Ordering-based Conditions for Global Convergence of Policy Gradient Methods

Jincheng Mei
Google DeepMind
jcmei@google.com

Bo Dai
Google DeepMind
bodai@google.com

Alekh Agarwal
Google Research
alekhagarwal@google.com

Mohammad Ghavamzadeh
Amazon[*]
ghavamza@amazon.com

Csaba Szepesvári
Google DeepMind
University of Alberta
szepi@google.com

Dale Schuurmans
Google DeepMind
University of Alberta
daes@ualberta.ca

Abstract

We prove that, for finite-arm bandits with linear function approximation, the global convergence of policy gradient (PG) methods depends on inter-related properties between the policy update and the representation. First, we establish a few key observations that frame the study: **(i)** Global convergence can be achieved under linear function approximation without policy or reward realizability, both for the standard Softmax PG and natural policy gradient (NPG). **(ii)** Approximation error is not a key quantity for characterizing global convergence in either algorithm. **(iii)** The conditions on the representation that imply global convergence are different between these two algorithms. Overall, these observations call into question approximation error as an appropriate quantity for characterizing the global convergence of PG methods under linear function approximation. Second, motivated by these observations, we establish new general results: **(i)** NPG with linear function approximation achieves global convergence *if and only if* the projection of the reward onto the representable space preserves the optimal action's rank, a quantity that is not strongly related to approximation error. **(ii)** The global convergence of Softmax PG occurs if the representation satisfies a non-domination condition and can preserve the ranking of rewards, which goes well beyond policy or reward realizability. We provide experimental results to support these theoretical findings.

1 Introduction

Policy gradient (PG) is a foundational concept in reinforcement learning (RL), centrally used in both policy-based and actor-critic methods [25]. Despite the non-convexity of the policy optimization objective [4], global convergence of PG methods has been recently established in the tabular case for standard configurations such as the softmax parameterization [4, 22] and stochastic on-policy sampling [20]. In practice, when an RL agent is faced with a problem with large state and/or action spaces, *function approximation* is needed to generalize across related states and actions. The behavior of PG methods in these settings is relatively under-explored. In this paper, we study this question for the case of linear function approximation, and establish a surprising result that

the classical Softmax PG method converges whenever there exists an adequate linear function that ranks actions in the same order as the ground-truth reward function.

[*] The work was done prior to joining Amazon, while the author was at Google Research.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Understanding the behavior of PG methods under function approximation is crucial for describing the behavior of RL in practice, since one rarely faces domains small enough to explicitly enumerate over states and actions in parameterizing the policy. It is well known that, standard Softmax PG converges to stationary points if a "compatible" function approximation is used [25]; i.e., one that is able to exactly represent policy value functions. However, when exact policy values are non-realizable, "approximation error" is typically considered to be the key quantity for characterizing how well a function approximation captures relevant problem quantities, including transition dynamics, rewards and policy values. This paper shows that such an approximation error perspective is *overly demanding* when attempting to characterize the conditions that lead to global convergence of PG methods.

Using the concept of approximation error, global convergence results for PG methods have been recently established in an additive form,

$$\text{sub-optimality gap} \leq \text{optimization error} + \text{approximation error}, \tag{1}$$

implying that if the approximation error is small, a diminishing optimization error implies a small sub-optimality gap. A representative result is the global convergence of natural policy gradient (NPG) [4, Table 2], where the optimization error will diminish as the algorithm updates. There have also been global convergence results for other PG variants under linear function approximation that follow a similar approximation error analysis [3, 8, 10, 28, 5, 1, 2]. However, an additive bound like Eq. (1) has the inherent weakness that the approximation error will never be zero if the function approximation is not able to perfectly represent the desired quantities. This prevents such a strategy from establishing global convergence in cases where the approximation error is non-zero but a PG method still reaches the best representable solution.

Therefore, in spite of this recent progress, using approximation error in PG global convergence with function approximations has left two major gaps in the literature. First, it has not been investigated whether small approximation error is *necessary* to achieve convergence to an optimal representable policy [4], diverting attention from feature designs that achieve useful properties beyond small approximation error. Second, it is not clear if standard Softmax PG (other than NPG) converges globally under small approximation errors. In particular, NPG contains a least squares regression step [4, Eq. (17)] that can be naturally characterized with an approximation error quantity. However, standard Softmax PG does not have such a projection step [25], and the results in [4] do not apply to this update. Whether standard Softmax PG can achieve global convergence with even linearly realizable rewards (zero approximation error) is still an open problem.

In this paper, we address the above questions and contribute the following results. First, we provide negative answers to questions on the role of approximation error in determining global convergence of PG methods:

(i) Global convergence can be achieved under linear function approximation with non-zero approximation error, for both the standard Softmax PG and natural policy gradient (NPG) updates.
(ii) Approximation error is not a key quantity for characterizing global convergence in either case.
(iii) The conditions that imply global convergence are different between these two algorithms.

Second, these results lead us to question whether approximation error is an appropriate quantity to consider the global convergence of PG methods under linear function approximation. We establish new general results that characterize the conditions for global convergence of PG methods:

(i) NPG with log-linear function approximation achieves global convergence if and only if the projection of the reward onto the representable space preserves the optimal action's rank. This result significantly extends previous results that use approximation error in the analysis [4, 3], since preserving the rank of the optimal action is not strongly related to approximation error (except in the realizable limit).
(ii) We show that the global convergence of Softmax PG follows if the representation satisfies a non-domination condition and can preserve the ranking of rewards, which goes well beyond policy or reward realizability. As a byproduct, we resolve an open question by showing that even for linearly realizable reward function, Softmax PG cannot always converge to globally optimal policies when the non-domination condition for representation is violated.

2 Settings

We study the policy optimization problem under one state with K actions. Given a reward vector $r \in \mathbb{R}^K$, the problem is to find a parametric policy π_θ to maximize the expected reward,

$$\sup_{\theta \in \mathbb{R}^d} \pi_\theta^\top r, \tag{2}$$

where $\theta \in \mathbb{R}^d$ with $d < K$ is the parameter, and $\pi_\theta = \text{softmax}(X\theta)$ is called a "log-linear policy" [4, 28] such that for all action $a \in [K] \coloneqq \{1, 2, \dots, K\}$,

$$\pi_\theta(a) = \frac{\exp\{[X\theta](a)\}}{\sum_{a' \in [K]} \exp\{[X\theta](a')\}}, \tag{3}$$

where $X \in \mathbb{R}^{K \times d}$ is the feature matrix with full column rank $d < K$. There are two major difficulties with the policy optimization problem. First, Eq. (2) is a non-concave maximization w.r.t. θ, due to the softmax transform [22, Proposition 1]. Second, the policy and reward can be unrealizable, in the sense that the parametric log-linear policy $\pi_\theta = \text{softmax}(X\theta)$ cannot well approximate every policy π in the K-dimensional probability simplex, and the score $X\theta \in \mathbb{R}^K$ cannot well approximate the true mean reward $r \in \mathbb{R}^K$. Such limitations arise in the linear function approximation case because π_θ and $X\theta$ are restricted to low-dimensional manifolds via $\theta \in \mathbb{R}^d$ for $d < K$.

To solve Eq. (2), we consider the standard Softmax PG [25] and NPG [13, 4] methods, shown in Algorithms 1 and 2. Softmax PG is an instance of gradient ascent, obtained by the chain rule,

$$\frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} = \frac{d\, X\theta_t}{d\theta_t} \left(\frac{d\, \pi_{\theta_t}}{d\, X\theta_t} \right)^\top \frac{d\, \pi_{\theta_t}^\top r}{d\pi_{\theta_t}} = X^\top (\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t} \pi_{\theta_t}^\top)\, r. \tag{4}$$

On the other hand, NPG conducts updates using least squares regression (i.e., projection),

$$\left(X^\top X\right)^{-1} X^\top r = \arg\min_{w \in \mathbb{R}^d} \| Xw - r \|_2^2. \tag{5}$$

As representative policy-based methods, in their general forms, Softmax PG and NPG lay the foundation for widely used RL methods, including REINFORCE [26], actor-critic [16, 7, 12], TRPO and PPO [23, 24]. The above Eqs. (4) and (5) are their updates applied to the one-state setting.

Algorithm 1 Softmax policy gradient (PG)

```
Input: Learning rate $\eta > 0$.
Output: Policies $\pi_{\theta_t} = \text{softmax}(X\theta_t)$.
Initialize parameter $\theta_1 \in \mathbb{R}^d$.
while $t \ge 1$ do
    $\theta_{t+1} \leftarrow \theta_t + \eta \cdot X^\top (\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t} \pi_{\theta_t}^\top) r$.
end while
```

Algorithm 2 Natural policy gradient (NPG)

```
Input: Learning rate $\eta > 0$.
Output: Policies $\pi_{\theta_t} = \text{softmax}(X\theta_t)$.
Initialize parameter $\theta_1 \in \mathbb{R}^d$.
while $t \ge 1$ do
    $\theta_{t+1} \leftarrow \theta_t + \eta \cdot (X^\top X)^{-1} X^\top r$.
end while
```

To understand the difficulty of the optimization problem in Eq. (2), it is helpful to consider previous work that has analyzed the convergence of PG methods.

In the tabular setting, where $d = K$, $X = \mathbf{Id}$, and $\pi_\theta = \text{softmax}(\theta)$ with $\theta \in \mathbb{R}^K$, both the rewards and optimal policy can be arbitrarily well approximated. In this case it is known that NPG enjoys a $O(1/t)$ global convergence rate [4, Table 1], which has been recently improved to $O(e^{-c \cdot t})$ [14, 20, 17, 27]. For the case of function approximation, such results have subsequently been extended to log-linear policies, where approximation error is used to characterize the projection step of Eq. (5) [4, 28]. In particular, NPG achieves the following sub-optimality gap for all $t \ge 1$ [4, Table 2],

$$\left(\pi^* - \pi_{\theta_t}\right)^\top r \le c_1 / \sqrt{t} + c_2 \cdot \epsilon_{\text{approx}}, \qquad (c_1 > 0,\ c_2 > 0) \tag{6}$$

where c_1 and c_2 are problem specific constants, π^* is the globally optimal policy, π_{θ_t} is produced by NPG, and ϵ_{approx} is the approximation error, i.e., the minimum error with which the policy values can be approximated using the features [4, Table 2]. The "optimization error" term $c_1/\sqrt{t}$ in Eq. (6) has since been improved to $O(e^{-c_3 \cdot t})$ with $c_3 > 0$ in [28, 5]. Note that if $\epsilon_{\text{approx}} > 0$ then Eq. (6) is insufficient for establishing $\pi_{\theta_t}^\top r \to r(a^*) \coloneqq \max_{a \in [K]} r(a)$ as $t \to \infty$ even when such global convergence is achieved.

The understanding for the standard Softmax PG is even less clear. In the tabular case, it is known that Softmax PG achieves global convergence asymptotically, i.e., $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ [4], with an $O(1/t)$ rate of convergence that exhibits undesirable problem and initialization dependent constants

[21, 18]. Directly extending this global convergence result to the case of function approximation, i.e., log-linear policies, is impossible without any additional assumptions on the features, since there can be exponentially many sub-optimal local maxima in the worst case [9]. In fact, even with linearly realizable rewards (zero approximation error), whether standard Softmax PG achieves global convergence still remains unsolved [4]. One intuitive reason why this is a difficult result to establish is that standard Softmax PG uses the gradient Eq. (4) rather than projection (regression) to perform updates, which is less directly connected to the concept of approximation error.

3 The Limitations of Approximation Error in Characterizing Convergence

It is known that there exist representations $X \in \mathbb{R}^{K \times d}$ with $d < K$ and $r \in \mathbb{R}^K$ that create exponentially many sub-optimal local maxima in Eq. (2) [9, Theorem 1], which makes it impossible to ensure global convergence of PG methods without imposing any structure on the function approximation. Before identifying specific conditions that ensure global convergence, we first explain how approximation error cannot be a useful structural measure for this purpose, by demonstrating that zero approximation error is not a necessary condition for global convergence, and illustrating problem instances with comparable approximation error that render starkly different convergence behaviors across different PG methods. Specifically, we illustrate these points with a set of concrete scenarios, each with 4 actions and 2-dimensional feature vectors describing each action. Since $d < K$, not every policy can be expressed in these representations, hence the problem instances are unrealizable.

3.1 Global Convergence is Achievable with Non-zero Approximation Error

The results of [9, Theorem 1] do not imply that sub-optimal local maxima always appear, as shown in the following.

