Francisco S. Melo
Fei Fang (Eds.)

LNAI 13441

Autonomous Agents and Multiagent Systems

Best and Visionary Papers

AAMAS 2022 Workshops
Virtual Event, May 9–13, 2022
Revised Selected Papers

Lecture Notes in Artificial Intelligence 13441

Subseries of Lecture Notes in Computer Science

TOPS: Transition-Based Volatility-Reduced Policy Search

Liangliang Xu[1], Daoming Lyu[1], Yangchen Pan[2], Aiwen Jiang[3], and Bo Liu[1(✉)]

[1] Auburn University, Auburn, AL, USA
{lxz0014,daoming.lyu,boliu}@auburn.edu
[2] University of Alberta, Edmonton, AB, Canada
pan6@ualberta.ca
[3] Jiangxi Normal University, Nanchang, Jiangxi, China
jiangaiwen@jxnu.edu.cn

Abstract. Existing risk-averse reinforcement learning approaches still face several challenges, including the lack of global optimality guarantee and the necessity of learning from long-term consecutive trajectories. Long-term consecutive trajectories are prone to involving visiting hazardous states, which is a major concern in the risk-averse setting. This paper proposes **T**ransition-based v**O**latility-controlled **P**olicy **S**earch (TOPS), a novel algorithm that solves risk-averse problems by learning from transitions. We prove that our algorithm—under the over-parameterized neural network regime—finds a globally optimal policy at a sublinear rate with proximal policy optimization and natural policy gradient. The convergence rate is comparable to the state-of-the-art risk-neutral policy-search methods. The algorithm is evaluated on challenging Mujoco robot simulation tasks under the mean-variance evaluation metric. Both theoretical analysis and experimental results demonstrate a state-of-the-art level of TOPS' performance among existing risk-averse policy search methods.

Keywords: Reinforcement learning · Risk control · Volatility control

1 Introduction

The world has witnessed the successes of reinforcement learning (RL, [46]) in multiple fields and domains [36]. However, there are still three concerns with existing RL approaches, which are *risk, long-term shocks*, and *global optimality*. The first concern, *risk*, refers to the instability with respect to the uncertainty of future outcomes [13], often measured by the variance of the future outcome (e.g., expected cumulative rewards). Most RL settings are risk-neutral [36,50, 53], meaning that an agent's goal is merely to learn to maximize the expected return (cumulative rewards) without considering the variance. Controlling risk is necessary in a variety of applications, including financial decision-making [27], healthcare [39], and robotics [32].

F. S. Melo and F. Fang (Eds.): AAMAS 2022 Workshops, LNAI 13441, pp. 3–47, 2022.
https://doi.org/10.1007/978-3-031-20179-0_1

The second concern, *long-term shocks*, is about visiting fatal or hazardous states (i.e., states with extremely low future outcomes) in the process of long-term interactions with the environment [18,20]. Unfortunately, avoiding hazardous state visitation is not always guaranteed for risk-averse RL. A key observation is that visiting hazardous states are often caused by the agent's long-term consecutive interactions with the environment [7,24,48]. Long-term consecutive interactions with the environment tend to generate trajectories with hazardous state visitations [7]. The possibility of triggering hazardous state visitations would be significantly reduced if the agent does not learn from long-term trajectories. However, most existing risk-averse RL algorithms require learning from long-term trajectories. Otherwise, one has to use an additional learning rate for the bootstrap-based critic learning, resulting in a multi-timescale step-size tuning scheme, which is quite inconvenient in practice.

The third concern regards global optimality. The theoretical understanding of policy gradient methods is under tentative study. Work on this topic has been done mostly in the tabular setting. [11] and [44] establish non-asymptotic convergence guarantees for various policy gradient methods with regularization. [35] show convergence rate for softmax parametrization. [1] analyze multiple policy gradient methods in the tabular setting as well as the linear approximation setting. [28,61] extend their work to an off-policy setting. A large spectrum of work has been done on the global optimality of policy gradient methods in a non-linear approximation setting with over-parameterized neural networks [31, 51,62].

