \in \mathcal{A}^+(i)} e^{\theta_t(a^+)} + \sum_{a^- \in \mathcal{A}^-(i)} e^{\theta_t(a^-)}}{\sum_{a \in [K]} e^{\theta_t(a)}} \tag{218}$$

$$\leq \frac{\sum_{a^+ \in \mathcal{A}^+(i)} e^{\theta_t(a^+)} + \sum_{a^- \in \mathcal{A}^-(i)} e^{c_2}}{\sum_{a \in [K]} e^{\theta_t(a)}} \tag{219}$$

$$= \frac{\sum_{a^+ \in \mathcal{A}^+(i)} e^{\theta_t(a^+)} + e^{c_2 - c_1} \cdot \sum_{a^- \in \mathcal{A}^-(i)} e^{c_1}}{\sum_{a \in [K]} e^{\theta_t(a)}} \tag{220}$$

$$= \frac{\sum_{a^+ \in \mathcal{A}^+(i)} e^{\theta_t(a^+)} + e^{c_2 - c_1} \cdot \frac{|\mathcal{A}^-(i)|}{|\mathcal{A}^+(i)|} \cdot \sum_{a^+ \in \mathcal{A}^+(i)} e^{c_1}}{\sum_{a \in [K]} e^{\theta_t(a)}} \qquad \left(|\mathcal{A}^+(i)| \geq 1\right) \tag{221}$$

$$\leq \frac{\sum_{a^+ \in \mathcal{A}^+(i)} e^{\theta_t(a^+)} + e^{c_2 - c_1} \cdot \frac{|\mathcal{A}^-(i)|}{|\mathcal{A}^+(i)|} \cdot \sum_{a^+ \in \mathcal{A}^+(i)} e^{\theta_t(a^+)}}{\sum_{a \in [K]} e^{\theta_t(a)}} \tag{222}$$

$$= \left(1 + e^{c_2 - c_1} \cdot \frac{|\mathcal{A}^-(i)|}{|\mathcal{A}^+(i)|}\right) \cdot \sum_{a^+ \in \mathcal{A}^+(i)} \pi_{\theta_t}(a^+). \tag{223}$$

According to Eq. (214), $N_\infty(a^+) < \infty$ for all $a^+ \in \mathcal{A}^+(i)$. Since

$$\sum_{t \geq 1} \sum_{a^- \in \mathcal{A}^-(i)} \pi_{\theta_t}(a^-) \leq e^{c_2 - c_1} \cdot \frac{|\mathcal{A}^-(i)|}{|\mathcal{A}^+(i)|} \cdot \sum_{t \geq 1} \sum_{a^+ \in \mathcal{A}^+(i)} \pi_{\theta_t}(a^+) < \infty, \tag{224}$$

we have $N_\infty(a^-) < \infty$ for all $a^- \in \mathcal{A}^-(i)$. Therefore. $\sum_{j \in \mathcal{A}(i)} N_\infty(i) = \infty$, which indicates that for all $j \in \mathcal{A}(i)$, the first update of i will be conducted for infinitely many times, and the second update will be conducted for finitely many times. According to Assumption 2.1, we have $|\mathcal{A}(i)| = 1$, and for all $j \in \mathcal{A}(i)$, we have

$$\theta_t(j) \leq \theta_1(j) + \eta \cdot r(j) \cdot \sum_{s=1}^{t-1} (1 - \pi_{\theta_s}(j)) \tag{225}$$

$$\leq \theta_1(j) + \eta \cdot r(j) \cdot \left(1 + e^{c_2 - c_1} \cdot \frac{|\mathcal{A}^-(i)|}{|\mathcal{A}^+(i)|}\right) \cdot \sum_{t \geq 1} \sum_{a^+ \in \mathcal{A}^+(i)} \pi_{\theta_t}(a^+) < \infty. \tag{226}$$

Hence, we have

$$c_4 \coloneqq \limsup_{t \to \infty} \theta_t(j) < \infty. \tag{227}$$

Therefore, we have

$$\sum_{j\in\mathcal{A}(i)}\pi_{\theta_t}(j) = \frac{\sum_{j\in\mathcal{A}(i)} e^{\theta_t(j)}}{\sum_{j\in\mathcal{A}(i)} e^{\theta_t(j)} + \sum_{a^+\in\mathcal{A}^+(i)} e^{\theta_t(a^+)} + \sum_{a^-\in\mathcal{A}^-(i)} e^{\theta_t(a^-)}} \tag{228}$$

$$\leq \frac{\sum_{j\in\mathcal{A}(i)} e^{\theta_t(j)}}{\sum_{j\in\mathcal{A}(i)} e^{\theta_t(j)} + \sum_{a^+\in\mathcal{A}^+(i)} e^{\theta_t(a^+)}} \qquad \left(e^{\theta_t(a^-)} > 0\right) \tag{229}$$

$$\leq \frac{\sum_{j\in\mathcal{A}(i)} e^{\theta_t(j)}}{\sum_{j\in\mathcal{A}(i)} e^{\theta_t(j)} + e^{c_2}\cdot|\mathcal{A}^+(i)|} \qquad \text{(by Eq. (212))} \tag{230}$$

$$\leq \frac{e^{c_4}\cdot|\mathcal{A}(i)|}{e^{c_4}\cdot|\mathcal{A}(i)| + e^{c_2}\cdot|\mathcal{A}^+(i)|} \qquad \text{(by Eq. (227))} \tag{231}$$

$$\not\to 1, \tag{232}$$

which is a contradiction with the assumption of Eq. (193), showing that $\mathbb{P}(\mathcal{E}_0\cap\mathcal{E}')=0$.

Second case. 2b). Consider the complement $\mathcal{E}_0^c$ of $\mathcal{E}_0$, where $\mathcal{E}_0$ is by Eq. (213). $\mathcal{E}_0^c$ indicates the event for at least one "good" action $a^+\in\mathcal{A}^+(i)$ has infinitely many updates as $t\to\infty$.

We now show that also $\mathbb{P}(\mathcal{E}')=0$ where $\mathcal{E}'=\mathcal{E}_0^c\cap\{I=i, i\neq a^*\} = (\cup_{a^+\in\mathcal{A}(i)}\{N_\infty(a^+)=\infty\})\cap\{I=i,i\neq a^*\}$.[3] Let $\tilde{\mathcal{A}}^+(i) \coloneqq \{a^+\in\mathcal{A}^+(i) : N_\infty(a^+)=\infty\}$, and

$$\mathcal{E}_\infty(a^+) \coloneqq \cup_{a^+\in\mathcal{A}(i)}\{N_\infty(a^+)=\infty\} = \cap_{a^+\in\tilde{\mathcal{A}}^+(i)}\{N_\infty(a^+)=\infty\}. \tag{233}$$

Then $\mathcal{E}'=\mathcal{E}_\infty(a^+)\cap\{I=i,i\neq a^*\}$. Since $\mathcal{E}'\subset\mathcal{E}_\infty(a^+)$, the statement follows if $\mathbb{P}(\mathcal{E}_\infty(a^+))=0$. Hence, assume that $\mathbb{P}(\mathcal{E}_\infty(a^+))>0$.