Example 1. *$K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 2 \\ -2 & 0 & 1 & 0 \end{bmatrix}$ and $r = (9, 8, 7, 6)^\top$. The approximation error is $\epsilon_{\textit{approx}} = \min_{w \in \mathbb{R}^d} \|Xw - r\|_2 = \left\| X \left(X^\top X \right)^{-1} X^\top r - r \right\|_2 = \sqrt{202.6} \approx 14.2338$.*

Note that the approximation error is larger than any sub-optimality gap, i.e., for any policy π,

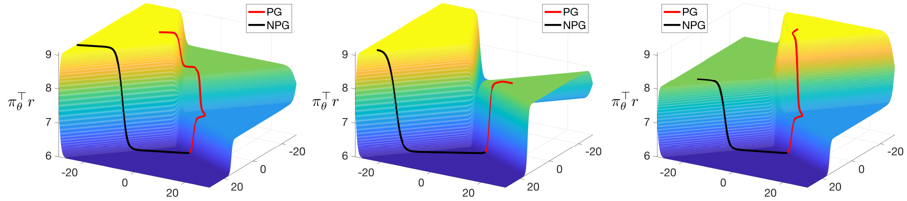
$$(\pi^* - \pi)^\top r \le 3 < \epsilon_{\text{approx}}, \tag{7}$$

hence the bound in Eq. (6) does not imply global convergence for NPG in this example. Yet, despite the non-zero approximation error and the inability of existing results including Eq. (6) to establish global convergence on Example 1, both Algorithms 1 and 2 can be shown to reach a global maximum.

Proposition 1. *Denote $a^* := \arg\max_{a \in [K]} r(a)$. With constant $\eta > 0$ and any initialization $\theta_1 \in \mathbb{R}^d$, both Algorithms 1 and 2 guarantee $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ on Example 1.*

All proofs can be found in the appendix due to space limits. The fact that Softmax PG achieves global convergence in Example 1 is much harder to establish than for NPG, since Eq. (4) involves a complex non-linearity given the presence of the softmax, unlike the linear least squares Eq. (5) used in NPG. To illustrate the intuition behind Proposition 1 we use a visualization of the optimization landscape.

Visualization. A visualization of the optimization landscape of Example 1 is shown in Figure 1(a). The bottom two-dimensional plane is the parameter space $\mathbb{R}^d$ where $d = 2$. For each $\theta \in \mathbb{R}^d$, we calculate π_θ using Eq. (3) and $\pi_\theta^\top r$ using Eq. (2), and use $\pi_\theta^\top r$ as the vertical axis value of θ.



(a) Example 1. (b) Example 2. (c) Example 3.

Figure 1: Visualizing the landscapes in the example problem instances.

To verify Proposition 1, we run Softmax PG and NPG on Example 1 with the same $\theta_1 = (6, 8)^\top \in \mathbb{R}^2$. In Figure 1(a), the optimization trajectories show 85 iterations of NPG and 8.5×10^6 iterations of

Softmax PG, both with learning rate $\eta = 0.2$. It can be clearly seen that both Softmax PG and NPG eventually achieve expected reward $\pi_{\theta_t}^\top r \to 9 = r(a^*)$, demonstrating global convergence (Figure 3(c) later shows that the sub-optimality gap $(\pi^* - \pi_{\theta_t})^\top r$ approaches 0).

In summary, Example 1 shows that both Softmax PG and NPG are able to achieve global convergence on unrealizable problem instances with non-zero approximation error. This raises the question:

Is non-zero approximation error useful for characterizing global convergence?

3.2 Global Convergence is Irrelevant to Non-zero Approximation Error

We answer the above question negatively. By comparing alternative problem instances with similar approximation errors but different convergence behaviors, we illustrate how approximation error is not able to distinguish between scenarios where global versus local convergence is obtained.

Example 2. *$K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & 0 & -1 & 2 \\ -2 & 1 & 0 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, and $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$. The approximation error is $\left\| X \left(X^\top X \right)^{-1} X^\top r - r \right\|_2 = \sqrt{205} \approx 14.3178$.*

The only difference between Examples 1 and 2 is that the second and third columns of $X^\top$ have been exchanged. The approximation error remains similar to that of Example 1. Using the upper bound of Eq. (6), one might therefore expect similar sub-optimality gaps $(\pi^* - \pi_{\theta_t})^\top r$ to be demonstrated by the algorithms, since the r.h.s. contains similar approximation errors. However, as shown in Figure 1(b), using the same initialization and learning rate, Softmax PG obtains $\pi_{\theta_t}^\top r \to 8 = r(2) < r(a^*)$ as it converges to a sub-optimal deterministic policy, while NPG continues to succeed.

Lest one believe that NPG is globally convergent, the following example, where the first and second columns of $X^\top$ are swapped, illustrate an analogous failure for NPG but not Softmax PG.

Example 3. *$K = 4$, $d = 2$, $X^\top = \begin{bmatrix} -1 & 0 & 0 & 2 \\ 0 & -2 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, and $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$. The approximation error is $\left\| X \left(X^\top X \right)^{-1} X^\top r - r \right\|_2 = \sqrt{212} \approx 14.5602$.*

Here again the approximation error is close to that of Example 1. Yet, Figure 1(c) shows that NPG achieves $\pi_{\theta_t}^\top r \to 8 < r(a^*)$ as it converges to a sub-optimal solution, while Softmax PG succeeds.

In summary, the Examples 1, 2 and 3 all have similar approximation errors, yet Softmax PG achieves global convergence on Example 1 but reaches a bad local maxima on Example 2, while NPG succeeds on Example 1 and fails on Example 3. Note that these examples can be re-scaled to have exactly the same approximation errors while demonstrating the same convergence behavior of the algorithms. From these findings we conclude that, if there is any quantity that can predict whether global versus local convergence is obtained by Softmax PG or NPG, that the quantity cannot be approximation error alone. This motivates to investigate the question: what is the right quantity to characterize global convergence for unrealizable problems?

3.3 Global Convergence Characterization is Algorithm Dependent

We make one more key point. From Figure 1(b) and Figure 1(c), NPG achieves global convergence on Example 2 but fails on Example 3, while, conversely, Softmax PG succeeds on Example 3 and fails on Example 2. This difference indicates that whatever condition characterizes global convergence, it must be *algorithm dependent*, even for the closely related algorithms Softmax PG and NPG. Therefore, one has to study the conditions for Softmax PG and NPG **respectively** (rather than one condition for both algorithms), which motivates the refined question:

What conditions characterize global convergence of Softmax PG and NPG in unrealizable problems?

4 New Characterizations of Global Convergence for PG Methods

From these observations, it is clear that whatever quantity characterizes the global convergence of PG methods, it cannot be based solely on approximation error and it must be algorithm dependent. Therefore, we study distinct global convergence conditions for Softmax PG and NPG respectively.

4.1 Reward Order Preservation with Adequate Features is Sufficient for PG Convergence

We now investigate a global convergence condition for Softmax PG under log-linear policies.

Intuition. Consider Example 1, where Softmax PG achieves global convergence. From the landscape shown in Figure 1(a), there appears to be a monotonic path from any initialization point that allows gradient ascent to reach the optimal plateau with reward $r(a^*) = 9$. Intuitively, this arises because the actions' rewards seem to be nicely "ordered". For example, starting from $\theta_1 = (6, 8)^\top \in \mathbb{R}^d$ such that $\pi_{\theta_1}^\top r \approx 6$, Softmax PG is able to improve its expected reward eventually to $\pi_{\theta_t}^\top r \approx 7$, since there exists a sub-optimal plateau with a higher reward 7 right beside the lowest plateau with reward 6. Next, Softmax PG continues to improve its expected reward eventually to $\pi_{\theta_t}^\top r \approx 8$ by "climbing" toward another neighboring plateau with a higher reward. Finally, this process ends with Softmax PG successfully arriving at the optimal plateau with reward $r(a^*) = 9$.

By contrast, in Example 2, as shown in Figure 1(b), Softmax PG gets stuck on a bad plateau with a local maximum reward of 8. Visually, Softmax PG stops improving its expected reward on this sub-optimal plateau, because it is "surrounded" by two lower plateaus with rewards 6 and 7, which breaks the nice "ordering" of the expected reward landscape and traps the gradient ascent trajectory on a sub-optimal plateau from which there is no monotonic ascent to global optimality.

Verifying reward order preservation. Based on the above intuition and observations, we conjecture that the ordering structure between the different rewards is a key property behind the global convergence of Softmax PG. We can verify this conjecture in each of the Examples 1 to 3 by determining whether the feature matrix $X \in \mathbb{R}^{K \times d}$ allows the same action ordering as the reward vector $r \in \mathbb{R}^K$ to be realized. For Example 1, note that with $w = (-1, -1)^\top \in \mathbb{R}^d$, we have

$$r' := Xw = (2, 1, -1, -2)^\top \in \mathbb{R}^K, \tag{8}$$

which preserves the ordering of $r \in \mathbb{R}^K$, such that for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$. Similarly, for Example 3, if we let $w = (-3, -1)^\top$ then we have $r' := Xw = (3, 2, -1, -6)^\top$, which also preserves the order of r over actions. Softmax PG converges to a globally optimal reward in both of these examples.

By contrast, for Example 2, it is impossible to find any $w \in \mathbb{R}^d$ such that Xw preserves the order of the rewards r. To see why, consider any $w = (w(1), w(2))^\top$ and note that

$$r' := Xw = (-2 \cdot w(2), w(2), -w(1), 2 \cdot w(1))^\top. \tag{9}$$

To preserve the reward order, we require both $-2 \cdot w(2) > w(2)$ (which would imply $w(2) < 0$) and $-w(1) > 2 \cdot w(1)$ (which would imply $w(1) < 0$), but these two conditions imply $w(2) < 0 < -w(1)$, which must reverse the order of the second and third actions. This is an example where PG can fail to reach a global optimum.

Main Softmax PG result. We formalize the above intuition by proving the following main result, which establishes that reward order preservation with adequate representations is a sufficient condition for the global convergence of Softmax PG under log-linear function approximation.

Theorem 1 (Reward order preservation, non-domination features)**.** *Given any reward $r \in \mathbb{R}^K$ and feature matrix $X \in \mathbb{R}^{K \times d}$. Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of X. If **(i)** $x_i^\top x_i > x_i^\top x_j$ for all $j \neq i$, and **(ii)** there exists at least one $w \in \mathbb{R}^d$, s.t., $r' := Xw$ preserves the order of r, i.e., for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$, then for any initialization $\theta_1 \in \mathbb{R}^d$, Algorithm 1 with a constant learning rate $\eta > 0$ achieves global convergence of $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$.*

A few remarks about this theorem are in order.

Examples 1 to 3 all satisfy the non-domination condition **(i)** on X, and their differences lie in satisfying reward order preservation or not. However, the following example shows that if the condition **(i)** on X is removed, then global convergence is not always achievable for even linearly realizable rewards (with zero approximation error).

Proposition 2. *Let $K = 3$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -10 & 0 \\ -2 & 4 & 1 \end{bmatrix} \in \mathbb{R}^{d \times K}$, and $r = Xw = (4, 2, -2)^\top$, where $w = (-1, -2)^\top \in \mathbb{R}^d$. With initialization $\theta_1 = (-\ln 2, \ln 2)^\top$, Algorithm 1 does not achieve global convergence, i.e., $\pi_{\theta_t}(1) \not\to 1$ as $t \to \infty$.*

Generalization of tabular and linear realizability. When $d = K$ and $X = \mathbf{Id}$, i.e., the softmax tabular parameterization $\pi_\theta = \mathrm{softmax}(\theta)$, it is always true that $Xr = r$ preserves the order of r. Consequently, Theorem 1 recovers the global convergence result for PG in the softmax tabular setting [4, 22] as a special case. More generally, for non-domination features, when the reward is linearly realizable, such that $Xw = r$ for some $w \in \mathbb{R}^d$, the global convergence of Softmax PG also follows from Theorem 1, since r preserves its own order when the approximation error is zero.

Corollary 1 (Linearly realizable rewards, non-domination features)**.** *Given any reward $r \in \mathbb{R}^K$ and feature matrix $X \in \mathbb{R}^{K\times d}$. Denote $x_i \in \mathbb{R}^d$ as i-th row vector of X. If **(i)** $x_i^\top x_i > x_i^\top x_j$ for all $j \neq i$, and **(ii)** there exists $w \in \mathbb{R}^d$, s.t., $Xw = r$, then for any initialization $\theta_1 \in \mathbb{R}^d$, Algorithm 1 with a constant learning rate $\eta > 0$ achieves global convergence of $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$.*

It is worth mentioning that Proposition 2 and Corollary 1 together answer a question which still remain unsolved in PG literature [4]: with linearly realizable rewards (zero approximation error), whether standard Softmax PG achieves global convergence? Proposition 2 shows that linearly realizable reward on its own is not enough to guarantee global convergence, while Corollary 1 shows that with adequate features, linearly realizable reward implies global convergence. Note that the NPG global convergence result in [4], such as Eq. (6), does not apply to standard Softmax PG.