In this paper, we aim to answer one question: *Can risk-aware policy gradient algorithms have global optimality convergence guarantee and learn safely without the need for long-term trajectories?* Motivated by addressing this question, we propose **T**ransition-based v**O**latility-controlled **P**olicy **S**earch (TOPS), a risk-averse RL framework with reward volatility [7] as its risk measurement and establish its global convergence and optimality. This paper makes two major contributions. *First*, instead of learning from long-term rollouts [55,57,62], our method TOPS does not require learning from long-term, uninterrupted trajectories. Instead, it can be either trained with segments of long-term rollouts, short-term trajectories, or a combination of them. This is achieved by using a lower-bound surrogate loss mean-volatility loss function (other than the original mean-variance loss function as in [14]) inspired by [7,60]. *Second*, we present a theoretical analysis of the global optimality of the proposed algorithm and prove that TOPS converges to a globally optimal policy at the rate of $1/\sqrt{K}$, where K is the number of iterations. This is achieved by the primal-dual formulation of the mean-volatility function used in [60] and the primal-dual sample complexity analysis inspired by [57,62].

The roadmap of the paper is as follows. We introduce the background in Sect. 2. In Sect. 3, we formulate the TOPS algorithm. We present the major result on its global convergence in Sect. 4. In Sect. 5, we perform experiments on benchmark domains and compare them with state-of-the-art methods. We discuss related work in Sect. 6 and conclude the paper in Sect. 7.

2 Background

This section introduces the background knowledge of the building blocks of this paper, such as reinforcement learning, policy gradient, over-parameterized neural networks. A detailed notation system is provided in Appendix A.

Reinforcement Learning. We consider the infinite-horizon discounted Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r(s,a), \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, the transition kernel $\mathcal{P} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$, the reward function $r(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the initial state $S_0 \in \mathcal{S}$ and its distribution $\mu_0 : \mathcal{S} \to [0,1]$, and the discounted factor $\gamma \to (0,1)$. At time step t, given a state s_t, an action a_t is taken according to policy $\pi(a_t|s_t) : \mathcal{S} \times \mathcal{A} \to [0,1]$, generating a reward $r_t := r(s_t, a_t)$ and the next state s_{t+1} based on $p(s_{t+1}|s_t, a_t)$ the reward function is assumed to be deterministic and bounded—a constant $r_{\max} > 0$ exists such that $r_{\max} = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}|r(s,a)|$. A change in states upon an action $(s, a, r(s,a), s')$ is termed a *transition*, where the state s' is the successive state of the state s. With a little bit abuse of notation, $r(s,a)$ is denoted as $r_{s,a}$, and $r(s_t, a_t)$ is denoted as r_t in the rest of the paper. A *trajectory* of length T is a consecutive sequence of transitions $\{(s_t, a_t, r_t, s'_t)\}_{t=0}^{T-1}$ over a set of contiguous timestamps, where $\forall t, s'_t = s_{t+1}$. Therefore, the trajectory is also equivalently denoted by $\{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{T-1}$. To evaluate the performance of policy π, we introduce state value function $V_\pi(s) := (1-\gamma)\mathbb{E}_{a\sim\pi(a|s)}\big[\sum_{t=0}^{\infty}\gamma^t r_t \big| S_0 = s, a_t \sim \pi(a|s_t), s_{t+1} \sim \mathcal{P}(s|s_t, a_t)\big]$ and state-action value function $Q_\pi(s,a) := (1-\gamma)\mathbb{E}_{a\sim\pi(a|s)}\big[\sum_{t=0}^{\infty}\gamma^t r_t \big| s_0 = s, a_0 = a, a_t \sim \pi(a|s_t), s_{t+1} \sim \mathcal{P}(s|s_t, a_t)\big]$. Bounded reward implies $|V_\pi(s)| \le r_{\max}$ and $|Q_\pi(s,a)| \le r_{\max}$ $\forall\pi$. Additionally, the advantage function $A_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of policy π is defined as $A_\pi(s,a) := Q_\pi(s,a) - V_\pi(s)$. The normalized state and state action occupancy measure of policy π is denoted by $\nu_\pi(s)$ and $\sigma_\pi(s,a) := \pi(a|s)\nu_\pi(s)$, respectively. Therefore, $\nu_\pi(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr(s_t = s|\mu_0, \pi, \mathcal{P})$ and $\sigma_\pi(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr(s_t = s, a_t = a|\mu_0, \pi, \mathcal{P})$, where $Pr(s_t = s|\mu_0, \pi, \mathcal{P})$ is the probability of $s_t = s$ given $\mu_0, \pi, \mathcal{P}$. Finally, the return is defined as $G := \sum_{t=0}^{\infty}\gamma^t r_t$.