Fix $\delta\in[0,1]$. Using a similar calculation to that of Eq. (208), there exists an event $\mathcal{E}_\delta$ such that $\mathbb{P}(\mathcal{E}_\delta)\geq 1-2\delta$, and on $\mathcal{E}_\delta$, for all $t\geq\tau$, for all $a^+\in\tilde{\mathcal{A}}^+(i)$,

$$\theta_t(a^+) = \mathbb{E}[\theta_1(a^+)] + Z_t(a^+) + P_1(a^+) + \cdots + P_{\tau-1}(a^+) \qquad \left(\text{by Eq. (160)}\right) \tag{234}$$

$$+ P_\tau(a^+) + \cdots + P_{t-1}(a^+) \tag{235}$$

$$\geq \mathbb{E}[\theta_1(a^+)] - \underbrace{\left\{\eta\cdot R_{\max}^2\cdot(1+C_3)\cdot\sqrt{\frac{1+\sum_{s=1}^{t-1}\pi_{\theta_s}(a^+)}{2}\log\left(\frac{(1+\sum_{s=1}^{t-1}\pi_{\theta_s}(a^+))^{1/2}}{\sqrt{2}\delta}\right)}\right\}}_{(\spadesuit)} \tag{236}$$

$$+ P_1(a^+) + \cdots + P_{\tau-1}(a^+) + \underbrace{\eta\cdot c'\cdot\sum_{s=\tau}^{t}\pi_{\theta_s}(a^+)}_{(\heartsuit)} \qquad \text{(by Eqs. (173) and (190))}\,. \tag{237}$$

On $\mathcal{E}_\infty(a^+)\cap\mathcal{E}_\delta$, $N_{t-1}(a^+)\to\infty$ as $t\to\infty$, which with Eq. (214) indicates that $\sum_{t\geq1}\pi_{\theta_t}(a^+)=\infty$. When $t\to\infty$, both ($\spadesuit$) and ($\heartsuit$) go to infinity while ($\heartsuit$) goes to infinity faster than ($\spadesuit$). Hence, we have $\theta_t(a^+)\to\infty$ as $t\to\infty$.

Since $\mathbb{P}(\mathcal{E}_\infty(a^+)\setminus(\mathcal{E}_\infty(a^+)\cap\mathcal{E}_\delta))\to 0$ as $\delta\to 0$, with an argument parallel to that used in the previous analysis (cf. the argument after Eq. (208)), we have, almost surely on $\mathcal{E}_\infty(a^+)$,

$$\lim_{t\to\infty}\theta_t(a^+)=\infty, \tag{238}$$

which implies that there exists $\tau_1\geq 1$ such that on $\mathcal{E}'(=\mathcal{E}_\infty(a^+)\cap\{I=i,i\neq a^*\})$ we have almost surely that $\tau_1<+\infty$ while we also have that for all $t\geq\tau_1$, for all $a^+\in\tilde{\mathcal{A}}^+(i)$,

$$\sum_{a^-\in\mathcal{A}^-(i)}\frac{r(i)-r(a^-)}{\exp\{\theta_t(a^+)-c_1\}} < c' \coloneqq \frac{r(a^+)-r(i)}{2}\,. \tag{239}$$

[3] Here, $\mathcal{E}'$ is redefined to minimize clutter; the previous definition is not used in this part of the proof.

For all $\bar{a}^+ \in \mathcal{A}^+(i)/\tilde{\mathcal{A}}^+(i)$, since $\lim_{t\to\infty} \bar{a}^+ < \infty$, we have

$$\max_{\bar{a}^+ \in \mathcal{A}^+(i)/\tilde{\mathcal{A}}^+(i)} \theta_t(\bar{a}^+) < \infty. \tag{240}$$

Recall in Eq. (212) we show $c_2 = \max_{a^- \in \mathcal{A}^-} \sup_{t\geq 1} \theta_t(a^-) < \infty$. We have

$$\max_{a \in (\mathcal{A}^+ \cup \mathcal{A}^-)/\tilde{\mathcal{A}}^+(i)} \theta_t(a) < \infty. \tag{241}$$

Fix any $\tilde{a}^+ \in \tilde{\mathcal{A}}^+$, since $\lim_{t\to\infty} \theta_t(a^+) = \infty$, there exists $\tau_2 < \infty$ such that when $t \geq \tau_2$,

$$\sum_{a \in (\mathcal{A}^+ \cup \mathcal{A}^-)/\tilde{\mathcal{A}}^+(i)} \exp\{\theta_t(a)\} \leq 0.1 \cdot \exp\{\theta_t(\tilde{a}^+)\} \tag{242}$$

$$\sum_{a \in (\mathcal{A}^+ \cup \mathcal{A}^-)/\tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(a) \leq 0.1 \cdot \pi_{\theta_t}(\tilde{a}^+) \tag{243}$$

$$1 - \sum_{j\in\mathcal{A}(i)} \pi_t(j) = \sum_{a \in (\mathcal{A}^+ \cup \mathcal{A}^-)/\tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(a) + \sum_{a^+ \in \tilde{\mathcal{A}}^+} \pi_{\theta_t}(a^+) \leq 0.1 \cdot \pi_{\theta_t}(\tilde{a}^+) + \sum_{a^+ \in \tilde{\mathcal{A}}^+} \pi_{\theta_t}(a^+) \leq 1.1 \cdot \sum_{a^+ \in \tilde{\mathcal{A}}^+} \pi_{\theta_t}(a^+). \tag{244}$$

Hence, on $\mathcal{E}'$, for $t \geq \tau_1$, almost surely,

$$\pi_{\theta_t}^\top r = \sum_{j\in\mathcal{A}(i)} \pi_{\theta_t}(j) \cdot r(i) + \sum_{a^- \in \mathcal{A}^-(i)} \pi_{\theta_t}(a^-) \cdot r(a^-) + \sum_{a^+ \in \mathcal{A}^+(i)} \pi_{\theta_t}(a^+) \cdot r(a^+) \tag{245}$$

$$= r(i) - \sum_{a^- \in \mathcal{A}^-(i)} \pi_{\theta_t}(a^-) \cdot \big(r(i) - r(a^-)\big) + \sum_{a^+ \in \mathcal{A}^+(i)} \pi_{\theta_t}(a^+) \cdot \big(r(a^+) - r(i)\big) \tag{246}$$

$$\geq r(i) - \sum_{a^- \in \mathcal{A}^-(i)} \pi_{\theta_t}(a^-) \cdot \big(r(i) - r(a^-)\big) + \sum_{\tilde{a}^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(\tilde{a}^+) \cdot \big(r(\tilde{a}^+) - r(i)\big) \tag{247}$$