Ordering does not determine approximation. As already illustrated in Section 3, approximation error is not adequate for capturing the global convergence of Softmax PG. It is important to emphasize that the existence of an order preserving reward r' is very different from having a small approximation error. When the approximation error is zero, then an order preserving reward (equal to r) always exists. However, in general, r' can take very different values than r, and hence have a very large approximation error, yet still enable global convergence as shown in Examples 1 and 3.

Proof idea. The idea behind the proof of the main theorem consists of three parts. We provide a sketch of the proof here; the full proof is given in Appendix A. **First**, starting from any initialization $\theta_t \in \mathbb{R}^d$, Algorithm 1 guarantees that π_{θ_t} will approach a (generalized) one-hot policy as $t \to \infty$. To see why, first note that $\pi_\theta^\top r$ is β-smooth over $\theta \in \mathbb{R}^d$ with some $\beta > 0$ (Lemma 3 in Appendix B), since the softmax transform is smooth [4, 22] and the feature matrix X has bounded values. This implies that using a sufficiently small constant learning rate $0 < \eta \le 2/\beta$ we obtain,

$$\pi_{\theta_{t+1}}^\top r - \pi_{\theta_t}^\top r \ge \frac{1}{2\,\beta} \cdot \left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \ge 0. \tag{10}$$

Note that $\pi_\theta^\top r$ is upper bounded by $r(a^*)$. According to the monotone convergence, $\pi_{\theta_t}^\top r \to c \le r(a^*)$ as $t \to \infty$. This fact combined with Eq. (10) implies $\left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \to 0$ as $t \to \infty$. Next, a special co-variance structure of softmax PG (Lemma 4) shows that $\left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \to 0$ implies that $\|\theta_t\|_2 \to \infty$ and π_{θ_t} approaches a (generalized) one-hot policy as $t \to \infty$.

Lemma 1. *Under the same conditions as Theorem 1, and $r(i) \neq r(j)$ for all $i \neq j$ (unique action reward), Algorithm 1 assures $\|\theta_t\|_2 \to \infty$ and $\pi_{\theta_t}(i) \to 1$ for an action $i \in [K]$ as $t \to \infty$.*

Remark 1. *Removing the unique action reward condition in Lemma 1 makes π_{θ_t} approach a generalized one-hot policy (rather than a strict one-hot in Lemma 1) as $t \to \infty$ as a result.*

According to Lemma 1, θ_t grows unboundedly. Intuitively, this can be seen in Figure 1(a), where there are no stationary points in a finite region.

Second, for any vector r' that preserves the order of r, we establish the following key lemma.

Lemma 2 (Non-negative covariance of order preservation)**.** *If $r' \in \mathbb{R}^K$ preserves the order of $r \in \mathbb{R}^K$, i.e., for all $i, j \in [K]$, $r(i) > r(j)$ iff $r'(i) > r'(j)$, then for any policy $\pi \in \Delta(K)$,*

$${r'}^\top \left(diag(\pi) - \pi\pi^\top\right) r = \mathrm{Cov}_\pi\left(r', r\right) \ge 0. \tag{11}$$

Now consider the direction $w \in \mathbb{R}^d$ such that $r' := Xw$ preserves the order of r. We have,

$$\begin{aligned} w^\top \theta_{t+1} &= w^\top \theta_t + \eta \cdot w^\top X^\top \left(\mathrm{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right) r \qquad \text{(by Algorithm 1)} && (12)\\ &= w^\top \theta_t + \eta \cdot {r'}^\top \left(\mathrm{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right) r \qquad (r' := Xw) && (13)\\ &\ge w^\top \theta_t. \qquad \text{(by Lemma 2)} && (14) \end{aligned}$$

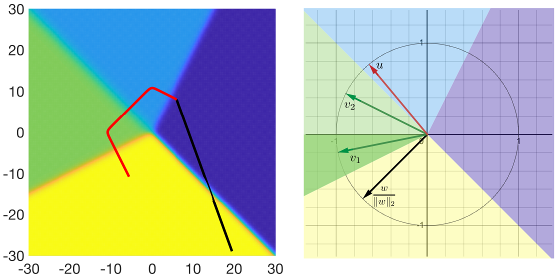
Third, take a sub-optimal action $i \in [K]$ with $r(i) < r(a^*)$, and we show that the assumption $\pi_{\theta_t}(i) \to 1$ as $t \to \infty$ leads to a contradiction.

To that end, first observe that this assumption implies that for all large enough time $t \geq 1$,

$$\left[\frac{X\theta_t}{\|\theta_t\|_2}\right](i) = \max_{a \in [K]} \left[\frac{X\theta_t}{\|\theta_t\|_2}\right](a), \tag{15}$$

which means that the sub-optimal action $i \in [K]$ always has the largest score (since its probability $\pi_{\theta_t}(i) \to 1$ is always the largest). Moreover, differences between actions' scores are unbounded, due to $\frac{\pi_{\theta_t}(i)}{\pi_{\theta_t}(j)} = \exp\left\{[X\theta_t](i) - [X\theta_t](j)\right\} \to \infty$ for all other actions $j \neq i$.

Consider Example 1 for illustration. The top view of Figure 1(a) is shown in Figure 2(a). Take $i = 2$ and $r(i) = 8$, and assume $\frac{\theta_t}{\|\theta_t\|_2}$ stays in the green (sub-optimal) region of Figure 2(a), excluding its boundaries. This green region is partitioned in Figure 2(b), where the dark sub-region contains $v_2 \in \mathbb{R}^d$ such that $[Xv_2](a^*)$ is the second largest component among all $a \neq 2$, and the light sub-region is the remaining.



(a) Top view of Example 1.

(b) Illustration for the third part.

Figure 2: Idea illustration.

The argument is completed by addressing the two cases: **(i)** If $\frac{\theta_t}{\|\theta_t\|_2}$ stays in the dark sub-region where $v_1 \in \mathbb{R}^d$ belongs to, then $\pi_{\theta_t}^\top r > r(i) = 8$ must occur in finite $t < \infty$, implying $\pi_{\theta_t}(i) \not\to 1$, contradicting the assumption. Intuitively, the contradiction occurs because the dark sub-region is closer to a higher plateau with reward 9, and scaling up θ_t's magnitude in this sub-region eventually ensures $\pi_{\theta_t}^\top r > r(i) = 8$. **(ii)** If $\frac{\theta_t}{\|\theta_t\|_2}$ stays in the light sub-region which contains $v_2 \in \mathbb{R}^d$, then $w^\top \theta_t > u^\top \theta_t$ must occur in finite time $t < \infty$, implying that $\frac{\theta_t}{\|\theta_t\|_2}$ will enter the dark sub-region, reducing to the first case. This argument depends on Eq. (12) and a key observation showing that $u^\top \theta_{t+1} < u^\top \theta_t$, where u is a "worse" direction such that $[Xu](a^-) = \max_{a \in [K]} [Xu](a)$ for some $a^- \in [K]$ with $r(a^-) < r(i)$.

To summarize, $\frac{\theta_t}{\|\theta_t\|_2}$ cannot always stay in the green sub-optimal region in Figure 2(a), which implies that $\frac{\theta_t}{\|\theta_t\|_2}$ must eventually enter the optimal region that contains w and stay in that region. By Lemma 1 we then obtain $\pi_{\theta_t}(a^*) \to 1$ and $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ (see appendix).

4.2 Optimal Action Preservation is Necessary and Sufficient for NPG Convergence

Next, we investigate the global convergence conditions for NPG under log-linear policies. Unlike Softmax PG, the key property for determining global convergence of NPG is whether the projection of the rewards r onto the feature representation X preserves the top ranking of the optimal action.

Intuition and demonstration. First consider Example 1 where NPG successfully converges to a global maximum. From Algorithm 2, a simple calculation shows,

$$X\theta_{t+1} = X\theta_t + \eta \cdot X(X^\top X)^{-1} X^\top r = X\theta_t + \eta \cdot \frac{1}{5} \cdot (22, -4, -11, 8)^\top, \tag{16}$$

which implies that the optimal action $a^* = 1$ always receives the largest update to its score $[X\theta_t](a^*)$ in each iteration. Next, take a sub-optimal action $a = 2$, as an example, and observe that,

$$\frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)} = \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} \cdot \exp\left\{\eta \cdot (\hat{r}(a^*) - \hat{r}(a))\right\} = \frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(a)} \cdot \exp\left\{\eta \cdot \frac{26}{5}\right\} \tag{17}$$

by Eq. (3), where $\hat{r} := X(X^\top X)^{-1} X^\top r$. Using a constant learning rate $\eta > 0$ and applying Eq. (17), we have that $\pi_{\theta_t}(a^*)$ grows exponentially with t, indicating that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$ since the same argument works for any sup-optimal action $a \neq a^*$. Moreover, the rate is $O(e^{-c \cdot t})$, since $(\pi^* - \pi_{\theta_t})^\top r \leq 2 \cdot \|r\|_\infty \cdot (1 - \pi_{\theta_t}(a^*))$. The $O(e^{-c \cdot t})$ rate matches the results in softmax tabular settings [14, 20, 17, 27].

Second, consider Example 2 where NPG fails to converge to a global maximum. Using similar calculations to Eq. (16) we obtain,

$$X\theta_{t+1} = X\theta_t + \eta \cdot \hat{r} = X\theta_t + \eta \cdot \frac{1}{5} \cdot (-3, 18, -9, -6)^\top \tag{18}$$

which implies that a sub-optimal action $a = 2$ always receives the largest update on its score $[X\theta_t](2)$ in each iteration. The failure in Figure 1(b) is then verified by similar arguments around Eq. (17).

Main NPG result. Based on these observations, it is evident that for NPG to converge globally, it is important for the optimal action to eventually always receive the largest update to its score, which makes it critical that the least square projection $X(X^\top X)^{-1}X^\top r$ preserves the top ranking of the optimal action. We formalize this intuition by establishing the following main result.

Theorem 2 (Optimal action preservation condition)**.** *For a constant learning rate $\eta > 0$, a necessary and sufficient condition for Algorithm 2 to achieve global convergence $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ from any initialization $\theta_1 \in \mathbb{R}^d$ is that $\hat{r}(a^*) > \hat{r}(a)$ for all $a \neq a^*$, such that $a^* := \arg\max_{a \in [K]} r(a)$, and $\hat{r} := X \left(X^\top X\right)^{-1} X^\top r$ is the least squares projection of r onto the column space of X. If the condition is satisfied, then the rate of convergence is $\left(\pi^* - \pi_{\theta_t}\right)^\top r \in O(e^{-c\cdot t})$ for some $c > 0$.*

Proof idea. When the optimal action preservation is satisfied, similar arguments to Eqs. (16) and (17) guarantee that $\pi_{\theta_t}(a^*)$ grows exponentially with t, indicating that $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$.

The constant $c > 0$ in Theorem 2 depends on the gap of $\hat{r}$, i.e., $\hat{r}(a^*) - \max_{a \neq a^*} \hat{r}(a)$, which finds similarities to NPG results in tabular settings [14, 15]. The main difference is that the gap of true reward r in tabular cases is replaced with the gap of least square projection $\hat{r}$ in function approximation settings in Theorem 2. This similarity is an evidence for improving the rate to super-linear by using geometrically increasing step sizes, as in tabular settings [17, 27, 19, 28, 6].

One-sided Approximation Error. For NPG, [4, Lemma 6.2] introduces a "one-sided approximation error" quantity, which aims to overestimate the advantage of the optimal action a^*,

$$\epsilon_t := r(a^*) - \pi_{\theta_t}^\top r - w^\top \left(x_{a^*} - X^\top \pi_{\theta_t}\right) = r(a^*) - \pi_{\theta_t}^\top r - \left(\hat{r}(a^*) - \pi_{\theta_t}^\top \hat{r}\right). \tag{19}$$

This quantity relaxes the notion of approximation error and still guarantees the global convergence of NPG, since if $\sum_{t=1} \epsilon_t \in o(T)$, then NPG with $\eta \in O(1/\sqrt{T})$ achieves global convergence [4, Lemma 6.2]. We note however that Eq. (19) has two limitations: **(i)** Eq. (19) depends on the entire update trajectory $\{\theta_t\}_{t \geq 1}$, which is hard to verify. By contrast, the optimal action preservation condition in Theorem 2 only involves problem quantities X and r. **(ii)** It is not clear whether Eq. (19) is a necessary condition for global convergence, while optimal action preservation is proved above to be both necessary and sufficient.

5 Simulation Study

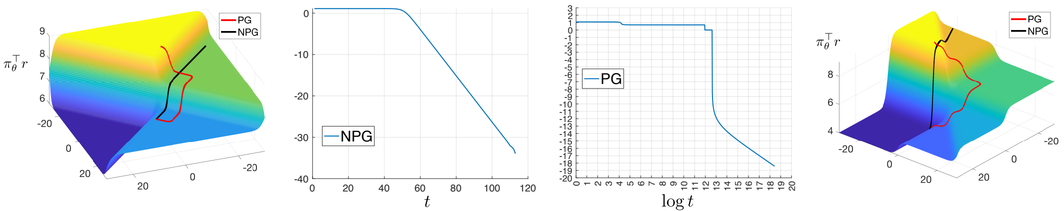
We conducted additional simulations to check the theoretical results. **First**, we check whether the strict inequality of $\hat{r}(a^*) > \hat{r}(a)$ for all $a \neq a^*$ in Theorem 2 is required for NPG global convergence.