Policy Gradient Methods. In the following, we discuss two policy gradient methods, where the policy π_θ is parameterized by the parameter θ. For natural policy gradient (NPG, [22]), we first define the Fisher information matrix,

$$F(\theta) := \mathbb{E}_{(s,a)\sim\sigma_{\pi_\theta}}\big[\nabla_\theta \log(\pi_\theta)(\nabla_\theta \log(\pi_\theta))^\top\big] \tag{1}$$

The update of parameter θ then takes the form,

$$\theta_{k+1} = \theta_k + \eta_{\mathrm{NPG}}\big(F(\theta_k)\big)^{-1}\nabla J_\theta(\pi_{\theta_k}) \tag{2}$$

where $\big(F(\theta_{k-1})\big)^{-1}$ is the inverse of the Fisher information matrix $F(\theta)$ in Eq. (1), ∇J_θ is the objective gradient and η_{NPG} the learning rate.

In proximal policy optimization (PPO, [43]), at the k-th iteration the update of policy parameter θ takes the following form, where β is the penalty hyperparameter:

$$\arg\max_\theta \mathbb{E}_{(s,a)\sim\sigma_k}\big[\pi_\theta A_{\theta_k/\pi_{\theta_k}} - \beta\mathrm{KL}(\pi_\theta\|\pi_{\theta_k})\big]. \tag{3}$$

Policy Network with Over-Parameterized Neural Networks. Policy π with the two-layer over-parameterized neural network is defined as: for $\forall(s,a) \in S \times A$,

$$f\big((s,a);\theta,b\big) := \frac{1}{\sqrt{m}} \sum_{v=1}^{m} b_v \texttt{ReLU}\big((s,a)^\top [\theta]_v\big). \tag{4}$$

Here (s,a) is the input and m is the width of the network. $\theta = ([\theta]_1^\top, \cdots, [\theta]_m^\top)^\top \in \mathbb{R}^{m\times d}$ is the input weight matrix in the first layer of the neural network. $b = (b_1, \cdots, b_m)^\top \in \mathbb{R}^{m\times 1}$ are the output weights in the second layer. We present a block diagram of a over-parameterized neural network with Fig. 4 in the Appendix B. At the start of training, the parameters θ, b are initialized by

$$\theta = \Theta_{\text{init}} \in \mathbb{R}^{m\times d}([\Theta_{\text{init}}]_v \sim \mathcal{N}(0, I_d/d), \tag{5}$$

and $b_v \sim \text{Unif}\big(\{-1,1\}\big), \forall v \in [m]$, respectively, where $\mathcal{N}$ denotes Gaussian distribution and Unif denotes uniform distribution. $f((s,a);\theta,b)$ can be simplified to $f((s,a);\theta)$ by updating only $[W]_v$ during training, and fixing b as its initialization [3]. We also restrict the possible value of θ within the space denoted by $\mathcal{D} = \{\xi \in \mathbb{R}^{md} : \|\xi - \Theta_{\text{init}}\|_2 \le \Upsilon, \Upsilon > 0\}$. Therefore the policy is defined $\pi_\theta(a|s)$ in the following form, where τ is the temperature parameter.

$$\pi_\theta(a|s) := \frac{\exp\big(\tau f\big((s,a);\theta\big)\big)}{\sum_{a'\in A} \exp\big(\tau f\big((s,a');\theta\big)\big)}.$$

Furthermore, the feature mapping of a two-layer neural network $f((s,a);\theta)$ is defined as, $\phi_\theta := \big([\phi_\theta]_1^\top, \cdots, [\phi_\theta]_m^\top\big)^\top$, where $[\phi_\theta]_v^\top = \frac{b_v}{\sqrt{m}} \texttt{ReLU}((s,a)^\top [\theta]_v)$, $\forall v \in [m]$. By Eq. (4), it holds that $f((s,a);\theta) = \phi(s,a)^\top \theta$ and $\nabla_\theta f((s,a),\theta) = \phi(s,a)$ [51]. Furthermore, we assume that there exists a constant $M > 0$ such that,

$$\mathbb{E}_{(s,a)\sim \text{init}}\Big[\sup_{(s,a)\in S\times A} \big|\phi((s,a)^\top \Theta_{\text{init}})\big|^2\Big] \le M^2.$$

Mean-Variance and Mean-Volatility RL. In a variance-constraint problem with the variance of the total reward, the objective can be formulated as,

$$\max_{\pi} J(\pi), \qquad \text{subject to } \mathbb{V}(G) \le Y \tag{6}$$

where $J(\pi) := \mathbb{E}_{(s,a)\sim\sigma_\pi}[G] = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\sigma_\pi}[r_{s,a}]$ is the expected return, $\mathbb{V}(\cdot)$ is the variance of a random variable and $Y > 0$ is the upper bound for this variance. The constrained formulation in Eq. (6) is NP-hard [45], and in reality, the relaxed formulation $J_\lambda^G(\pi)$ defined in Eq. (7) is often solved instead [14,26,55] as follows, where λ is called variance-controlling parameter.