$$\big(r(a^+) - r(i) > 0, \text{ Eq. (195)}\big) \tag{248}$$

$$= r(i) + \sum_{\tilde{a}^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(\tilde{a}^+) \cdot \left[\big(r(a^+) - r(i)\big) - \sum_{a^- \in \mathcal{A}^-(i)} \frac{\pi_{\theta_t}(a^-)}{\pi_{\theta_t}(\tilde{a}^+) \cdot |\tilde{\mathcal{A}}^+(i)|} \cdot \big(r(i) - r(a^-)\big) \right] \tag{249}$$

$$\geq r(i) + \sum_{\tilde{a}^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(\tilde{a}^+) \cdot \left[\big(r(a^+) - r(i)\big) - \sum_{a^- \in \mathcal{A}^-(i)} \frac{\pi_{\theta_t}(a^-)}{\pi_{\theta_t}(\tilde{a}^+)} \cdot \big(r(i) - r(a^-)\big) \right] \tag{250}$$

$$= r(i) + \sum_{\tilde{a}^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(\tilde{a}^+) \cdot \left[\big(r(\tilde{a}^+) - r(i)\big) - \sum_{a^- \in \mathcal{A}^-(i)} \frac{r(i) - r(a^-)}{\exp\{\theta_t(\tilde{a}^+) - \theta_t(a^-)\}} \right] \tag{251}$$

$$\geq r(i) + \sum_{a^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(a^+) \cdot \left[\big(r(a^+) - r(i)\big) - \sum_{a^- \in \mathcal{A}^-(i)} \frac{r(i) - r(a^-)}{\exp\{\theta_t(a^+) - c_2\}} \right] \qquad \text{(by Eq. (212))} \tag{252}$$

$$> r(i) + c' \cdot \sum_{a^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(a^+)\,. \qquad \text{(by Eq. (239))} \tag{253}$$

Therefore, on $\mathcal{E}'$, for all $t \geq \tau_1$, for any $j \in \mathcal{A}(i)$, almost surely, by Eq. (169), we have

$$P_t(j) = \eta \cdot \pi_{\theta_t}(j) \cdot (r(j) - \pi_{\theta_t}^\top r) < -c' \cdot \sum_{a^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(i) \cdot \pi_{\theta_t}(a^+) < 0. \tag{254}$$

Since $|\mathcal{A}(i)| = 1$, by Assumption 2.1, $\lim_{t\to\infty} \pi_{\theta_t}(i) = 1$, there exists τ_2 such that when $t \geq \tau_3$, $\pi_{\theta_t}(i) > 1/2$. Hence, when $t \geq \max\{\tau_1, \tau_3\}$, we have $P_t(j) < -c' \cdot \sum_{a^+ \in \tilde{\mathcal{A}}^+(i)} \pi_{\theta_t}(a^+)/2$. Let $\tau' = \max\{\tau_1, \tau_2, \tau_3\}$. When $t > \tau'$, for

$j \in \mathcal{A}(i)$,

$$S_t^1(j) = \sum_{s=1}^{t-1} \pi_{\theta_s}(j)^2 \cdot (r(j) - \pi_{\theta_s}^\top r)^2 = \sum_{s=1}^{t-1} \left(\frac{P_s(j)}{\eta}\right)^2, \tag{255}$$

$$S_t^5(j) = \sum_{s=1}^{t-1} (1 - \pi_{\theta_s}(j)) = \sum_{s=1}^{\tau'-1} (1 - \pi_{\theta_s}(j)) + \sum_{s=\tau'}^{t-1} \Big[1 - \sum_{j \in \mathcal{A}(i)} \pi_{\theta_s}(j)\Big] \tag{256}$$

$$\leq \sum_{s=1}^{\tau'-1} (1 - \pi_{\theta_s}(j)) + 1.1 \cdot \sum_{s=\tau'}^{t-1} \sum_{a^+ \in \tilde{\mathcal{A}}^+} \pi_{\theta_s}(a^+). \qquad \text{(by Eq. (244))} \tag{257}$$

Hence, for $t \geq \tau'$, when $\mathcal{E}_1$ and $\mathcal{E}_4$ hold, we have

$$\theta_t(j) = \mathbb{E}[\theta_1(j)] + Z_t(j) + P_1(j) + \cdots + P_{\tau-1}(j) + P_\tau(a^+) + \cdots + P_{t-1}(j) \qquad \big(\text{by Eq. (160)}\big) \tag{258}$$

$$\leq \mathbb{E}[\theta_1(j)] + \eta \cdot \sqrt{\frac{1 + S_t^1(j)}{2} \log\left(\frac{(1 + S_t^1(j))^{1/2}}{\sqrt{2}\delta}\right)} + \eta \cdot C_3 \cdot R_{\max} \cdot \sqrt{\frac{1 + S_t^5(j)}{2} \log\left(\frac{(1 + S_t^5(j))^{1/2}}{\sqrt{2}\delta}\right)} \tag{259}$$

$$+ P_1(j) + \cdots + P_{\tau'-1}(j) + P_{\tau'}(j) + \cdots + P_{t-1}(j) \qquad \text{(by Eqs. (173) and (192))} \tag{260}$$

$$\leq \mathbb{E}[\theta_1(j)] + \underbrace{\eta \cdot \sqrt{\frac{1 + S_t^1(j)}{2} \log\left(\frac{(1 + S_t^1(j))^{1/2}}{\sqrt{2}\delta}\right)}}_{(\spadesuit)} + \underbrace{\eta \cdot C_3 \cdot R_{\max} \cdot \sqrt{\frac{1 + S_t^5(j)}{2} \log\left(\frac{(1 + S_t^5(j))^{1/2}}{\sqrt{2}\delta}\right)}}_{(\heartsuit)} \tag{261}$$

$$+ P_1(j) + \cdots + P_{\tau'-1}(j) - \underbrace{\frac{1}{2} \cdot \sum_{j=\tau'}^{t} |P_s(j)|}_{(\clubsuit)} - \underbrace{\frac{c'}{4} \cdot \sum_{a^+ \in \tilde{A}^+} \pi_{\theta_t}(a^+)}_{(\diamondsuit)}. \tag{262}$$

Since ($\clubsuit$) and ($\diamondsuit$) go to infinity when $t \to \infty$, ($\clubsuit$) goes to infinity faster than ($\spadesuit$), ($\diamondsuit$) goes to infinity faster than ($\heartsuit$), we have $\sup_{t \geq 1} \theta_t(j) < \infty$.