Example 4. *$K = 4$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix} \in \mathbb{R}^{d \times K}$, and $r = (9, 8, 7, 6)^\top \in \mathbb{R}^K$. The best fit for r is $\hat{r} = X \left(X^\top X\right)^{-1} X^\top r = (1, 1, -1, -1)^\top$.*

Example 4 has $\hat{r}(a^*) = \hat{r}(1) = \hat{r}(2)$, which violates the strict inequality condition of $\hat{r}(a^*) > \hat{r}(a)$ for all $a \neq a^*$. The consequence is that NPG guarantees $\frac{\pi_{\theta_t}(a^*)}{\pi_{\theta_t}(2)} = \frac{\pi_{\theta_1}(a^*)}{\pi_{\theta_1}(2)}$ for all $t \geq 1$, which makes it impossible for $\pi_{\theta_t}(a^*) \to 1$ as $t \to \infty$. This is observed in Figure 3(a), supporting that the strictly inequality condition in Theorem 2 is indeed necessary. The initialization is $\theta_1 = (4, 10)^\top$, and $\eta = 0.2$. We run 150 iterations for NPG and 1.5×10^7 iterations for Softmax PG.

Second, we run 150 iterations of NPG on Example 1. As shown in Figure 3(b), the quantity $\log \left(\pi^* - \pi_{\theta_t}\right)^\top r$ is a linear function of time t, implying that $\left(\pi^* - \pi_{\theta_t}\right)^\top r \in O(e^{-c\cdot t})$ with $c > 0$. This supports the convergence rate results in Theorem 2. Here θ_1 and η are the same as in Figure 1(a).

Third, we check whether the condition in Theorem 1 is required for Softmax PG global convergence.



(a) Example 4. (b) $\log{(\pi^* - \pi_{\theta_t})}^\top r$ (c) $\log{(\pi^* - \pi_{\theta_t})}^\top r$ (d) Example 5.

Figure 3: Simulations for verifying theoretical results.

Example 5. *$K = 6$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -1 & -1 & 0 & 1 & 1 \\ -1 & 0 & 1 & 1 & 0 & -1 \end{bmatrix}$, and $r = (9, 8, 7, 6, 5, 4)^\top$.*

Similar to Eq. (9), it is impossible to find any $w \in \mathbb{R}^d$, such that $r' := Xw$ preserves the order of r in Example 5. However, as shown in Figure 3(d), Softmax PG achieves $\pi_{\theta_t}^\top \to r(a^*) = 9$, indicating that the reward order preservation condition in Theorem 1 is sufficient but not necessary for PG to achieve global convergence. The initialization is $\theta_1 = (10, -2)^\top$, and $\eta = 0.2$. We run 100 iterations for NPG and 2×10^6 iterations for Softmax PG. Note that NPG behaves erratically on Example 5 (which does not satisfy its global convergence conditions), by first entering then leaving the optimal plateau, eventually approaching a sub-optimal solution.

Finally, we run 10^8 iterations of Softmax PG on Example 1, using the same η and θ_1 as in Figure 1(a). Figure 3(c) shows that the slope of $\log{(\pi^* - \pi_{\theta_t})}^\top r$ over $\log t$ approaches -1, indicating that the global convergence rate is $(\pi^* - \pi_{\theta_t})^\top r \in O(1/t)$, matching the softmax tabular setting results [22].

6 Discussions

Checking ordering-based conditions. Checking the existence of $w \in \mathbb{R}^d$ in Theorem 1 is known as linear feasibility in literature [11], i.e., determining whether a set of inequalities has a non-empty intersection. In particular, suppose $X \in \mathbb{R}^{K \times d}$, and $r \in \mathbb{R}^K$ is sorted, i.e., $r(1) \ge r(2) \ge \cdots \ge r(K)$. Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of X. The linear feasibility problem in this case is to check if there exists $w \in \mathbb{R}^d$, such that for all $i \in [K-1]$, $x_i^\top w \ge x_{i+1}^\top w$. Linear feasibility can be cast as linear programming (LP) using a dummy objective and keeping the constraints, hence any LP technique, such as the ellipsoid method, can be used to solve it [11]. On the other hand, checking the optimal action preservation in Theorem 2 requires the same information as in calculating approximation error $\|\hat{r} - r\|_2 = \min_{w \in \mathbb{R}^d} \|Xw - r\|_2$, since $\arg\max_{a \in [K]} \hat{r}(a) = \arg\max_{a \in [K]} r(a)$ can be immediately verified after calculating the projection $\hat{r} := X^\top (X^\top X)^{-1} X^\top r$.

Generalization to Markov decision processes (MDPs). Our work provides some new and useful insights for understanding more complex settings, but it requires further investigation to resolve this highly non-trivial problem for general MDPs. See Appendix C for detailed discussions.

7 Conclusions and Future Work

We believe this work opens new directions for understanding PG-based methods under function approximation, going well beyond the conventional approximation error based analysis. The major technical findings involve ordering-based conditions and relevant techniques (covariance and global convergence). Identifying exact necessary and sufficient conditions for the global convergence of Softmax PG remains future work. Extending the results and techniques to general MDPs is another important and challenging next step. Combining function approximation with recent results on stochastic on-policy sampling [20] is another interesting direction for agnostic learning. Investigating whether these new global convergence conditions might be used to achieve better representation learning is of great interest for algorithm design. Generalizing the proof techniques to other scenarios where non-linear transforms (activation functions) interact with low-dimensional features through gradient descent, such as in neural networks, is another lofty ambition.

Acknowledgments and Disclosure of Funding

The authors would like to thank anonymous reviewers for their valuable comments. CS, DS gratefully acknowledge funding from the Canada CIFAR AI Chairs Program, Amii and NSERC.

References

[1] Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

[2] Yasin Abbasi-Yadkori, Nevena Lazic, Csaba Szepesvari, and Gellert Weisz. Exploration-enhanced politex. *arXiv preprint arXiv:1908.10479*, 2019.

[3] Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. PC-PG: Policy cover directed exploration for provable policy gradient learning. *Advances in neural information processing systems*, 33:13399–13412, 2020.

[4] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

[5] Carlo Alfano and Patrick Rebeschini. Linear convergence for natural policy gradient with log-linear policy parametrization. *arXiv preprint arXiv:2209.15382*, 2022.

[6] Carlo Alfano, Rui Yuan, and Patrick Rebeschini. A novel framework for policy mirror descent with general parametrization and linear convergence. *arXiv preprint arXiv:2301.13139*, 2023.

[7] Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

[8] Semih Cayci, Niao He, and Rayadurgam Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

[9] Minmin Chen, Ramki Gummadi, Chris Harris, and Dale Schuurmans. Surrogate objectives for batch policy optimization in one-step decision making. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Zaiwei Chen, Sajad Khodadadian, and Siva Theja Maguluri. Finite-sample analysis of off-policy natural actor–critic with linear function approximation. *IEEE Control Systems Letters*, 6:2611–2616, 2022.

[11] Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

[12] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[13] Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002.

[14] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On the linear convergence of natural policy gradient algorithm. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 3794–3799. IEEE, 2021.

[15] Sajad Khodadadian, Prakirt Raj Jhunjhunwala, Sushil Mahavir Varma, and Siva Theja Maguluri. On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214, 2022.

[16] Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

[17] Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

[18] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge. In *Conference on Learning Theory*, pages 3107–3110. PMLR, 2021.

[19] Yan Li, Guanghui Lan, and Tuo Zhao. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity. *arXiv preprint arXiv:2201.09457*, 2022.

[20] Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvari, and Dale Schuurmans. The role of baselines in policy gradient optimization. *Advances in Neural Information Processing Systems*, 35:17818–17830, 2022.

[21] Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. *Advances in Neural Information Processing Systems*, 33:21130–21140, 2020.

[22] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

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

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

[25] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

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

[27] Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

[28] Rui Yuan, Simon S Du, Robert M Gower, Alessandro Lazaric, and Lin Xiao. Linear convergence of natural policy gradient methods with log-linear policies. *arXiv preprint arXiv:2210.01400*, 2022.

A Proofs for Main Results

Proposition 1. Denote $a^* := \arg\max_{a \in [K]} r(a)$. With constant $\eta > 0$ and any initialization $\theta_1 \in \mathbb{R}^d$, both Algorithms 1 and 2 guarantee $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ on Example 1.

Proof. **First part.** Algorithm 1 guarantees $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ on Example 1.

Let $w = (-1, -1)^\top \in \mathbb{R}^d$. We have

$$r' := Xw = (2, 1, -1, -2)^\top, \tag{20}$$

which preserves the ordering of $r \in \mathbb{R}^K$, such that for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$, which means Example 1 satisfies the conditions in Theorem 1. The results then follow by using Theorem 1.

Second part. Algorithm 2 guarantees $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ on Example 1.

First, note that $r = (9, 8, 7, 6)^\top$, and $a^* = \arg\max_{a \in [K]} r(a) = 1$. Next, by calculation, we have,

$$\hat{r} := X(X^\top X)^{-1} X^\top r = \frac{1}{5} \cdot (22, -4, -11, 8)^\top. \tag{21}$$

Therefore, we have, $\hat{r}(a^*) = \hat{r}(1) > \hat{r}(a)$ for all $a \neq a^*$, which means Example 1 satisfies the conditions in Theorem 2. The results then follow by using Theorem 2. □

Lemma 1 (No stationary points in finite region)**.** Under the same conditions as Theorem 1, and $r(i) \neq r(j)$ for all $i \neq j$ (unique action reward), Algorithm 1 assures $\|\theta_t\|_2 \to \infty$ and $\pi_{\theta_t}(i) \to 1$ for an action $i \in [K]$ as $t \to \infty$.

Proof. According to Lemma 3, we have, for all $t \geq 1$,

$$\left| (\pi_{\theta_{t+1}} - \pi_{\theta_t})^\top r - \Big\langle \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle \right| \leq \frac{9}{4} \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X) \cdot \|\theta_{t+1} - \theta_t\|_2^2, \tag{22}$$

which implies that,

$$\begin{align} \pi_{\theta_{t+1}}^\top r - \pi_{\theta_t}^\top r &\geq \Big\langle \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t}, \theta_{t+1} - \theta_t \Big\rangle - \frac{9}{4} \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X) \cdot \|\theta_{t+1} - \theta_t\|_2^2 \tag{23} \\ &= \Big(\eta - \eta^2 \cdot \frac{9}{4} \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X) \Big) \cdot \left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2. \tag{24} \end{align}$$

Using a constant learning rate,

$$0 < \eta < \frac{4}{9 \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X)}, \tag{25}$$

we have,

$$\pi_{\theta_{t+1}}^\top r - \pi_{\theta_t}^\top r \geq \eta \cdot \Big(1 - \eta \cdot \frac{9 \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X)}{4} \Big) \cdot \left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 \geq 0. \tag{26}$$

Note that $\pi_{\theta_t}^\top r \leq r(a^*) < \infty$. According to the monotone convergence, $\pi_{\theta_t}^\top r \to c \leq r(a^*)$ as $t \to \infty$. According to Eq. (26), we have,

$$\lim_{t \to \infty} \left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2^2 = 0. \tag{27}$$

Next, we prove that there is no stationary points in finite region by contradiction. Suppose there exists $\theta' \in \mathbb{R}^d$ ($\|\theta'\|_2 < \infty$), such that,

$$\frac{d\, \pi_{\theta'}^\top r}{d\theta'} = X^\top \left(\text{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) r = \mathbf{0}. \tag{28}$$

Taking inner product with $w \in \mathbb{R}^K$ on both sides of Eq. (28), we have,

$$\begin{align} w^\top X^\top \left(\text{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) r &= {r'}^\top \left(\text{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) r \qquad (r' := Xw) \tag{29} \\ &= w^\top \mathbf{0} = 0. \tag{30} \end{align}$$

Since $\|\theta'\|_2 < \infty$ and X is bounded ($\max_{i \in [K],\, j \in [d]} |X_{i,j}| \le C$ for some $C < \infty$), we have, for all $i \in [K]$,

$$\pi_{\theta'}(i) = \frac{\exp\{[X\theta'](i)\}}{\sum_{j \in [K]} \exp\{[X\theta'](j)\}} > 0. \tag{31}$$

Next, according to Lemma 4, we have,

$$r'^\top \left(\text{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) r = \sum_{i=1}^{K-1} \pi_{\theta'}(i) \cdot \sum_{j=i+1}^{K} \pi_{\theta'}(j) \cdot (r'(i) - r'(j)) \cdot (r(i) - r(j)). \tag{32}$$

Given any non-trivial reward vector, i.e., $r \ne c \cdot \mathbf{1}$ for any $c \in \mathbb{R}$, since $r' \in \mathbb{R}^K$ preserves the order of $r \in \mathbb{R}^K$, i.e., for all $i, j \in [K]$, $r(i) > r(j)$ iff $r'(i) > r'(j)$, we have, for all $i, j \in [K]$,

$$(r'(i) - r'(j)) \cdot (r(i) - r(j)) \ge 0. \tag{33}$$

On the other hand, since $r \ne c \cdot \mathbf{1}$, there exists at least one pair of $i \ne j$, such that,

$$(r'(i) - r'(j)) \cdot (r(i) - r(j)) > 0. \tag{34}$$

Combining Eqs. (28), (29) and (31) to (34), we have,

$$\begin{align} 0 = w^\top \mathbf{0} = w^\top \left(\frac{d\, \pi_{\theta'}^\top r}{d\theta'} \right) \tag{35} \\ = w^\top X^\top \left(\text{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) r \tag{36} \\ = r'^\top \left(\text{diag}(\pi_{\theta'}) - \pi_{\theta'} \pi_{\theta'}^\top \right) r \tag{37} \\ > 0, \tag{38} \end{align}$$

which is a contradiction. Thus we have, for any $\theta' \in \mathbb{R}^d$ ($\|\theta'\|_2 < \infty$), θ' is not a stationary point.