$$J_\lambda^G(\pi) := \mathbb{E}[G] - \lambda\mathbb{V}(G) = \mathbb{E}[G] - \lambda\mathbb{E}[G^2] + \lambda(\mathbb{E}[G])^2 \tag{7}$$

Meanwhile, [7] proposed a reward-volatility risk measure. *Volatility* is defined as the variance of per-step reward—per-step reward R is a discrete random variable

with a probability mass function of $p(R = x) = \sum_{s,a} \sigma_\pi(s,a)\mathbb{1}\{r_{s,a} = x\}$, where $\mathbb{1}\{\cdot\}$ is the indicator function. It is easy to see that $\mathbb{E}[R] = (1-\gamma)J(\pi)$ [60]. $\mathbb{V}(R)$ is the variance of R. Likewise, $J_\lambda(\pi)$ is proposed as a counterpart of Eq. (7) in the sequel, which is defined with respect to R.

$$J_\lambda(\pi) := \mathbb{E}[R] - \lambda\mathbb{V}(R) = \mathbb{E}[R] - \lambda\mathbb{E}[R^2] + \lambda(\mathbb{E}[R])^2$$

We first present the following lemma based on Lemma 1 in [7] to show that $J_\lambda(\pi)$ is a reasonable counterpart to $J_\lambda^G(\pi)$.

Lemma 1. *Given* $\lambda \geq 0$, $\frac{1}{(1-\gamma)}J_{\frac{\lambda}{(1-\gamma)}}(\pi)$ *is a lower-bound of* $J_\lambda^G(\pi)$, *i.e.,* $\frac{1}{(1-\gamma)}J_{\frac{\lambda}{(1-\gamma)}}(\pi) \leq J_\lambda^G(\pi)$.

A detailed proof is provided in Appendix D.5. Given Lemma 1, maximizing $J_\lambda^G(\pi)$ can be reduced to maximizing its lower bound $\frac{1}{(1-\gamma)}J_{\frac{\lambda}{(1-\gamma)}}(\pi)$. There are several advantages of optimizing $J_\lambda(\pi)$. Compared to optimizing $\mathbb{V}(G)$, optimizing $\mathbb{V}(R)$ is computationally easier [60]. [7] argue that $\mathbb{V}(R)$ is better at capturing short-term risk and leads to smoother trajectories that avoid possible "shocks" caused by long-horizon trajectories [7].

3 Algorithm Formulation

In this section, we present our risk-averse policy-search algorithm. In particular, we use **(i)** reward volatility to construct the mean-volatility objective, which circumvents the long-horizon reward issue and avoids large variance, and **(ii)** over-parameterized neural network [10] as the neural network architecture of the actor and the critic to facilitate global convergence analysis.

3.1 Augmented MDP

As [7] shows, reward volatility has advantages over mean-variance methods, including a smoother trajectory and much-reduced variance. Therefore, in our paper, we choose volatility as the risk measurement. Compared with the conventional mean-variance objective, the mean-volatility objective function enables the agent to learn from transitions instead of trajectories and greatly reduces the chance of getting into hazardous states due to consecutive long-horizon explorations. This can greatly help improve safety. Note that because of *compositional* expectations $(\mathbb{E}[R])^2$, double-sampling is needed, which is a heavy burden for sampling.[1]

[1] For more details on double-sampling and the more general compositional expectations, please refer to [30,52].

To avoid double-sampling, we resort to augmented MDP. First, note that it holds $(\mathbb{E}[R])^2 = \max_{y\in\mathbb{R}}(2\mathbb{E}[R]y - y^2)$. Then the optimization objective transforms into:

$$\max_{\pi,y} J_\lambda^y(\pi) := \mathbb{E}_{(s,a)\sim\sigma_\pi}(r_{s,a} - \lambda r_{s,a}^2 + 2\lambda r_{s,a} y) - \lambda y^2 \tag{8}$$

We now introduce the augmented MDP, with the augmented reward defined as follows:

$$\tilde{r}_{s,a} := r_{s,a} - \lambda r_{s,a}^2 + 2\lambda r_{s,a} y \tag{9}$$

We refer to this new MDP as the *augmented MDP* $\tilde{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \tilde{r}(s,a), \gamma\}$, and denote corresponding terms by the ˜ sign—for example, the associated state value function and state-action value function are $\tilde{V}_\pi(s)$ and $\tilde{Q}_\pi(s,a)$. $\tilde{r}(s,a)$ is denoted by $\tilde{r}_{s,a}$) for notation simplicity in the remainder of this paper. We solve Eq. (8) by maximizing y and π of the augmented MDP in two steps iteratively.