Let $\mathcal{E}'_\delta = \mathcal{E}_1 \cap \mathcal{E}_4$. Since $\mathbb{P}(\mathcal{E}'^c_\delta) \leq 2\delta \to 0$ as $\delta \to 0$, with an argument parallel to that used in the previous analysis (cf. the argument after Eq. (208)), we get that there exists a random constant $c_5(j)$ such that almost surely on $\mathcal{E}'$, $c_5(j) < \infty$ and $\sup_{t \geq \tau_1} \theta_t(j) \leq c_5(j)$. Define $c_5 \coloneqq \max_{j \in \mathcal{A}(i)} c_5(j)$. Then, almost surely on $\mathcal{E}'$, $c_5 < \infty$ and

$$\sup_{t \geq \tau_1} \max_{j \in \mathcal{A}(i)} \theta_t(j) \leq c_5\,. \tag{263}$$

By Eq. (238), there exists $a^+ \in \mathcal{A}^+(i)$, $\tau'' \geq 1$, such that almost surely on $\mathcal{E}'$, $\tau'' < \infty$ while we also have

$$\inf_{t \geq \tau''} \theta_t(a^+) \geq 0, \tag{264}$$

for all $t \geq \tau''$. Hence, on $\mathcal{E}'$, almost surely for all $t \geq \max(\tau', \tau'')$,

$$\sum_{j \in \mathcal{A}(i)} \pi_{\theta_t}(j) = \frac{\sum_{j \in \mathcal{A}(i)} e^{\theta_t(j)}}{\sum_{j \in \mathcal{A}(i)} e^{\theta_t(j)} + \sum_{\tilde{a}^+ \in \mathcal{A}^+(i)} e^{\theta_t(\tilde{a}^+)} + \sum_{a^- \in \mathcal{A}^-(i)} e^{\theta_t(a^-)}} \tag{265}$$

$$\leq \frac{\sum_{j \in \mathcal{A}(i)} e^{\theta_t(j)}}{\sum_{j \in \mathcal{A}(i)} e^{\theta_t(j)} + e^{\theta_t(a^+)}} \qquad \big(e^{\theta_t(k)} > 0 \text{ for any } k \in [K]\big) \tag{266}$$

$$\leq \frac{\sum_{j \in \mathcal{A}(i)} e^{\theta_t(j)}}{\sum_{j \in \mathcal{A}(i)} e^{\theta_t(j)} + 1} \qquad \text{(by Eq. (264))} \tag{267}$$

$$\leq \frac{e^{c_5} \cdot |\mathcal{A}(i)|}{e^{c_5} \cdot |\mathcal{A}(i)| + 1} \qquad \text{(by Eq. (263))} \tag{268}$$

$$\not\to 1\,. \tag{269}$$

Hence, $\mathbb{P}(\mathcal{E}') = 0$, finishing the proof. □

Lemma 5.4 (Non-uniform Łojasiewicz (NŁ), Mei et al. (2020b, Lemma 3))**.** Assume r has a unique maximizing action a^*. Let $\pi^* = \arg\max_{\pi \in \Delta} \pi^\top r$. Then,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2 \geq \pi_\theta(a^*) \cdot (\pi^* - \pi_\theta)^\top r \,. \tag{270}$$

Proof. Using the definition of softmax Jacobian, we have,

$$\left\| \frac{d\pi_\theta^\top r}{d\theta} \right\|_2^2 = \sum_{a \in [K]} \pi_\theta(a)^2 \cdot \left(r(a) - \pi_\theta^\top r\right)^2 \tag{271}$$

$$\geq \pi_\theta(a^*)^2 \cdot \left(r(a^*) - \pi_\theta^\top r\right)^2, \qquad \text{(fewer terms)} \tag{272}$$

which implies Eq. (270). □

A.2. Proof of Theorem 5.5

Theorem 5.5 (Convergence rate and regret)**.** Using Algorithm 1 with $\eta = \frac{\Delta^2}{40 \cdot K^{3/2} \cdot R_{\max}^3}$, we have, for all $t \geq 1$,

$$\mathbb{E}[(\pi^* - \pi_{\theta_t})^\top r] \leq \frac{C}{t}, \quad \text{and} \tag{273}$$

$$\mathbb{E}\left[\sum_{t=1}^{T} (\pi^* - \pi_{\theta_t})^\top r\right] \leq \min\{\sqrt{2\, R_{\max}\, C\, T}, C\, \log T + 1\}, \tag{274}$$

where $C \coloneqq \frac{80 \cdot K^{3/2} \cdot R_{\max}^3}{\Delta^2 \cdot \mathbb{E}[c^2]}$, and $c \coloneqq \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0$ is from Theorem 5.1.

Proof. **First part, Eq. (273).** According to Lemma 4.6, we have,

$$\mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] - \pi_{\theta_t}^\top r \geq \frac{\Delta^2}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \left\| \frac{d\pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \tag{275}$$

$$\geq \frac{\Delta^2 \cdot \pi_{\theta_t}(a^*)^2}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right)^2 \qquad \text{(by Lemma 5.4)} \tag{276}$$

$$\geq \frac{\Delta^2 \cdot \inf_{t \geq 1} \pi_{\theta_t}(a^*)^2}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right)^2 \tag{277}$$

$$= \frac{\Delta^2 \cdot c^2}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \left(r(a^*) - \pi_{\theta_t}^\top r\right)^2. \qquad \text{(by Theorem 5.1)} \tag{278}$$

Denote $\delta(\theta_t) \coloneqq (\pi^* - \pi_{\theta_t})^\top r$ as the sub-optimality gap. We have,

$$\delta(\theta_t) - \mathbb{E}_t[\delta(\theta_{t+1})] = (\pi^* - \pi_{\theta_t})^\top r - \left(\pi^* - \mathbb{E}_t[\pi_{\theta_{t+1}}]\right)^\top r \tag{279}$$

$$= \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] - \pi_{\theta_t}^\top r \tag{280}$$

$$\geq \frac{\Delta^2 \cdot c^2}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \delta(\theta_t)^2. \tag{281}$$

Taking expectation, we have,

$$\mathbb{E}\left[\delta(\theta_t)\right] - \mathbb{E}\left[\delta(\theta_{t+1})\right] \geq \frac{\Delta^2 \cdot \mathbb{E}[c^2]}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \mathbb{E}\left[\delta(\theta_t)^2\right] \qquad \left(c \coloneqq \inf_{t \geq 1} \pi_{\theta_t}(a^*) > 0 \text{ is independent with } t\right) \tag{282}$$

$$\geq \frac{\Delta^2 \cdot \mathbb{E}[c^2]}{80 \cdot K^{3/2} \cdot R_{\max}^3} \cdot \left(\mathbb{E}\left[\delta(\theta_t)\right]\right)^2 \qquad \text{(by Jensen's inequality)} \tag{283}$$

$$= \frac{1}{C} \cdot \left(\mathbb{E}\left[\delta(\theta_t)\right]\right)^2. \tag{284}$$