Next, we show that $\|\theta_t\|_2 \to \infty$ as $t \to \infty$ also by contradiction. Suppose there exists $C < 0$, such that for all $t \ge 1$,

$$\theta_t \in S_C \coloneqq \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le C\}. \tag{39}$$

From the above arguments, we have, for all $\theta \in S_C$, $\left\| \frac{d\, \pi_\theta^\top r}{d\theta} \right\|_2 > 0$. Since S_C is compact, we have,

$$\inf_{\theta \in S_C} \left\| \frac{d\, \pi_\theta^\top r}{d\theta} \right\|_2 \ge \varepsilon > 0, \tag{40}$$

for some $\varepsilon > 0$, which implies that, for all $t \ge 1$,

$$\left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \ge \varepsilon > 0, \tag{41}$$

contradicting Eq. (27). Therefore, we have, $\|\theta_t\|_2 \to \infty$ as $t \to \infty$.

Next, we show that $\pi_{\theta_t}(i) \to 1$ for an action $i \in [K]$ as $t \to \infty$. Suppose $\pi_{\theta_t}(i) \not\to 1$ for any action $i \in [K]$, then there exists at least two different actions $j \ne k$ such that $\pi_{\theta_t}(j) \not\to 0$ and $\pi_{\theta_t}(k) \not\to 0$. Using similar calculations in Eq. (32), we have, $\left\| \frac{d\, \pi_{\theta_t}^\top r}{d\theta_t} \right\|_2 \not\to 0$ as $t \to \infty$, contradicting Eq. (27). Therefore, $\pi_{\theta_t}(i) \to 1$ for an action $i \in [K]$ as $t \to \infty$, i.e. π_{θ_t} approaches a one-hot policy. □

Lemma 2 (Non-negative co-variance of order preservation)**.** If $r' \in \mathbb{R}^K$ preserves the order of $r \in \mathbb{R}^K$, i.e., for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$, then for any policy $\pi \in \Delta(K)$,

$$r'^\top \left(\text{diag}(\pi) - \pi \pi^\top \right) r = \text{Cov}_\pi \left(r', r \right) \ge 0. \tag{42}$$

Proof. According to Lemma 4, we have, for all policy $\pi \in \Delta(K)$,

$$r'^\top \left(\text{diag}(\pi) - \pi \pi^\top \right) r = \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot (r'(i) - r'(j)) \cdot (r(i) - r(j)). \tag{43}$$

Since $r' \in \mathbb{R}^K$ preserves the order of $r \in \mathbb{R}^K$, i.e., for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$, we have, for all $i \ne j$,

$$(r'(i) - r'(j)) \cdot (r(i) - r(j)) \ge 0. \tag{44}$$

Combining Eqs. (43) and (44), we have Eq. (42). □

Theorem 1 (Reward order preservation, non-domination features)**.** Given any reward $r \in \mathbb{R}^K$ and feature matrix $X \in \mathbb{R}^{K \times d}$. Denote $x_i \in \mathbb{R}^d$ as the i-th row vector of X. If **(i)** $x_i^\top x_i > x_i^\top x_j$ for all $j \neq i$, and **(ii)** there exists at least one $w \in \mathbb{R}^d$, s.t., $r' \coloneqq Xw$ preserves the order of r, i.e., for all $i, j \in [K]$, $r(i) > r(j)$ if and only if $r'(i) > r'(j)$, then for any initialization $\theta_1 \in \mathbb{R}^d$, Algorithm 1 with a constant learning rate $\eta > 0$ achieves global convergence of $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$.

Proof. **First part.** According to Lemma 1, using any constant learning rate,

$$0 < \eta < \frac{4}{9 \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X)}, \tag{45}$$

Algorithm 1 guarantees that $\|\theta_t\|_2 \to \infty$ as $t \to \infty$, and $\pi_{\theta_t}(i) \to 1$ for an action $i \in [K]$ as $t \to \infty$.

Second part. For the direction $w \in \mathbb{R}^d$ such that $r' \coloneqq Xw$ preserves the order of r. We have,

$$w^\top \theta_{t+1} = w^\top \theta_t + \eta \cdot w^\top X^\top \left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t} \pi_{\theta_t}^\top \right) r \qquad \text{(by Algorithm 1)} \tag{46}$$

$$= w^\top \theta_t + \eta \cdot {r'}^\top \left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t} \pi_{\theta_t}^\top \right) r \qquad (r' \coloneqq Xw) \tag{47}$$

$$\geq w^\top \theta_t. \qquad \text{(by Lemma 2)} \tag{48}$$

Third part. Suppose there exists a sub-optimal action $i \in [K]$ with $r(i) < r(a^*)$, and $\pi_{\theta_t}(i) \to 1$ as $t \to \infty$. Then we have, for all large enough $t \geq 1$,

$$\pi_{\theta_t}(i) > \pi_{\theta_t}(j), \tag{49}$$

for all $j \neq i$, which implies that,

$$\left[\frac{X\theta_t}{\|\theta_t\|_2} \right](i) = \max_{a \in [K]} \left[\frac{X\theta_t}{\|\theta_t\|_2} \right](a). \tag{50}$$

Now we prove by contradiction that the assumption of $\pi_{\theta_t}(i) \to 1$ as $t \to \infty$ cannot be true for any sub-optimal action $i \in [K]$ with $r(i) < r(a^*)$.

Using the sub-optimal action's reward $r(i)$, the action set $[K]$ can be partitioned as follows,

$$\mathcal{A}(i) \coloneqq \{ j \in [K] : r(j) = r(i) \}, \tag{51}$$

$$\mathcal{A}^+(i) \coloneqq \left\{ a^+ \in [K] : r(a^+) > r(i) \right\}, \tag{52}$$

$$\mathcal{A}^-(i) \coloneqq \left\{ a^- \in [K] : r(a^-) < r(i) \right\}. \tag{53}$$

According to Eq. (50), for all large enough $t \geq 1$, $i = \arg\max_{a \in [K]} [X\theta_t](a)$. Take the second largest component of $X\theta_t$, and denote the corresponding action index as j.

Case 1. $j \in \mathcal{A}^+(i)$. This means $j = a^+$ for a "good" action with $r(a^+) > r(i)$.

We have, for all large enough $t \geq 1$, for all "bad" action $a^- \in \mathcal{A}^-(i)$,

$$[X\theta_t](a^+) - [X\theta_t](a^-) = \|\theta_t\|_2 \cdot \left(\left[\frac{X\theta_t}{\|\theta_t\|_2} \right](a^+) - \left[\frac{X\theta_t}{\|\theta_t\|_2} \right](a^-) \right) \tag{54}$$

$$\geq c \cdot \|\theta_t\|_2, \tag{55}$$

for some $c > 0$ according to $j = a^+$. Next, we have,

$$\frac{\pi_{\theta_t}(a^+)}{\pi_{\theta_t}(a^-)} = \exp\left\{ [X\theta_t](a^+) - [X\theta_t](a^-) \right\} \geq \exp\left\{ c \cdot \|\theta_t\|_2 \right\}, \tag{56}$$

which implies that,

$$r(i) - \pi_{\theta_t}^\top r = \sum_{k \neq i} \pi_{\theta_t}(k) \cdot (r(i) - r(k)) \tag{57}$$

$$= - \sum_{\tilde{a}^+ \in \mathcal{A}^+(i)} \pi_{\theta_t}(\tilde{a}^+) \cdot \left(r(\tilde{a}^+) - r(i) \right) + \sum_{a^- \in \mathcal{A}^-(i)} \pi_{\theta_t}(a^-) \cdot \left(r(i) - r(a^-) \right) \tag{58}$$

$$\leq -\pi_{\theta_t}(a^+) \cdot \left(r(a^+) - r(i) \right) + \sum_{a^- \in \mathcal{A}^-(i)} \pi_{\theta_t}(a^-) \cdot \left(r(i) - r(a^-) \right) \tag{59}$$

$$= -\pi_{\theta_t}(a^+) \cdot \Bigg[\underbrace{r(a^+) - r(i)}_{>0} - \sum_{a^- \in \mathcal{A}^-(i)} \underbrace{\frac{\pi_{\theta_t}(a^-)}{\pi_{\theta_t}(a^+)}}_{\to 0} \cdot \left(r(i) - r(a^-) \right) \Bigg] \tag{60}$$

$$< 0, \tag{61}$$

where $r(a^+) - r(i) > 0$ is from Eq. (52), $\frac{\pi_{\theta_t}(a^-)}{\pi_{\theta_t}(a^+)} \to 0$ as $t \to \infty$ is by Eq. (56) and Lemma 1. Eq. (57) means that $\pi_{\theta_t}^\top r > r(i)$ happens at a finite time $t < \infty$. According to Eq. (26), we have, for all large enough $t \ge 1$, $\pi_{\theta_t}^\top r > r(i)$, which is a contradiction with $\pi_{\theta_t}^\top r \to r(i)$ implied by the assumption of $\pi_{\theta_t}(i) \to 1$ as $t \to \infty$.

Case 2. $j \in \mathcal{A}^-(i)$. This means $j = a^-$ for a "bad" action $r(a^-) < r(i)$.

Using similar arguments around Eq. (56), we have, for all $a \in [K]$ such that $a \ne i$ and $a \ne a^-$,

$$\frac{\pi_{\theta_t}(a^-)}{\pi_{\theta_t}(a)} = \exp\left\{ [X\theta_t](a^-) - [X\theta_t](a) \right\} \ge \exp\left\{ c \cdot \|\theta_t\|_2 \right\}, \tag{62}$$

for some $c > 0$. Consider a direction $u \in \mathbb{R}^d$, $\|u\|_2 = 1$, such that,

$$[Xu](a^-) = \max_{a \in [K]} [Xu](a). \tag{63}$$

According to Algorithm 1, we have,

$$\begin{aligned} u^\top \theta_{t+1} &= u^\top \theta_t + \eta \cdot u^\top X^\top \left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top \right) \, r & (64) \\ &= u^\top \theta_t + \eta \cdot u^\top X^\top \left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top \right) \left(r - r(i) \cdot \mathbf{1} \right), & (65) \end{aligned}$$

where the last equation is because of,

$$\left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top \right) \, \mathbf{1} = \pi_{\theta_t} - \pi_{\theta_t} \cdot (\pi_{\theta_t}^\top \mathbf{1}) = \pi_{\theta_t} - \pi_{\theta_t} = \mathbf{0}. \tag{66}$$

Denote $y \coloneqq Xu$. We have,

$$\left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top \right) Xu = \begin{bmatrix} \pi_{\theta_t}(1) \cdot \left(y(1) - \pi_{\theta_t}^\top y \right) \\ \pi_{\theta_t}(2) \cdot \left(y(2) - \pi_{\theta_t}^\top y \right) \\ \vdots \\ \pi_{\theta_t}(K) \cdot \left(y(K) - \pi_{\theta_t}^\top y \right) \end{bmatrix} \in \mathbb{R}^K. \tag{67}$$

Therefore, from Eqs. (64) and (67), we have,

$$\begin{aligned} & u^\top X^\top \left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top \right) \, r = \sum_{a \ne i} \pi_{\theta_t}(a) \cdot \left(y(a) - \pi_{\theta_t}^\top y \right) \cdot \left(r(a) - r(i) \right) & (68) \\ & \quad = \sum_{a \ne i, \ a \ne a^-} \pi_{\theta_t}(a) \cdot \left(y(a) - \pi_{\theta_t}^\top y \right) \cdot \left(r(a) - r(i) \right) + \pi_{\theta_t}(a^-) \cdot \left(y(a^-) - \pi_{\theta_t}^\top y \right) \cdot \left(r(a^-) - r(i) \right) & (69) \\ & \quad = -\pi_{\theta_t}(a^-) \cdot \left[\underbrace{\left(y(a^-) - \pi_{\theta_t}^\top y \right)}_{>0} \cdot \underbrace{\left(r(i) - r(a^-) \right)}_{>0} - \sum_{\substack{a \ne i, \\ a \ne a^-}} \underbrace{\frac{\pi_{\theta_t}(a)}{\pi_{\theta_t}(a^-)}}_{\to 0} \cdot \underbrace{\left(y(a) - \pi_{\theta_t}^\top y \right)}_{\text{bounded}} \cdot \left(r(a) - r(i) \right) \right] & (70) \\ & \quad < 0, & (71) \end{aligned}$$

where $y(a^-) - \pi_{\theta_t}^\top y > 0$ is by Eq. (63), $r(i) - r(a^-) > 0$ is from Eq. (53), $\frac{\pi_{\theta_t}(a)}{\pi_{\theta_t}(a^-)} \to 0$ as $t \to \infty$ is according to Eq. (62) and Lemma 1, and $y(a) - \pi_{\theta_t}^\top y$ is bounded is because of X and u are bounded ($\max_{i \in [K], \ j \in [d]} |X_{i,j}| \le C$, and $\max_{j \in [d]} |u(j)| \le C$ for some $C < \infty$).