3.2 Proposed Algorithms

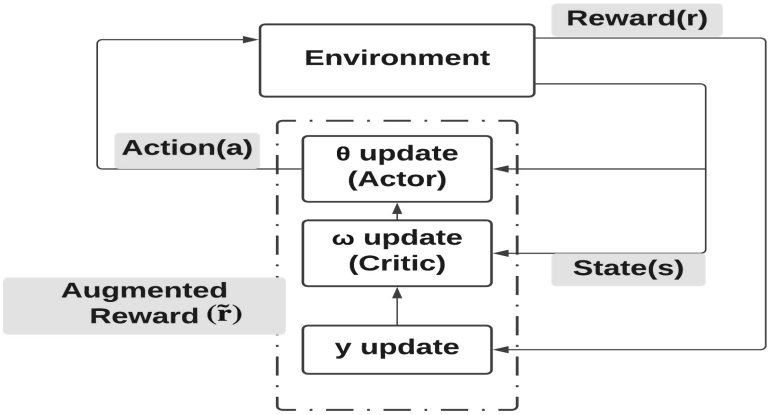

Fig. 1. A simple block diagram of TOPS

We present TOPS in Algorithm 1. A block diagram of TOPS is illustrated in Fig. 1, where there are three sets of parameters to update, i.e., θ for the actor, ω for the critic, and the auxiliary variable y. Note that the mean-volatility framework allows incorporating any off-the-shelf policy optimization methods as pointed out by [60]. Since most global optimality analysis literature is based on NPG and PPO [9,31,51,62], we also use NPG and PPO as inner policy search algorithms for a fair comparison.

y Update: Since Eq. (9) is quadratic in y, to update y in each iteration, we have $y_k = (1-\gamma)J(\pi_k)$. However, we do not have direct access to the exact value of

$J(\pi_k)$. As an alternative, we estimate this value with a sample average in the k-th iteration,

$$\hat{y}_k := \frac{1}{T}\sum_{t=1}^{T} r_t, \tag{10}$$

as an estimator of y_k, where T is the sample (batch) size.

θ Update: We update π_θ with an actor-critic scheme, particularly NPG and PPO. NPG and PPO are the two most widely used policy gradient methods. According to empirical studies [54], PPO usually achieves state-of-the-art performance among on-policy policy gradient methods, and NPG has the advantage of easy hyperparameter tuning compared with PPO. On the other hand, NPG and PPO's global convergences under the risk-neutral setting have been studied intensively and can show good results [1,51]. Therefore, PPO and PNG are used as the inner actor algorithm of the TOPS framework to make fair comparisons with existing approaches. Additionally, over-parameterized neural networks are widely used in proving global convergence of gradient-based methods under the risk-neutral setting, which show impressive results [17,31,51]. The capability of a gradient-based neural network method to reach the global optimum in an over-parameterization setting is explained in theory [10]. Therefore, we parameterize the policy in the paper with a two-layer over-parameterized neural network. We first introduce the actor part of the two methods, respectively.

θ Update for Neural NPG. Per the update rule for NPG in Eq. (2), we need to estimate the natural policy gradient $\big(F(\theta_k)\big)^{-1}\nabla_\theta J(\pi_\theta)$. However $F(\theta_k)$ is difficult to invert due to its high-dimensionality. Instead the gradient is estimated by solving $\min_{\xi\in\mathcal{D}} \|\hat{F}(\theta_k)\xi - \tau_k\hat{\nabla}_\theta J(\pi_{\theta_k})\|_2$, where

$$\begin{aligned}\hat{\nabla}_\theta J(\pi_{\theta_k}) &:= \frac{\tau_k}{T}\sum_{t=1}^{T} Q_{\omega_k}(s_t,a_t)\big(\phi_{\theta_k}(s_t,a_t)\\ &\quad - \mathbb{E}_{a\sim\pi_{\theta_k}}[\phi_{\theta_k}(s_t,a_t')]\big),\\ \hat{F}(\theta_k) &:= \frac{\tau_k^2}{T}\sum_{t=1}^{T}\Big(\big(\phi_{\theta_k}(s_t,a_t) - \mathbb{E}_{a\sim\pi_{\theta_k}}[\phi_{\theta_k}(s_t,a_t')]\big)\\ &\quad \big(\phi_{\theta_k}(s_t,a_t) - \mathbb{E}_{a\sim\pi_{\theta_k}}[\phi_{\theta_t}(s_t,a_t')]\big)^\top\Big),\end{aligned}$$