Therefore, we have,

$$\frac{1}{\mathbb{E}\left[\delta(\theta_t)\right]} = \frac{1}{\mathbb{E}\left[\delta(\theta_1)\right]} + \sum_{s=1}^{t-1}\left[\frac{1}{\mathbb{E}\left[\delta(\theta_{s+1})\right]} - \frac{1}{\mathbb{E}\left[\delta(\theta_s)\right]}\right] \tag{285}$$

$$= \frac{1}{\mathbb{E}\left[\delta(\theta_1)\right]} + \sum_{s=1}^{t-1}\frac{1}{\mathbb{E}\left[\delta(\theta_{s+1})\right]\cdot\mathbb{E}\left[\delta(\theta_s)\right]}\cdot\left(\mathbb{E}\left[\delta(\theta_s)\right] - \mathbb{E}\left[\delta(\theta_{s+1})\right]\right) \tag{286}$$

$$\geq \frac{1}{\mathbb{E}\left[\delta(\theta_1)\right]} + \sum_{s=1}^{t-1}\frac{1}{\mathbb{E}\left[\delta(\theta_{s+1})\right]\cdot\mathbb{E}\left[\delta(\theta_s)\right]}\cdot\frac{1}{C}\cdot\left(\mathbb{E}\left[\delta(\theta_s)\right]\right)^2 \qquad \text{(by Eq. (282))} \tag{287}$$

$$\geq \frac{1}{\mathbb{E}\left[\delta(\theta_1)\right]} + \sum_{s=1}^{t-1}\frac{1}{C} \qquad \left(\mathbb{E}\left[\delta(\theta_s)\right] \geq \mathbb{E}\left[\delta(\theta_{s+1})\right] > 0\right) \tag{288}$$

$$= \frac{1}{\mathbb{E}\left[\delta(\theta_1)\right]} + \frac{1}{C}\cdot(t-1) \tag{289}$$

$$\geq \frac{t}{C}, \qquad \left(\mathbb{E}\left[\delta(\theta_1)\right] \leq 2\,R_{\max} \leq C = \max_{s\leq t}\frac{80\cdot K^{3/2}\cdot R_{\max}^3}{\Delta^2\cdot\mathbb{E}[c^2]}\right) \tag{290}$$

which implies Eq. (273).

Second part, Eq. (274). According to Eq. (285),we have,

$$\mathbb{E}\left[\sum_{t=1}^{T}\delta(\theta_t)\right] = \sum_{t=1}^{T}\mathbb{E}[\delta(\theta_t)] \leq \sum_{t=1}^{T}\frac{C}{t} \leq C\cdot\log T + 1. \tag{291}$$

On the other hand, we have

$$\sum_{t=1}^{T}\mathbb{E}[\delta(\theta_t)] \leq \sqrt{T}\cdot\left[\sum_{t=1}^{T}\left(\mathbb{E}\left[\delta(\theta_t)\right]\right)^2\right]^{\frac{1}{2}} \qquad \text{(by Cauchy–Schwarz)} \tag{292}$$

$$\leq \sqrt{T}\cdot\left[\sum_{t=1}^{T}C\cdot\left(\mathbb{E}[\delta(\theta_t)] - \mathbb{E}[\delta(\theta_{t+1})]\right)\right]^{\frac{1}{2}} \qquad \text{(by Eq. (282))} \tag{293}$$

$$= \sqrt{C\cdot T\cdot\left(\mathbb{E}[\delta(\theta_1)] - \mathbb{E}[\delta(\theta_{T+1})]\right)} \tag{294}$$

$$\leq \sqrt{C\cdot T\cdot 2\cdot R_{\max}}, \qquad \left(\mathbb{E}[\delta(\theta_{T+1})] \geq 0.\text{ and }\mathbb{E}[\delta(\theta_1)] \leq 2\,R_{\max}\right) \tag{295}$$

Combining Eqs. (291) and (292), we have Eq. (274). □

B. Proofs for Using Baselines

The following Algorithm 2 is same as the gradient bandit algorithm in Sutton & Barto (2018, Section 2.8).

Proposition B.1. *Algorithm 2 is equivalent to the following stochastic gradient ascent update on $\pi_\theta^\top r$.*

$$\theta_{t+1} \leftarrow \theta_t + \eta\cdot\frac{d\pi_{\theta_t}^\top\left(\hat{r}_t - \hat{b}_t\right)}{d\theta_t} \tag{296}$$

$$= \theta_t + \eta\cdot\left(diag(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right)\left(\hat{r}_t - \hat{b}_t\right), \tag{297}$$

where $\left(\frac{d\pi_\theta}{d\theta}\right)^\top = diag(\pi_\theta) - \pi_\theta\pi_\theta^\top$ is the Jacobian of $\theta \mapsto \pi_\theta \coloneqq \mathrm{softmax}(\theta)$, and $\hat{r}_t(a) \coloneqq \frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)}\cdot R_t(a)$ for all $a \in [K]$ is the importance sampling (IS) estimator, and we set $R_t(a) = 0$ for all $a \neq a_t$. The baseline is defined as $\hat{b}_t(a) \coloneqq \frac{\mathbb{I}\{a_t=a\}}{\pi_{\theta_t}(a)}\cdot B_t$ for all $a \in [K]$.

Algorithm 2 Gradient bandit algorithm with baselines

Input: initial parameters $\theta_1 \in \mathbb{R}^K$, learning rate $\eta > 0$.
Output: policies $\pi_{\theta_t} = \mathrm{softmax}(\theta_t)$.
while $t \geq 1$ **do**
  Sample one action $a_t \sim \pi_{\theta_t}(\cdot)$.
  Observe one reward sample $R_t(a_t) \sim P_{a_t}$.
  Choose a baseline $B_t \in \mathbb{R}$.
  for all $a \in [K]$ **do**
    if $a = a_t$ **then**
      $\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot (1 - \pi_{\theta_t}(a)) \cdot (R_t(a_t) - B_t)$.
    else
      $\theta_{t+1}(a) \leftarrow \theta_t(a) - \eta \cdot \pi_{\theta_t}(a) \cdot (R_t(a_t) - B_t)$.
    end if
  end for
end while

Proof. Using the definition of softmax Jacobian, $\hat{r}_t$ and $\hat{b}_t$, we have, for all $a \in [K]$,

$$\theta_{t+1}(a) \leftarrow \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \Big(\hat{r}_t(a) - \hat{b}_t(a) - \pi_{\theta_t}^\top \big(\hat{r}_t - \hat{b}_t\big)\Big) \tag{298}$$

$$= \theta_t(a) + \eta \cdot \pi_{\theta_t}(a) \cdot \Big(\hat{r}_t(a) - \hat{b}_t(a) - (R_t(a_t) - B_t)\Big) \tag{299}$$

$$= \theta_t(a) + \begin{cases} \eta \cdot (1 - \pi_{\theta_t}(a)) \cdot (R_t(a_t) - B_t)\,, & \text{if } a_t = a\,, \\ -\eta \cdot \pi_{\theta_t}(a) \cdot (R_t(a_t) - B_t)\,, & \text{otherwise}\,. \end{cases}$$ □