Combining Eqs. (64) and (68), we have, for all large enough $t \ge 1$,

$$\begin{aligned} u^\top \theta_{t+1} &= u^\top \theta_t + \eta \cdot u^\top X^\top \left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top \right) \, r & (72) \\ &< u^\top \theta_t, & (73) \end{aligned}$$

which implies that $u^\top \theta_t \to -\infty$ as $t \to \infty$ according to Lemma 1. On the other hand, according to Eq. (46), we have, for all large enough $t \ge 1$,

$$w^\top \theta_{t+1} > w^\top \theta_t, \tag{74}$$

which implies that $w^\top \theta_t \to \infty$ as $t \to \infty$ according to Lemma 1. According to the non-domination feature condition, i.e., $x_i^\top x_i > x_i^\top x_j$ for all $i \ne j$, we have, for any "bad" action $a^- \in \mathcal{A}^-(i)$,

$$[Xx_{a^-}](a^-) = \max_{a \in [K]} [Xx_{a^-}](a), \tag{75}$$

which implies that,

$$[X\theta_{t+1}](a^-) = x_{a^-}^\top \theta_{t+1} < x_{a^-}^\top \theta_t = [X\theta_t](a^-), \tag{76}$$

by taking $u = x_{a^-}$ in Eq. (72). This means the score of a "bad" action $a^- \in \mathcal{A}^-(i)$ is monotonically decreasing, and it approaches $-\infty$ due to Lemma 1. On the other hand, take $w = x_{a^*}$ in Eq. (74) (or take $u = -x_{a^*}$ in Eq. (72)), we have,

$$[X\theta_{t+1}](a^*) = x_{a^*}^\top \theta_{t+1} > x_{a^*}^\top \theta_t = [X\theta_t](a^*), \tag{77}$$

which means the optimal action's score is monotonically increasing, it approaches ∞ due to Lemma 1. Therefore, we have, for all large enough $t \ge 1$,

$$[X\theta_t](a^*) > [X\theta_t](a^-), \tag{78}$$

for any "bad" action $a^- \in \mathcal{A}^-(i)$, contradicting the assumption of $j = a^- \in \mathcal{A}^-(i)$. □

Proposition 2. Let $K = 3$, $d = 2$, $X^\top = \begin{bmatrix} 0 & -10 & 0 \\ -2 & 4 & 1 \end{bmatrix} \in \mathbb{R}^{d \times K}$, and $r = Xw = (4, 2, -2)^\top$, where $w = (-1, -2)^\top \in \mathbb{R}^d$. With initialization $\theta_1 = (-\ln 2, \ln 2)^\top$, Algorithm 1 does not achieve global convergence, i.e., $\pi_{\theta_t}(1) \not\to 1$ as $t \to \infty$.

Proof. Define the following region,

$$\mathcal{R} \coloneqq \left\{ \theta \in \mathbb{R}^d : \frac{\pi_\theta(1)}{\pi_\theta(3)} < \frac{1}{2}, \text{ and } \frac{\pi_\theta(2)}{\pi_\theta(3)} > 3 \right\}. \tag{79}$$

We show that, **(i)** $\theta_1 = (-\ln 2, \ln 2)^\top \in \mathcal{R}$, and **(ii)** if $\theta_t \in \mathcal{R}$, then $\theta_{t+1} \in \mathcal{R}$, and,

$$\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} < \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)}, \tag{80}$$

which means that, for all $t \ge 1$, we have, $\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < 1/2$, implying that $\pi_{\theta_t}(1) \not\to 1$ as $t \to \infty$.

First part. (i) $\theta_1 = (-\ln 2, \ln 2)^\top \in \mathcal{R}$.

For the initialization $\theta_1 = (-\ln 2, \ln 2)^\top$, we have,

$$\exp\{X\theta_1\} = (2^{-2}, 2^{14}, 2)^\top. \tag{81}$$

By calculation, we have,

$$\frac{\pi_{\theta_1}(1)}{\pi_{\theta_1}(3)} = \frac{\exp\{[X\theta_1](1)\}}{\exp\{[X\theta_1](3)\}} = \frac{1}{8} < \frac{1}{4} \cdot \frac{r(2) - r(3)}{r(1) - r(2)} = \frac{1}{2}, \text{ and} \tag{82}$$

$$\frac{\pi_{\theta_1}(2)}{\pi_{\theta_1}(3)} = \frac{\exp\{[X\theta_1](2)\}}{\exp\{[X\theta_1](3)\}} = 2^{13} > \frac{r(1) - r(3)}{r(1) - r(2)} = 3, \tag{83}$$

which verifies that $\theta_1 = (-\ln 2, \ln 2)^\top \in \mathcal{R}$.

Second part. (ii) If $\theta_t \in \mathcal{R}$, then $\theta_{t+1} \in \mathcal{R}$, and $\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} < \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)}$.

Suppose $\theta_t \in \mathcal{R}$, we have,

$$\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < \frac{1}{2}, \text{ and } \frac{\pi_{\theta_t}(2)}{\pi_{\theta_t}(3)} > 3. \tag{84}$$

Next, we have,

$$\begin{align} r(2) - \pi_{\theta_t}^\top r &= \pi_{\theta_t}(1) \cdot (r(2) - r(1)) + \pi_{\theta_t}(3) \cdot (r(2) - r(3)) \tag{85} \\ &= -2 \cdot \pi_{\theta_t}(1) + 4 \cdot \pi_{\theta_t}(3) \tag{86} \\ &= 2 \cdot \pi_{\theta_t}(3) \cdot \left(-\frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} + 2 \right) > 0, \tag{87} \end{align}$$

and

$$\begin{aligned}\frac{r(1)-\pi_{\theta_t}^\top r}{\pi_{\theta_t}^\top r - r(3)} &= \frac{\pi_{\theta_t}(2)\cdot (r(1)-r(2)) + \pi_{\theta_t}(3)\cdot (r(1)-r(3))}{\pi_{\theta_t}(1)\cdot (r(1)-r(3)) + \pi_{\theta_t}(2)\cdot (r(2)-r(3))} && (88)\\ &= \frac{2\cdot \pi_{\theta_t}(2) + 6\cdot \pi_{\theta_t}(3)}{6\cdot \pi_{\theta_t}(1) + 4\cdot \pi_{\theta_t}(2)} && (89)\\ &< \frac{2\cdot \pi_{\theta_t}(2) + 6\cdot \pi_{\theta_t}(3)}{4\cdot \pi_{\theta_t}(2)} && (90)\\ &= \frac{1}{2} + \frac{3}{2}\cdot \frac{\pi_{\theta_t}(3)}{\pi_{\theta_t}(2)} < 1. && (91)\end{aligned}$$

According to Algorithm 1, we have,

$$\begin{aligned}\theta_{t+1}-\theta_t &= \eta \cdot X^\top \left(\text{diag}(\pi_{\theta_t}) - \pi_{\theta_t}\pi_{\theta_t}^\top\right)\ r && (92)\\ &= \eta\cdot \begin{bmatrix} 0 & -10 & 0 \\ -2 & 4 & 1\end{bmatrix} \begin{bmatrix} \pi_{\theta_t}(1)\cdot \left(r(1)-\pi_{\theta_t}^\top r\right) \\ \pi_{\theta_t}(2)\cdot \left(r(2)-\pi_{\theta_t}^\top r\right) \\ \pi_{\theta_t}(3)\cdot \left(r(3)-\pi_{\theta_t}^\top r\right)\end{bmatrix} && (93)\\ &= \eta\cdot \begin{bmatrix} -10\cdot \pi_{\theta_t}(2)\cdot \left(r(2)-\pi_{\theta_t}^\top r\right) \\ -2\cdot \pi_{\theta_t}(1)\cdot \left(r(1)-\pi_{\theta_t}^\top r\right) + 4\cdot \pi_{\theta_t}(2)\cdot \left(r(2)-\pi_{\theta_t}^\top r\right) + \pi_{\theta_t}(3)\cdot \left(r(3)-\pi_{\theta_t}^\top r\right)\end{bmatrix}. && (94)\end{aligned}$$

Next, we have,

$$\begin{aligned}& -2\cdot \pi_{\theta_t}(1)\cdot \left(r(1)-\pi_{\theta_t}^\top r\right) + 4\cdot \pi_{\theta_t}(2)\cdot \left(r(2)-\pi_{\theta_t}^\top r\right) + \pi_{\theta_t}(3)\cdot \left(r(3)-\pi_{\theta_t}^\top r\right) && (95)\\ &= -6\cdot \pi_{\theta_t}(1)\cdot \left(r(1)-\pi_{\theta_t}^\top r\right) - 3\cdot \pi_{\theta_t}(3)\cdot \left(r(3)-\pi_{\theta_t}^\top r\right) && (96)\\ &= 3\cdot \pi_{\theta_t}(3)\cdot \left(\pi_{\theta_t}^\top r - r(3)\right)\cdot \left[-2\cdot \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)}\cdot \frac{r(1)-\pi_{\theta_t}^\top r}{\pi_{\theta_t}^\top r - r(3)} + 1\right] && (97)\\ &> 3\cdot \pi_{\theta_t}(3)\cdot \left(\pi_{\theta_t}^\top r - r(3)\right)\cdot \left(-2\cdot \frac{1}{2}\cdot 1 + 1\right) = 0, && (98)\end{aligned}$$

which implies that,

$$\theta_{t+1}(2) > \theta_t(2). \tag{99}$$

Therefore, we have,

$$\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} = \frac{\exp\{[X\theta_{t+1}](1)\}}{\exp\{[X\theta_{t+1}](3)\}} = \frac{\exp\{-2\cdot \theta_{t+1}(2)\}}{\exp\{\theta_{t+1}(2)\}} < \frac{\exp\{-2\cdot \theta_t(2)\}}{\exp\{\theta_t(2)\}} = \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)} < \frac{1}{2}. \tag{100}$$

On the other hand, we have,

$$-10\cdot \pi_{\theta_t}(2)\cdot \left(r(2)-\pi_{\theta_t}^\top r\right) < 0, \tag{101}$$

which implies that,

$$\theta_{t+1}(1) < \theta_t(1). \tag{102}$$

Therefore, we have,

$$\begin{aligned}\frac{\pi_{\theta_{t+1}}(2)}{\pi_{\theta_{t+1}}(3)} &= \frac{\exp\{[X\theta_{t+1}](2)\}}{\exp\{[X\theta_{t+1}](3)\}} = \frac{\exp\{-10\cdot \theta_{t+1}(1) + 4\cdot \theta_{t+1}(2)\}}{\exp\{\theta_{t+1}(2)\}} && (103)\\ &> \frac{\exp\{[X\theta_{t+1}](2)\}}{\exp\{[X\theta_{t+1}](3)\}} = \frac{\exp\{-10\cdot \theta_t(1) + 4\cdot \theta_t(2)\}}{\exp\{\theta_t(2)\}} && (104)\\ &= \frac{\pi_{\theta_t}(2)}{\pi_{\theta_t}(3)} > 3, && (105)\end{aligned}$$

which proves that $\theta_{t+1} \in \mathcal{R}$ and $\frac{\pi_{\theta_{t+1}}(1)}{\pi_{\theta_{t+1}}(3)} < \frac{\pi_{\theta_t}(1)}{\pi_{\theta_t}(3)}$. □

Theorem 2 (Optimal action preservation condition)**.** For a constant learning rate $\eta > 0$, a necessary and sufficient condition for Algorithm 2 to achieve global convergence $\pi_{\theta_t}^\top r \to r(a^*)$ as $t \to \infty$ from any initialization $\theta_1 \in \mathbb{R}^d$ is that $\hat{r}(a^*) > \hat{r}(a)$ for all $a \neq a^*$, such that $a^* := \arg\max_{a\in[K]} r(a)$, and $\hat{r} := X\left(X^\top X\right)^{-1} X^\top r$ is the least squares projection of r onto the column space of X. If the condition is satisfied, then the rate of convergence is $\left(\pi^* - \pi_{\theta_t}\right)^\top r \in O(e^{-c\cdot t})$ for some $c > 0$.