are unbias estimations of $\nabla_\theta J(\pi_\theta)$ and $F(\theta_k)$ respectively, with the help of feature mapping, θ is, therefore, updated as,

$$\begin{aligned}\tau_{k+1} &= \tau_k + \eta_{\text{NPG}},\\ \theta_{k+1} &= \big(\tau_k\theta_k + \eta_{\text{NPG}}\arg\min_{\xi\in\mathcal{D}}\|\hat{F}(\theta_k)\xi - \tau_k\hat{\nabla}_\theta J(\pi_{\theta_k})\|_2\big)/\tau_{k+1}\end{aligned} \tag{11}$$

θ Update for Neural PPO. Given the update rule of PPO in Eq. (3), the PPO's objective function $L(\theta)$ can be rewritten as $L(\theta) := \mathbb{E}_{s\sim\nu_{\pi_k}}\big[\mathbb{E}_{a\sim\pi_\theta}[Q_{\pi_k}] -$

$\beta \mathrm{KL}(\pi_\theta \| \pi_{\theta_k})\big]$. With energy-based policy $\pi \propto \exp\{\tau^{-1} f\}$, the solution to the subproblem $\hat{\pi}_{k+1} = \arg\max_\pi L(\theta)$ can be obtained by solving the following [31]:

$$\theta_{k+1} = \arg\min_{\theta \in \mathcal{D}} \mathbb{E}_{(s,a)\sim\sigma}\big[\big(f_\theta - \tau_{k+1}(\beta^{-1} Q_{\pi_k} + \tau_k^{-1} f_{\theta_k})\big)^2\big] \tag{12}$$

The stochastic gradient method can be used to solve Eq. (12).

ω Update. To estimate the state-action function value of the augmented MDP $\tilde{Q}_\pi$, a critic network parameterized by ω is constructed, denoted as $\tilde{Q}_\omega$. Note that the critic uses the same two-layer neural network architecture as the actor defined in Eq. (4), which indicates that the policy network π_θ's parameter θ and critic network $\tilde{Q}_\pi$'s parameter ω are of identical dimensions, i.e., $\theta \in \mathbb{R}^d, \omega \in \mathbb{R}^d$. The critic network is parameterized with a different set of parameters $\omega = ([\omega]_1^\top, \cdots, [\omega]_m^\top)^\top \in \mathbb{R}^{md}$, denoted by $f((s,a);\omega)$. For simplicity, we then learn $\tilde{Q}_\omega$ by applying the semi-gradient TD method. Other approaches, such as the Gradient TD (GTD) algorithm family [47], can also be applied. For each iteration t of the TD update,

$$\begin{aligned} \omega_{t+1} = \omega_t - \eta_{\mathrm{TD}}\big(&\tilde{Q}_{\omega_t}(s,a) \\ &- (1-\gamma)\tilde{r}_{s,a} - \gamma \tilde{Q}_{\omega_t}(s',a')\big)\nabla_\omega Q_{\omega_t}(s,a), \end{aligned} \tag{13}$$

where η_{TD} is the learning rate for TD update.

4 Theoretical Analysis

Although NPG and PPO's global convergences under the risk-neutral setting show prominent result [1,51], the techniques used by these methods only apply to the primal constrained-MDP case and remain challenging to apply to the analysis of the primal-dual case as in our augmented MDP, where the dual variable y is critical. For example, Lemma (5.2) of [31], a critical step in the error-bound analysis of risk-neutral PPO, cannot be applied to our primal-dual risk-averse case. In this section, we establish the global convergence rate of TOPS with both NPG and PPO.

4.1 Assumptions

We first impose regularity condition assumptions, which are common in the literature on TD analysis with a neural network approximation [9,31,51,62].

Assumption 1 *(Variance upper bound)* [51]. *Let $\mathcal{D} = \{\alpha \in \mathbb{R}^{md} : \|\alpha - \Theta_{\mathrm{init}}\|_2 \le \Upsilon\}$. For all $k \in [K]$, We assume that for all $k \in [K]$, there exists an absolute constant $\sigma_\xi > 0$ such that,*

$$\mathbb{E}[\|\xi_k(\delta_k)\|_2^2] \le \tau_k^4 \sigma_\xi^2 / T, \quad \mathbb{E}[\|\xi_k(\omega_k)\|_2^2] \le \tau_k^4 \sigma_\xi^2 / T.$$

where $\delta_k = argmin_{\delta\in\mathcal{D}} \|\hat{F}(\theta_k)\delta - \tau_k \hat{\nabla}_\theta J(\pi_{\theta_k})\|_2$ and $\xi_k(\delta) = \hat{F}(\theta_k)\delta - \tau_k \hat{\nabla}_\theta \tilde{J}(\pi_{\theta_k}) - \mathbb{E}[\hat{F}(\theta_k)\delta - \tau_k \hat{\nabla}_\theta \tilde{J}(\pi_{\theta_k})]$. The expectation is taken over σ given θ_k and ω_k.