Lemma B.2 (Unbiased stochastic gradient with bounded variance / scale)**.** *Using Algorithm 2, we have, for all $t \geq 1$,*

$$\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \big(\hat{r}_t - \hat{b}_t\big)}{d\theta_t}\right] = \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \tag{300}$$

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \big(\hat{r}_t - \hat{b}_t\big)}{d\theta_t}\right\|_2^2\right] \leq 2\,\bar{R}_{\max}^2, \tag{301}$$

where $\mathbb{E}_t[\cdot]$ is on randomness from the on-policy sampling $a_t \sim \pi_{\theta_t}(\cdot)$ and reward sampling $R_t(a_t) \sim P_{a_t}$, and $\bar{R}_{\max}$ is the range of reward minus baselines, i.e.,

$$R_t(a_t) - B_t \in [-\bar{R}_{\max}, \bar{R}_{\max}]. \tag{302}$$

Proof. **First part, Eq. (300).** For all action $a \in [K]$, the true softmax PG is,

$$\frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)} = \pi_{\theta_t}(a) \cdot \big(r(a) - \pi_{\theta_t}^\top r\big)\,. \tag{303}$$

For all $a \in [K]$, the stochastic softmax PG is,

$$\frac{d\pi_{\theta_t}^\top \big(\hat{r}_t - \hat{b}_t\big)}{d\theta_t(a)} = \pi_{\theta_t}(a) \cdot \Big(\hat{r}_t(a) - \hat{b}_t(a) - \pi_{\theta_t}^\top \big(\hat{r}_t - \hat{b}_t\big)\Big) \tag{304}$$

$$= \pi_{\theta_t}(a) \cdot \Big(\hat{r}_t(a) - \hat{b}_t(a) - (R_t(a_t) - B_t)\Big) \tag{305}$$

$$= \left(\mathbb{I}\left\{a_t = a\right\} - \pi_{\theta_t}(a)\right) \cdot (R_t(a_t) - B_t)\,. \tag{306}$$

For the sampled action a_t, we have,

$$\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a_t)}\right] = \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\Big[\,(1-\pi_{\theta_t}(a_t))\cdot(R_t(a_t)-B_t)\,\Big] \tag{307}$$

$$= (1-\pi_{\theta_t}(a_t))\cdot\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\Big[R_t(a_t)-B_t\Big] \tag{308}$$

$$= (1-\pi_{\theta_t}(a_t))\cdot(r(a_t)-B_t)\,. \tag{309}$$

For any other not sampled action $a\neq a_t$, we have,

$$\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a)}\right] = \mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\Big[\,-\pi_{\theta_t}(a)\cdot(R_t(a_t)-B_t)\,\Big] \tag{310}$$

$$= -\pi_{\theta_t}(a)\cdot\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\Big[R_t(a_t)-B_t\Big] \tag{311}$$

$$= -\pi_{\theta_t}(a)\cdot(r(a_t)-B_t)\,. \tag{312}$$

Combing Eqs. (307) and (310), we have, for all $a\in[K]$,

$$\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a)}\right] = (\mathbb{I}\{a_t=a\}-\pi_{\theta_t}(a))\cdot(r(a_t)-B_t)\,. \tag{313}$$

Taking expectation over $a_t\sim\pi_{\theta_t}(\cdot)$, we have,

$$\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a)}\right] = \Pr(a_t=a)\cdot\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a)}\;\Big|\;a_t=a\right] \tag{314}$$

$$+\Pr(a_t\neq a)\cdot\mathop{\mathbb{E}}_{R_t(a_t)\sim P_{a_t}}\left[\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a)}\;\Big|\;a_t\neq a\right] \tag{315}$$

$$= \pi_{\theta_t}(a)\cdot(1-\pi_{\theta_t}(a))\cdot(r(a)-B_t)+\sum_{a'\neq a}\pi_{\theta_t}(a')\cdot(-\pi_{\theta_t}(a))\cdot(r(a')-B_t) \tag{316}$$

$$= \pi_{\theta_t}(a)\cdot\sum_{a'\neq a}\pi_{\theta_t}(a')\cdot\Big[(r(a)-B_t)-(r(a')-B_t)\Big] \tag{317}$$

$$= \pi_{\theta_t}(a)\cdot\big(r(a)-\pi_{\theta_t}^\top r\big)\,. \tag{318}$$

Combining Eqs. (303) and (314), we have, for all $a\in[K]$,

$$\mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a)}\right] = \frac{d\pi_{\theta_t}^\top r}{d\theta_t(a)}, \tag{319}$$

which implies Eq. (300) since $a\in[K]$ is arbitrary.

Second part, Eq. (301). The squared stochastic PG norm is,

$$\left\|\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t}\right\|_2^2 = \sum_{a\in[K]}\left(\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t(a)}\right)^2 \tag{320}$$

$$= \sum_{a\in[K]}(\mathbb{I}\{a_t=a\}-\pi_{\theta_t}(a))^2\cdot(R_t(a_t)-B_t)^2 \qquad \text{(by Eq. (304))} \tag{321}$$

$$\leq \bar{R}_{\max}^2\cdot\sum_{a\in[K]}(\mathbb{I}\{a_t=a\}-\pi_{\theta_t}(a))^2 \qquad \text{(by Eq. (302))} \tag{322}$$

$$= \bar{R}_{\max}^2\cdot\left[(1-\pi_{\theta_t}(a_t))^2+\sum_{a\neq a_t}\pi_{\theta_t}(a)^2\right] \tag{323}$$

$$\leq \bar{R}_{\max}^2\cdot\left[(1-\pi_{\theta_t}(a_t))^2+\Big(\sum_{a\neq a_t}\pi_{\theta_t}(a)\Big)^2\right] \qquad (\|x\|_2\leq\|x\|_1) \tag{324}$$

$$= 2\cdot\bar{R}_{\max}^2\cdot(1-\pi_{\theta_t}(a_t))^2\,. \tag{325}$$

Therefore, we have, for all $a \in [K]$, conditioning on $a_t = a$,

$$\left[\left\|\frac{d\pi_{\theta_t}^\top\left(\hat{r}_t - \hat{b}_t\right)}{d\theta_t}\right\|_2^2 \,\middle|\, a_t = a\right] \le 2 \cdot \bar{R}_{\max}^2 \cdot \left(1 - \pi_{\theta_t}(a)\right)^2. \tag{326}$$