Proof. **First part.** Sufficiency. Suppose that $\hat{r}(a^*) > \hat{r}(a)$ for all $a \neq a^*$. Denote

$$\hat{\Delta} \coloneqq \hat{r}(a^*) - \max_{a \neq a^*} \hat{r}(a). \tag{106}$$

According to Algorithm 2, we have, for all $t \geq 1$,

$$X\theta_{t+1} = X\theta_t + \eta \cdot X \left(X^\top X \right)^{-1} X^\top r = X\theta_t + \eta \cdot \hat{r}. \tag{107}$$

Next, we have, for all $a \neq a^*$,

$$\begin{aligned}
\frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)} &= \exp\left\{ [X\theta_{t+1}](a^*) - [X\theta_{t+1}](a) \right\} \qquad \text{(by Eq. (3))} && (108)\\
&= \exp\left\{ [X\theta_t](a^*) - [X\theta_t](a) + \eta \cdot (\hat{r}(a^*) - \hat{r}(a)) \right\} \qquad \text{(by Eq. (107))} && (109)\\
&= \exp\left\{ [X\theta_1](a^*) - [X\theta_1](a) + \eta \cdot (\hat{r}(a^*) - \hat{r}(a)) \cdot t \right\} && (110)\\
&\geq \exp\left\{ [X\theta_1](a^*) - [X\theta_1](a) + \eta \cdot \hat{\Delta} \cdot t \right\} \qquad \text{(by Eq. (106))} && (111)\\
&= \frac{\pi_{\theta_1}(a^*)}{\pi_{\theta_1}(a)} \cdot e^{\eta \cdot \hat{\Delta} \cdot t}, && (112)
\end{aligned}$$

which implies that,

$$\begin{aligned}
\frac{1}{\pi_{\theta_t}(a^*)} - 1 &= \sum_{a \neq a^*} \frac{\pi_{\theta_t}(a)}{\pi_{\theta_t}(a^*)} && (113)\\
&\leq \sum_{a \neq a^*} \frac{\pi_{\theta_1}(a)}{\pi_{\theta_1}(a^*)} \cdot e^{-\eta \cdot \hat{\Delta} \cdot (t-1)} \qquad \text{(by Eq. (108))} && (114)\\
&\leq c(X, \theta_1) \cdot K \cdot e^{-\eta \cdot \hat{\Delta} \cdot (t-1)}, && (115)
\end{aligned}$$

where

$$c(X, \theta_1) \coloneqq \max_{a \neq a^*} \frac{\pi_{\theta_1}(a)}{\pi_{\theta_1}(a^*)}. \tag{116}$$

Therefore, we have,

$$\begin{aligned}
\left(\pi^* - \pi_{\theta_t} \right)^\top r &= \sum_{a \neq a^*} \pi_{\theta_t}(a) \cdot (r(a^*) - r(a)) && (117)\\
&\leq 2 \cdot \|r\|_\infty \cdot (1 - \pi_{\theta_t}(a^*)) \qquad \left(\text{using } r \in \left[-\|r\|_\infty, \|r\|_\infty \right]^K \right) && (118)\\
&= 2 \cdot \|r\|_\infty \cdot \left(1 - \frac{1}{\frac{1}{\pi_{\theta_t}(a^*)} - 1 + 1} \right) && (119)\\
&\leq 2 \cdot \|r\|_\infty \cdot \left(1 - \frac{1}{c(X, \theta_1) \cdot K \cdot e^{-\eta \cdot \hat{\Delta} \cdot (t-1)} + 1} \right) \qquad \text{(by Eq. (113))} && (120)\\
&= \frac{2 \cdot \|r\|_\infty \cdot c(X, \theta_1) \cdot K}{c(X, \theta_1) \cdot K + \exp\left\{ \eta \cdot \hat{\Delta} \cdot (t-1) \right\}}, && (121)
\end{aligned}$$

which proves the sufficiency and the convergence rate $\left(\pi^* - \pi_{\theta_t} \right)^\top r \in O(e^{-c \cdot t})$ for some $c > 0$.

Second part. Necessity. Suppose the condition is not satisfied, i.e., there exists one sub-optimal action $a \neq a^*$, such that $\hat{r}(a^*) \leq \hat{r}(a)$. We have, for all $t \geq 1$,

$$\begin{aligned}
\frac{\pi_{\theta_{t+1}}(a^*)}{\pi_{\theta_{t+1}}(a)} &= \exp\left\{ [X\theta_{t+1}](a^*) - [X\theta_{t+1}](a) \right\} \qquad \text{(by Eq. (3))} && (122)\\
&= \exp\left\{ [X\theta_t](a^*) - [X\theta_t](a) + \eta \cdot (\hat{r}(a^*) - \hat{r}(a)) \right\} \qquad \text{(by Eq. (107))} && (123)\\
&\leq \exp\left\{ [X\theta_1](a^*) - [X\theta_1](a) + \eta \cdot (\hat{r}(a^*) - \hat{r}(a)) \cdot t \right\} && (124)\\
&\leq \frac{\pi_{\theta_1}(a^*)}{\pi_{\theta_1}(a)}, \qquad \text{(using } \hat{r}(a^*) \leq \hat{r}(a)) && (125)
\end{aligned}$$

which implies that,

$$\frac{1}{\pi_{\theta_t}(a^*)} - 1 = \sum_{a' \neq a^*} \frac{\pi_{\theta_t}(a')}{\pi_{\theta_t}(a^*)} \tag{126}$$

$$\geq \frac{\pi_{\theta_t}(a)}{\pi_{\theta_t}(a^*)} \qquad (\pi_{\theta_t}(a') > 0 \text{ for all } a' \in [K]) \tag{127}$$

$$\geq \frac{\pi_{\theta_1}(a)}{\pi_{\theta_1}(a^*)}. \qquad \text{(by Eq. (122))} \tag{128}$$

Therefore, we have,

$$(\pi^* - \pi_{\theta_t})^\top r = \sum_{a \neq a^*} \pi_{\theta_t}(a) \cdot (r(a^*) - r(a)) \tag{129}$$

$$\geq \Delta \cdot (1 - \pi_{\theta_t}(a^*)) \qquad \left(\Delta \coloneqq r(a^*) - \max_{a \neq a^*} r(a)\right) \tag{130}$$

$$= \Delta \cdot \left(1 - \frac{1}{\frac{1}{\pi_{\theta_t}(a^*)} - 1 + 1}\right) \tag{131}$$

$$\geq \Delta \cdot \left(1 - \frac{1}{\frac{\pi_{\theta_1}(a)}{\pi_{\theta_1}(a^*)} + 1}\right) \qquad \text{(by Eq. (126))} \tag{132}$$

$$= \frac{\Delta \cdot \pi_{\theta_1}(a)}{\pi_{\theta_1}(a) + \pi_{\theta_1}(a^*)} > 0, \tag{133}$$

i.e., $\pi_{\theta_t}^\top r \not\to r(a^*)$ as $t \to \infty$, which proves the necessity of the condition. □

B Miscellaneous Extra Supporting Results

Lemma 3 (Smoothness). *Given any reward vector $r \in \mathbb{R}^K$ and feature matrix $X \in \mathbb{R}^{K \times d}$. The expected reward function $\theta \mapsto \pi_\theta^\top r$ with $\pi_\theta = \mathrm{softmax}(X\theta)$ is β-smooth with*

$$\beta = \frac{9}{2} \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X), \tag{134}$$

i.e., for all θ, $\theta' \in \mathbb{R}^d$,

$$\left| (\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle \right| \leq \frac{9}{4} \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X) \cdot \|\theta' - \theta\|_2^2. \tag{135}$$

Proof. Let $S \coloneqq S(X, r, \theta) \in \mathbb{R}^{d \times d}$ be the second-order derivative of the value map $\theta \mapsto \pi_\theta^\top r$. By Taylor's theorem, it suffices to show that the spectral radius of S (regardless of θ) is bounded by β. Now, by its definition we have

$$S = \frac{d}{d\theta} \left\{ \frac{d\, \pi_\theta^\top r}{d\theta} \right\} \tag{136}$$

$$= \frac{d}{d\theta} \left\{ X^\top (\mathrm{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top)\, r \right\}. \qquad \text{(by Eq. (4))} \tag{137}$$

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

$$S_{i,j} = \frac{d\{ \sum_{a=1}^K X_{a,i} \cdot \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r)\}}{d\theta(j)} \tag{138}$$

$$= \sum_{a=1}^K X_{a,i} \cdot \frac{d\pi_\theta(a)}{d\theta(j)} \cdot \left(r(a) - \pi_\theta^\top r\right) - \sum_{a=1}^K X_{a,i} \cdot \pi_\theta(a) \cdot \sum_{a'=1}^K \frac{d\pi_\theta(a')}{d\theta(j)} \cdot r(a'). \tag{139}$$

We have, for all $a \in [K]$ and $j \in [d]$,

$$\frac{d\pi_\theta(a)}{d\theta(j)} = \frac{d}{d\theta(j)}\left\{\frac{\exp\{[X\theta](a)\}}{\sum_{a' \in [K]} \exp\{[X\theta](a')\}}\right\} \tag{140}$$

$$= \frac{\frac{d\ \exp\{[X\theta](a)\}}{d\theta(j)} \cdot \sum_{a' \in [K]} \exp\{[X\theta](a')\} - \exp\{[X\theta](a)\} \cdot \frac{d\ \sum_{a' \in [K]} \exp\{[X\theta](a')\}}{d\theta(j)}}{\left(\sum_{a' \in [K]} \exp\{[X\theta](a')\}\right)^2} \tag{141}$$

$$= \frac{\exp\{[X\theta](a)\} \cdot X_{a,j} \cdot \sum_{a' \in [K]} \exp\{[X\theta](a')\} - \exp\{[X\theta](a)\} \cdot \sum_{a' \in [K]} \exp\{[X\theta](a')\} \cdot X_{a',j}}{\left(\sum_{a' \in [K]} \exp\{[X\theta](a')\}\right)^2} \tag{142}$$

$$= \frac{\exp\{[X\theta](a)\} \cdot X_{a,j} - \exp\{[X\theta](a)\} \cdot \sum_{a' \in [K]} \pi_\theta(a') \cdot X_{a',j}}{\sum_{a' \in [K]} \exp\{[X\theta](a')\}} \tag{143}$$

$$= \pi_\theta(a) \cdot \Big(X_{a,j} - \sum_{a' \in [K]} \pi_\theta(a') \cdot X_{a',j}\Big). \tag{144}$$

Combining Eqs. (138) and (140), we have,

$$S_{i,j} = \sum_{a=1}^{K} X_{a,i} \cdot \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r) \cdot X_{a,j} - \sum_{a=1}^{K} X_{a,i} \cdot \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r) \cdot \sum_{a'=1}^{K} \pi_\theta(a') \cdot X_{a',j} \tag{145}$$

$$- \sum_{a=1}^{K} X_{a,i} \cdot \pi_\theta(a) \cdot \sum_{a'=1}^{K} \pi_\theta(a') \cdot \Big(X_{a',j} - \sum_{a''=1}^{K} \pi_\theta(a'') \cdot X_{a'',j}\Big) \cdot r(a'). \tag{146}$$

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

$$\left|y^\top S y\right| = \left|\sum_{i=1}^{d}\sum_{j=1}^{d} S_{i,j} \cdot y(i) \cdot y(j)\right| \tag{147}$$

$$= \Bigg|\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{a=1}^{K} y(i) \cdot X_{a,i} \cdot \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r) \cdot X_{a,j} \cdot y(j) \tag{148}$$

$$- \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{a=1}^{K} y(i) \cdot X_{a,i} \cdot \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r) \cdot \sum_{a'=1}^{K} \pi_\theta(a') \cdot X_{a',j} \cdot y(j) \tag{149}$$

$$- \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{a=1}^{K} y(i) \cdot X_{a,i} \cdot \pi_\theta(a) \cdot \sum_{a'=1}^{K} \pi_\theta(a') \cdot \Big(X_{a',j} - \sum_{a''=1}^{K} \pi_\theta(a'') \cdot X_{a'',j}\Big) \cdot r(a') \cdot y(j)\Bigg|, \tag{150}$$

which is equal to,

$$\left|y^\top S y\right| = \Bigg|\sum_{a=1}^{K} [Xy](a) \cdot \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r) \cdot [Xy](a) \tag{151}$$

$$- \sum_{a=1}^{K} [Xy](a) \cdot \pi_\theta(a) \cdot (r(a) - \pi_\theta^\top r) \cdot \sum_{a'=1}^{K} \pi_\theta(a') \cdot [Xy](a') \tag{152}$$

$$- \sum_{a=1}^{K} [Xy](a) \cdot \pi_\theta(a) \cdot \sum_{a'=1}^{K} \pi_\theta(a') \cdot r(a') \cdot \Big([Xy](a') - \sum_{a''=1}^{K} \pi_\theta(a'') \cdot [Xy](a'')\Big)\Bigg|. \tag{153}$$