Algorithm 1: TOPS: Transition-based VOlatility-controlled Policy Search

Input: number of iteration K, learning rate for natural policy gradient (resp. PPO) and neural TD η_{NPG} (resp. η_{PPO}), temperature parameters $\{\tau_k\}_{k=1}^K$;
Initialize policy network $f((s,a);\theta,b)$ as defined in Eq. (5). Set $\tau_1 = 1$. Initialize Q-network with (b,ω_1) similarly;
for $k = 1, \cdots, K$ **do**
  Sample a batch of transitions $\{s_t, a_t, r_t, s_t'\}_{t=1}^T$ following current policy with size of T;
  $y = \frac{1}{T}\sum_{t=1}^T r_t$;
  for $t = 1, \cdots, T$ **do**
    $\tilde{r}_t = r_t - \lambda r_t^2 + 2\lambda r_t y$, $a_t' \sim \pi(a|s_t')$;
  end
  Q-value update: update ω_k according to Eq. (13);
  if *select NPG update* **then**
    update θ_k according to Eq. (11) ;
  else if *select PPO update* **then**
    update θ_k according to Eq. (12) ;
end
Output: π_{θ_K};

Note that δ_k and ω_k have the same dimension due to the compatible neural network setting. $\xi_k(\delta)$ can be generalized to both δ_k and ω_k. We then introduce a regularity condition assumption on visitation measures and stationary distributions in the sequel, respectively.

Assumption 2 *(Upper bounded concentrability coefficient)* [51]. *ν^* and σ^* are denoted as the state and state-action visitation measures corresponding to the global optimum π^*. For all $k \in [K]$, we define the following terms:*

$$\varphi_k = \left\{\mathbb{E}_{(s,a)\sim\sigma_{\pi_k}}[(\frac{d\sigma^*}{d(\sigma_{\pi_k})})^2]\right\}^{1/2}, \ \psi_k = \left\{\mathbb{E}_{s\sim\nu_{\pi_k}}[(\frac{d\nu^*}{d(\nu_{\pi_k})})^2]\right\}^{1/2},$$
$$\varphi_k' = \left\{\mathbb{E}_{(s,a)\sim\sigma'_{\pi_k}}[(\frac{d\sigma^*}{d(\sigma'_{\pi_k})})^2]\right\}^{1/2}, \ \psi_k' = \left\{\mathbb{E}_{s\sim\nu'_{\pi_k}}[(\frac{d\nu^*}{d(\nu'_{\pi_k})})^2]\right\}^{1/2}.$$

We assume that $\varphi_k, \psi_k, \varphi_k', \psi_k'$ are uniformly upper bounded by an absolute constant $c_0 > 0$.

σ'_{π_k} and ν'_{π_k} are state-action and state stationary distribution. $\varphi_k, \psi_k, \varphi_k', \psi_k'$ are the concentrability coefficients, which reflects how much the starting state and state-action distribution diverge from the state and state-action distribution under the optimal policy [37]. Assumption 2 impose a upper bound on such divergence. This regularity condition is commonly used in the literature [4,

16,38,51,58]. More over, we define $\varphi_k^* = \mathbb{E}_{(s,a)\sim\sigma_\pi}\left[\left(\frac{d\pi^*}{d\pi_0} - \frac{d\pi_{\theta_k}}{d\pi_0}\right)^2\right]^{1/2}, \psi_k^* = \mathbb{E}_{(s,a)\sim\sigma_\pi}\left[\left(\frac{d\sigma_{\pi^*}}{d\sigma_\pi} - \frac{d\nu_{\pi^*}}{d\nu_\pi}\right)^2\right]^{1/2}$.

4.2 Major Theoretical Results

In the following, we present the major theoretical results, i.e., the global optimality and convergence rate of TOPS with neural PPO. We define the optimality gap $\min_{k\in[K]}\left(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\right)$, where π^*, y^* are respectively defined as $\pi^* := \arg\max_\pi J_\lambda(\pi)$, $\hat{y}^* := (1-\gamma)J(\pi^*)$, and y^k is defined in Eq. (10). $J_\lambda^{y^*}(\pi^*)$ (resp. $J_\lambda^{\hat{y}_k}(\pi_k)$) represents the risk-averse objective under π^* (resp.π_k, i.e., the policy at the k-th iteration).