Taking expectation over $a_t \sim \pi_{\theta_t}(\cdot)$, we have,

$$\begin{aligned}
\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top\left(\hat{r}_t - \hat{b}_t\right)}{d\theta_t}\right\|_2^2\right] &= \sum_{a\in[K]} \Pr\left(a_t = a\right) \cdot \left[\left\|\frac{d\pi_{\theta_t}^\top\left(\hat{r}_t - \hat{b}_t\right)}{d\theta_t}\right\|_2^2 \,\middle|\, a_t = a\right] && (327)\\
&\le \sum_{a\in[K]} \pi_{\theta_t}(a) \cdot 2 \cdot \bar{R}_{\max}^2 \cdot \left(1 - \pi_{\theta_t}(a)\right)^2 && (328)\\
&\le 2 \cdot \bar{R}_{\max}^2 \cdot \sum_{a\in[K]} \pi_{\theta_t}(a) \qquad (\pi_{\theta_t}(a) \in (0,1) \text{ for all } a \in [K]) && (329)\\
&= 2\,\bar{R}_{\max}^2. && \square
\end{aligned}$$

Lemma B.3 (NS between iterates)**.** *Using Algorithm 2 with $\eta \in \left(0, 2/(9\,\bar{R}_{\max})\right)$, we have, for all $t \ge 1$,*

$$D(\theta_{t+1}, \theta_t) \coloneqq \left|(\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \right\rangle\right| \le \frac{\beta(\theta_t)}{2} \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{330}$$

where $\bar{R}_{\max}$ is from Eq. (302), and

$$\beta(\theta_t) = \frac{6}{2 - 9 \cdot \bar{R}_{\max} \cdot \eta} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2. \tag{331}$$

Proof. In the proofs for Lemma 4.2, replacing $R_{\max}$ with $\bar{R}_{\max}$, and replacing $\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}$ with $\frac{d\pi_{\theta_t}^\top (\hat{r}_t - \hat{b}_t)}{d\theta_t}$, we have the results. $\square$

Lemma 6.1 (Strong growth conditions / Self-bounding noise property)**.** Using Algorithm 2, we have, for all $t \ge 1$,

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top\left(\hat{r}_t - \hat{b}_t\right)}{d\theta_t}\right\|_2^2\right] \le \frac{8 \cdot \bar{R}_{\max}^2 \cdot R_{\max} \cdot K^{3/2}}{\Delta^2} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2, \tag{332}$$

where $\Delta \coloneqq \min_{i\neq j} |r(i) - r(j)|$, and $\bar{R}_{\max}$ is from Eq. (302).

Proof. Given $t \ge 1$, denote k_t as the action with largest probability, i.e., $k_t \coloneqq \arg\max_{a\in[K]} \pi_{\theta_t}(a)$. We have,

$$\pi_{\theta_t}(k_t) \ge \frac{1}{K}. \tag{333}$$

According to Eq. (327), we have,

$$\begin{aligned}
\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top\left(\hat{r}_t - \hat{b}_t\right)}{d\theta_t}\right\|_2^2\right] &= \sum_{a\in[K]} \Pr\left(a_t = a\right) \cdot \left[\left\|\frac{d\pi_{\theta_t}^\top\left(\hat{r}_t - \hat{b}_t\right)}{d\theta_t}\right\|_2^2 \,\middle|\, a_t = a\right] && (334)\\
&\le \sum_{a\in[K]} \pi_{\theta_t}(a) \cdot 2 \cdot \bar{R}_{\max}^2 \cdot \left(1 - \pi_{\theta_t}(a)\right)^2 && (335)\\
&= 2 \cdot \bar{R}_{\max}^2 \cdot \left[\pi_{\theta_t}(k_t) \cdot \left(1 - \pi_{\theta_t}(k_t)\right)^2 + \sum_{a\neq k_t} \pi_{\theta_t}(a) \cdot \left(1 - \pi_{\theta_t}(a)\right)^2\right] && (336)\\
&\le 2 \cdot \bar{R}_{\max}^2 \cdot \left[1 - \pi_{\theta_t}(k_t) + \sum_{a\neq k_t} \pi_{\theta_t}(a)\right] \qquad (\pi_{\theta_t}(a) \in (0,1) \text{ for all } a \in [K]) && (337)\\
&= 4 \cdot \bar{R}_{\max}^2 \cdot \left(1 - \pi_{\theta_t}(k_t)\right). && (338)
\end{aligned}$$

Therefore, we have,

$$\mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t}\right\|_2^2\right] \le 4\cdot\bar{R}_{\max}^2\cdot(1-\pi_{\theta_t}(k_t)) \tag{339}$$

$$\le \frac{4\cdot\bar{R}_{\max}^2\cdot K}{\Delta^2}\cdot\sum_{a\in[K]}\pi_{\theta_t}(a)\cdot(r(a)-\pi_{\theta_t}^\top r)^2 \qquad \text{(by Eq. (117))} \tag{340}$$

$$\le \frac{4\cdot\bar{R}_{\max}^2\cdot K}{\Delta^2}\cdot 2\cdot\sqrt{K}\cdot R_{\max}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \qquad \text{(by Eq. (110))} \tag{341}$$

$$= \frac{8\cdot\bar{R}_{\max}^2\cdot R_{\max}\cdot K^{3/2}}{\Delta^2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2. \qquad \square$$

Lemma B.4 (Constant learning rate)**.** *Using Algorithm 2 with* $\eta=\frac{\Delta^2}{40\cdot K^{3/2}\cdot\bar{R}_{\max}^2\cdot R_{\max}}$*, we have, for all* $t\ge 1$*,*

$$\pi_{\theta_t}^\top r-\mathbb{E}_t[\pi_{\theta_{t+1}}^\top r]\le -\frac{\Delta^2}{80\cdot K^{3/2}\cdot\bar{R}_{\max}^2\cdot R_{\max}}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2, \tag{342}$$

where $\bar{R}_{\max}$ *is from Eq.* (302)*.*

Proof. Using the learning rate,

$$\eta=\frac{\Delta^2}{40\cdot K^{3/2}\cdot\bar{R}_{\max}^2\cdot R_{\max}} \tag{343}$$

$$=\frac{4}{45\cdot\bar{R}_{\max}}\cdot\frac{\Delta^2}{\bar{R}_{\max}\cdot R_{\max}}\cdot\frac{1}{K^{3/2}}\cdot\frac{45}{4}\cdot\frac{1}{40} \tag{344}$$

$$\le\frac{4}{45\cdot\bar{R}_{\max}}\cdot 4\cdot\frac{1}{2\cdot\sqrt{2}}\cdot\frac{45}{4}\cdot\frac{1}{40}, \qquad \big(\Delta\le 2\cdot R_{\max},\ \Delta\le 2\cdot\bar{R}_{\max},\ \text{and } K\ge 2\big) \tag{345}$$

$$<\frac{4}{45\cdot\bar{R}_{\max}}, \tag{346}$$

we have $\eta\in\big(0,2/(9\,\bar{R}_{\max})\big)$. According to Lemma B.3, we have,

$$\left|(\pi_{\theta_{t+1}}-\pi_{\theta_t})^\top r-\left\langle\frac{d\pi_{\theta_t}^\top r}{d\theta_t},\theta_{t+1}-\theta_t\right\rangle\right|\le\frac{3}{2-9\cdot\bar{R}_{\max}\cdot\eta}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\cdot\|\theta_{t+1}-\theta_t\|_2^2 \tag{347}$$