Denote

$$H(\pi_\theta) \coloneqq \text{diag}(\pi_\theta) - \pi_\theta \pi_\theta^\top \in \mathbb{R}^{K \times K}. \tag{154}$$

We have,

$$\left|y^\top S y\right| = \left|\left(H(\pi_\theta)\, r\right)^\top \left(Xy \odot Xy\right) - \left(H(\pi_\theta)\, r\right)^\top \left(Xy\right) \cdot \left(\pi_\theta^\top Xy\right) - \left(\pi_\theta^\top Xy\right) \cdot \left(H(\pi_\theta) Xy\right)^\top r\right| \tag{155}$$

$$= \left|\left(H(\pi_\theta)\, r\right)^\top \left(Xy \odot Xy\right) - 2 \cdot \left(H(\pi_\theta)\, r\right)^\top \left(Xy\right) \cdot \left(\pi_\theta^\top Xy\right)\right|, \tag{156}$$

where $\odot$ is Hadamard (component-wise) product. According to the triangle inequality and Hölder's inequality, we have,

$$\left|y^\top S y\right| \le \left|\left(H(\pi_\theta)\, r\right)^\top \left(Xy \odot Xy\right)\right| + 2 \cdot \left|\left(H(\pi_\theta)\, r\right)^\top \left(Xy\right)\right| \cdot \left|\pi_\theta^\top Xy\right| \tag{157}$$

$$\le \left\|H(\pi_\theta) r\right\|_\infty \cdot \left\|Xy \odot Xy\right\|_1 + 2 \cdot \left\|H(\pi_\theta) r\right\|_1 \cdot \left\|Xy\right\|_\infty \cdot \left\|\pi_\theta\right\|_1 \cdot \left\|Xy\right\|_\infty \tag{158}$$

$$= \left\|H(\pi_\theta) r\right\|_\infty \cdot \left\|Xy\right\|_2^2 + 2 \cdot \left\|H(\pi_\theta) r\right\|_1 \cdot \left\|Xy\right\|_\infty^2 \qquad \left(\|Xy \odot Xy\|_1 = \|Xy\|_2^2,\ \|\pi_\theta\|_1 = 1\right) \tag{159}$$

$$\le \left\|H(\pi_\theta) r\right\|_\infty \cdot \left\|Xy\right\|_2^2 + 2 \cdot \left\|H(\pi_\theta) r\right\|_1 \cdot \left\|Xy\right\|_2^2. \qquad \left(\|Xy\|_\infty \le \|Xy\|_2\right) \tag{160}$$

For $a \in [K]$, denote by $H_{a,:}(\pi_\theta)$ the a-th row of $H(\pi_\theta)$ as a row vector. Then,

$$\left\|H_{a,:}(\pi_\theta)\right\|_1 = \pi_\theta(a) - \pi_\theta(a)^2 + \pi_\theta(a) \cdot \sum_{a' \ne a} \pi_\theta(a') \tag{161}$$

$$= \pi_\theta(a) - \pi_\theta(a)^2 + \pi_\theta(a) \cdot (1 - \pi_\theta(a)) \tag{162}$$

$$= 2 \cdot \pi_\theta(a) \cdot (1 - \pi_\theta(a)) \tag{163}$$

$$\le \frac{1}{2}. \qquad (\text{using } x \cdot (1-x) \le 1/4 \text{ for all } x \in [0,1]) \tag{164}$$

On the other hand,

$$\left\|H(\pi_\theta) r\right\|_1 = \sum_{a \in [K]} \pi_\theta(a) \cdot \left|r(a) - \pi_\theta^\top r\right| \tag{165}$$

$$\le \max_{a \in [K]} \left|r(a) - \pi_\theta^\top r\right| \tag{166}$$

$$\le 2 \cdot \|r\|_\infty. \qquad \left(\text{using } r \in \left[-\|r\|_\infty, \|r\|_\infty\right]^K\right) \tag{167}$$

Therefore, we have,

$$\left|y^\top S(X, r, \theta)\, y\right| \le \left\|H(\pi_\theta) r\right\|_\infty \cdot \left\|Xy\right\|_2^2 + 2 \cdot \left\|H(\pi_\theta) r\right\|_1 \cdot \left\|Xy\right\|_2^2 \tag{168}$$

$$= \max_{a \in [K]} \left|\left(H_{a,:}(\pi_\theta)\right)^\top r\right| \cdot \left\|Xy\right\|_2^2 + 2 \cdot \left\|H(\pi_\theta) r\right\|_1 \cdot \left\|Xy\right\|_2^2 \tag{169}$$

$$\le \max_{a \in [K]} \left\|H_{a,:}(\pi_\theta)\right\|_1 \cdot \|r\|_\infty \cdot \left\|Xy\right\|_2^2 + 4 \cdot \|r\|_\infty \cdot \left\|Xy\right\|_2^2 \tag{170}$$

$$\le \left(\frac{1}{2} + 4\right) \cdot \|r\|_\infty \cdot \left\|Xy\right\|_2^2 \tag{171}$$

$$\le \frac{9}{2} \cdot \|r\|_\infty \cdot \|X\|_{\text{op}}^2 \cdot \|y\|_2^2 \tag{172}$$

$$= \frac{9}{2} \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X) \cdot \|y\|_2^2, \tag{173}$$

where $\|X\|_{\text{op}}$ is the operator norm of $X \in \mathbb{R}^{K \times d}$ (squared root of largest eigenvalue of $X^\top X$),

$$\|X\|_{\text{op}} = \sup\left\{\|Xv\|_2 : \|v\|_2 \le 1,\ v \in \mathbb{R}^d\right\}. \tag{174}$$

According to Taylor's theorem, for all $\theta,\ \theta' \in \mathbb{R}^d$, there exists $\theta_\zeta := \zeta \cdot \theta + (1-\zeta) \cdot \theta'$ with $\zeta \in [0,1]$, such that,

$$\left|(\pi_{\theta'} - \pi_\theta)^\top r - \left\langle \frac{d\pi_\theta^\top r}{d\theta}, \theta' - \theta \right\rangle\right| = \frac{1}{2} \cdot \left|(\theta' - \theta)^\top S(X, r, \theta_\zeta)\, (\theta' - \theta)\right| \tag{175}$$

$$\le \frac{9}{4} \cdot \|r\|_\infty \cdot \lambda_{\max}(X^\top X) \cdot \|\theta' - \theta\|_2^2. \qquad \square$$

Lemma 4 (Alternative expression of co-variance)**.** *Given any vectors $x \in \mathbb{R}^K$, $y \in \mathbb{R}^K$, we have, for all policy $\pi \in \Delta(K)$,*

$$\operatorname{Cov}_\pi(x, y) = \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot (x(i) - x(j)) \cdot (y(i) - y(j)). \tag{176}$$

Proof. Note that, $\operatorname{Cov}_\pi(x, y) = x^\top \left(\operatorname{diag}(\pi) - \pi \pi^\top \right) y$. Next, we have,

$$x^\top \left(\operatorname{diag}(\pi) - \pi \pi^\top \right) y = \sum_{i=1}^{K} \pi(i) \cdot x(i) \cdot y(i) - \sum_{i=1}^{K} \pi(i) \cdot y(i) \cdot \sum_{j=1}^{K} \pi(j) \cdot x(j) \tag{177}$$

$$= \sum_{i=1}^{K} \pi(i) \cdot x(i) \cdot y(i) - \sum_{i=1}^{K} \pi(i)^2 \cdot x(i) \cdot y(i) - \sum_{i=1}^{K} \pi(i) \cdot y(i) \cdot \sum_{j \not= i} \pi(j) \cdot x(j) \tag{178}$$

$$= \sum_{i=1}^{K} \pi(i) \cdot x(i) \cdot y(i) \cdot (1 - \pi(i)) - \sum_{i=1}^{K} \pi(i) \cdot y(i) \cdot \sum_{j \not= i} \pi(j) \cdot x(j) \tag{179}$$

$$= \sum_{i=1}^{K} \pi(i) \cdot x(i) \cdot y(i) \cdot \sum_{j \not= i} \pi(j) - \sum_{i=1}^{K} \pi(i) \cdot y(i) \cdot \sum_{j \not= i} \pi(j) \cdot x(j) \tag{180}$$

$$= \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot (x(i) \cdot y(i) + x(j) \cdot y(j)) - \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot (x(j) \cdot y(i) + x(i) \cdot y(j)) \tag{181}$$

$$= \sum_{i=1}^{K-1} \pi(i) \cdot \sum_{j=i+1}^{K} \pi(j) \cdot (x(i) - x(j)) \cdot (y(i) - y(j)), \tag{182}$$

finishing the proofs. □

C Generalization to MDPs

We discuss some research plans for generalizing the results to MDPs, considering Softmax PG for illustration. The discussion provides some new ideas, but resolving this problem is highly non-trivial and requires further investigation. We omit the introduction of notations for general finite MDPs.

According to the policy gradient theorem [25, Theorem 1], we have, for all $\theta \in \mathbb{R}^d$,

$$\theta_{t+1} = \theta_t + \eta \cdot \sum_{s \in \mathcal{S}} d^{\pi_{\theta_t}}(s) \cdot \sum_{a \in \mathcal{A}} \frac{\partial\, \pi_{\theta_t}(a|s)}{\partial\, \theta_t} \cdot Q^{\pi_{\theta_t}}(s, a) \tag{183}$$

$$= \theta_t + \eta \cdot \sum_{s \in \mathcal{S}} d^{\pi_{\theta_t}}(s) \cdot X_s^\top \left(\operatorname{diag}(\pi_{\theta_t}(\cdot|s)) - \pi_{\theta_t}(\cdot|s) \pi_{\theta_t}(\cdot|s)^\top \right) Q^{\pi_{\theta_t}}(s, \cdot), \tag{184}$$

where $X_s \in \mathbb{R}^{|\mathcal{A}| \times d}$ is the feature matrix under state $s \in \mathcal{S}$ and can be shared across multiple states. Comparing with Eq. (4), for all $s \in \mathcal{S}$, the reward vector $r \in \mathbb{R}^K$ is replaced with $Q^{\pi_{\theta_t}}(s, \cdot) \in \mathbb{R}^{|\mathcal{A}|}$, which provides some new ideas as well as difficulties.

The idea is that preserving the order of $Q^*(s, \cdot)$ (value of the optimal policy π^* under state $s \in \mathcal{S}$) might be enough to achieve global convergence. Suppose that there exists $w \in \mathbb{R}^d$, such that for all $s \in \mathcal{S}$, $X_s w \in \mathbb{R}^{|\mathcal{A}|}$ preserves the order of $Q^*(s, \cdot)$. If $\operatorname{softmax}(X_s \theta_t)$ is close enough to $\pi^*(\cdot|s)$, such that $Q^{\pi_{\theta_t}}(s, \cdot)$ preserves the order of $Q^*(s, \cdot)$, then we have,

$$\theta_{t+1}^\top w = \theta_t^\top w + \eta \cdot \sum_{s \in \mathcal{S}} d^{\pi_{\theta_t}}(s) \cdot w^\top X_s^\top \left(\operatorname{diag}(\pi_{\theta_t}(\cdot|s)) - \pi_{\theta_t}(\cdot|s) \pi_{\theta_t}(\cdot|s)^\top \right) Q^{\pi_{\theta_t}}(s, \cdot) \tag{185}$$

$$\geq \theta_t^\top w, \tag{186}$$

which is similar to and generalizes Eq. (12). If one can show that θ_t approaches w in direction, then $\pi_{\theta_t}(a^*(s)|s) = \operatorname{softmax}(X_s \theta_t)(a^*(s)) \to \pi^*(a^*(s)|s) = 1$. This means that preserving the order of $Q^*(s, \cdot)$ could be enough for π^* to be a local attractor for Softmax PG updates. One challenge is to generalize the arguments for arbitrary initialization $\theta_1 \in \mathbb{R}^d$ rather than θ_t being close enough to optimal solution, and the difficulty is that $Q^{\pi_{\theta_t}}(s, \cdot)$ does not necessarily preserve the order of $Q^*(s, \cdot)$, and the above inequality does not necessarily hold.