$$\le\frac{3}{2-9\cdot\bar{R}_{\max}\cdot\frac{4}{45\cdot\bar{R}_{\max}}}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\cdot\|\theta_{t+1}-\theta_t\|_2^2 \qquad \text{(by Eq. (346))} \tag{348}$$

$$=\frac{5}{2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\cdot\|\theta_{t+1}-\theta_t\|_2^2, \tag{349}$$

which implies that,

$$\pi_{\theta_t}^\top r-\pi_{\theta_{t+1}}^\top r\le-\left\langle\frac{d\pi_{\theta_t}^\top r}{d\theta_t},\theta_{t+1}-\theta_t\right\rangle+\frac{5}{2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\cdot\|\theta_{t+1}-\theta_t\|_2^2 \tag{350}$$

$$=-\eta\cdot\left\langle\frac{d\pi_{\theta_t}^\top r}{d\theta_t},\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t}\right\rangle+\frac{5}{2}\cdot\left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2\cdot\eta^2\cdot\left\|\frac{d\pi_{\theta_t}^\top(\hat{r}_t-\hat{b}_t)}{d\theta_t}\right\|_2^2, \tag{351}$$

where the last equation uses Algorithm 2. Taking expectation over $a_t \sim \pi_{\theta_t}(\cdot)$ and $R_t(a_t) \sim P_{a_t}$, we have,

$$\pi_{\theta_t}^\top r - \mathbb{E}_t[\pi_{\theta_{t+1}}^\top r] \le -\eta \cdot \left\langle \frac{d\pi_{\theta_t}^\top r}{d\theta_t}, \mathbb{E}_t\left[\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right]\right\rangle + \frac{5}{2}\cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \eta^2 \cdot \mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top \hat{r}_t}{d\theta_t}\right\|_2^2\right] \tag{352}$$

$$= -\eta \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 + \frac{5}{2}\cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \eta^2 \cdot \mathbb{E}_t\left[\left\|\frac{d\pi_{\theta_t}^\top (\hat{r}_t - \hat{b}_t)}{d\theta_t}\right\|_2^2\right] \quad \text{(by Lemma B.2)} \tag{353}$$

$$\le -\eta \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 + \frac{5}{2}\cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \cdot \eta^2 \cdot \frac{8\cdot \bar{R}_{\max}^2 \cdot R_{\max}\cdot K^{3/2}}{\Delta^2} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2 \quad \text{(by Lemma 6.1)} \tag{354}$$

$$= \left(-\eta + \eta^2 \cdot \frac{20\cdot \bar{R}_{\max}^2 \cdot R_{\max}\cdot K^{3/2}}{\Delta^2}\right) \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2 \tag{355}$$

$$= -\frac{\Delta^2}{80\cdot K^{3/2}\cdot \bar{R}_{\max}^2 \cdot R_{\max}} \cdot \left\|\frac{d\pi_{\theta_t}^\top r}{d\theta_t}\right\|_2^2. \quad \text{(by Eq. (343))} \tag{356}$$

□

Theorem B.5. *Using Algorithm 2, we have, the sequence $\{\pi_{\theta_t}^\top r\}_{t\ge 1}$ converges with probability one.*

Proof. As in the proof for Theorem 5.1, we set

$$W_{t+1}(a) = \theta_{t+1}(a) - \mathbb{E}_t[\theta_{t+1}(a)] \tag{357}$$

$$= \theta_t(a) + \eta\cdot[I_t(a) - \pi_{\theta_t}(a)]\cdot(R_t(a_t) - B_t) - \left[\theta_t(a) + \eta\cdot\pi_{\theta_t}(a)\cdot\left(r(a) - \pi_{\theta_t}^\top r\right)\right] \tag{358}$$

$$= \eta\cdot[I_t(a) - \pi_{\theta_t}(a)]\cdot(R_t(a_t) - B_t) - \eta\cdot\pi_{\theta_t}(a)\cdot\left[r(a) - \pi_{\theta_t}^\top r\right], \tag{359}$$

which implies that,

$$Z_t(a) = W_1(a) + \cdots + W_t(a) \tag{360}$$

$$= \sum_{s=1}^t \eta\cdot[I_s(a) - \pi_{\theta_s}(a)]\cdot(R_s(a_s) - B_s) - \eta\cdot\pi_{\theta_s}(a)\cdot\left[r(a) - \pi_{\theta_s}^\top r\right]. \tag{361}$$

We also have,

$$P_t(a) = \mathbb{E}_t[\theta_{t+1}(a)] - \theta_t(a) = \eta\cdot\pi_{\theta_t}(a)\cdot\left[r(a) - \pi_{\theta_t}^\top r\right]. \tag{362}$$

In the remaining part of the proofs for Theorem 5.1, replacing $R_{\max}$ with $\bar{R}_{\max}$, we have the results. □

C. Miscellaneous Extra Supporting Results

Recall that $(X_t, \mathcal{F}_t)_{t\ge 1}$ is a *sub-martingale* (super-martingale, martingale) if $(X_t)_{t\ge 1}$ is adapted to the filtration $(\mathcal{F}_t)_{t\ge 1}$ and $\mathbb{E}[X_{t+1}|\mathcal{F}_t] \ge X_t$ ($\mathbb{E}[X_{t+1}|\mathcal{F}_t] \le X_t$, $\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t$, respectively) holds almost surely for any $t \ge 1$. For brevity, let $\mathbb{E}_t[\cdot]$ denote $\mathbb{E}[\cdot|\mathcal{F}_t]$ where the filtration should be clear from the context and we also extend this notation to $t = 0$ such that $\mathbb{E}_0 U = \mathbb{E}[U]$.

Theorem C.1 (Doob's supermartingale convergence theorem (Doob, 2012))**.** *If $(Y_t)_{t\ge 1}$ is an $\{\mathcal{F}_t\}_{t\ge 1}$-adapted sequence such that $\mathbb{E}[Y_{t+1}|\mathcal{F}_t] \le Y_t$ and $\sup_t \mathbb{E}[|Y_t|] < \infty$ then $\{Y_t\}_{t\ge 1}$ almost surely converges (a.s.) and, in particular, $Y_t \to Y$ a.s. as $t \to \infty$ where $Y = \limsup_{t\to\infty} Y_t$ is such that $\mathbb{E}[|Y|] < \infty$.*

Lemma C.2 (Extended Borel-Cantelli Lemma, Corollary 5.29 of (Breiman, 1992))**.** *Let $(\mathcal{F}_n)_{n\ge 1}$ be a filtration, $A_n \in \mathcal{F}_n$. Then, almost surely,*

$$\{\omega : \omega \in A_n \text{ infinitely often }\} = \left\{\omega : \sum_{n=1}^\infty \mathbb{P}(A_n|\mathcal{F}_n)\right\}.$$