Theorem 1 *(Global Optimality and Rate of Convergence on neural PPO). We set the learning rate of PPO* $\eta_{\text{PPO}} = \min\{(1-\gamma)/3(1+\gamma)^2, 1/\sqrt{K_{\text{TD}}}\}$*, the learning rate of TD update* $\eta_{\text{TD}} = \min\{(1-\gamma)/3(1+\gamma)^2, 1/\sqrt{K_{\text{TD}}}\}$ *where* K_{TD} *is the total iteration of TD update, and* $\beta_0 := \beta/\sqrt{K}$*. Under Assumptions 3–4, we have, with a probability of* $1-\delta$*,*

$$
\begin{aligned}
&\min_{k\in[K]}\left(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\right)\\
\leq\ &\frac{\beta_0^2\log|\mathcal{A}| + U + \beta_0^2\sum_{k=1}^K(\varepsilon_k+\varepsilon_k')}{(1-\gamma)\beta_0\sqrt{K}}\\
&+\frac{4\lambda c_3 r_{\max}(1-\gamma)}{\sqrt{K}}\\
\text{where } \varepsilon_k'' &= \mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta) + \Upsilon^{5/2}m^{-1/4}\cdot\\
&\quad\sqrt{\log(1/\delta)} + \Upsilon\cdot r_{\max}^2 m^{-1/4} + \Upsilon^2 K_{\text{TD}}^{-1/2} + \Upsilon),\\
\varepsilon_k &= \tau_{k+1}^{-1}\varepsilon_k''\varphi_{k+1}^* + \beta^{-1}\varepsilon_k''\psi_k^*,\\
\varepsilon_k' &= |\mathcal{A}|\tau_{k+1}^{-2}\epsilon_{k+1}^2,\\
U &= 2\mathbb{E}_{s\sim\nu_{\pi^*}}[\max_{a\in\mathcal{A}}(\tilde{Q}_{\omega_0})^2] + 2\Upsilon^2
\end{aligned}
$$

Similarly, we present TOPS global optimality and convergence rate with neural NPG.

Theorem 2 *(Global Optimality and Rate of Convergence for neural NPG). We set the learning rate of NPG* $\eta_{\text{NPG}} = 1/\sqrt{K}$*, the learning rate of TD update* $\eta_{\text{TD}} = \min\{(1-\gamma)/3(1+\gamma)^2, 1/\sqrt{K_{\text{TD}}}\}$ *where* K_{TD} *is the total iteration of TD update, and the temperature parameters* $\tau_k = (k-1)\eta_{\text{NPG}}$*. Under Assumptions 3–4, with a probability of* $1-\delta$*, we have*

$$\min_{k\in[K]}\left(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\right)$$

$$
\begin{aligned}
&\leq \frac{1}{(1-\gamma)\sqrt{K}}\Big(\log|\mathcal{A}| + 9\Upsilon^2 + M^2 + 4c_3 M(1-\gamma)^2\lambda\Big) \\
&\quad + \frac{1}{K}\sum_{k=1}^{K}\big(\epsilon_k\big) \\
&\quad \textit{where } \epsilon_k = \sqrt{8}c_0\Upsilon^{1/2}\sigma_\xi^{1/2}T^{-1/4} \\
&\quad + \mathcal{O}\big((\tau_{k+1}K^{1/2}+1)\Upsilon^{3/2}m^{-1/4} + \Upsilon^{5/4}m^{-1/8}\big) \\
&\quad + c_0\mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta) + \Upsilon^{5/2}m^{-1/4}\sqrt{\log(1/\delta)} \\
&\quad + \Upsilon r_{\max}^2 m^{-1/4} + \Upsilon^2 K_{\mathrm{TD}}^{-1/2} + \Upsilon)
\end{aligned}
$$

Remark 1. Theorem 1 and 2 show the upper bound of the optimality gap

$$\min_{k\in[K]}\big(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \sim O(\frac{1}{K}),$$

where K is the maximum number of updates. It reflects how close the policy produced by TOPS can achieve to the global optimal policy.

From Theorem 1 and 2, we can conclude that our risk-averse algorithm TOPS with PPO and NPG version of the actor both converge to the global optimal policy at a $\mathcal{O}(1/\sqrt{K})$ rate. We provide a detailed proof in Appendix D.

5 Experiments

In this section, we aim to empirically examine the performance of our algorithm on Mujoco robot manipulation benchmark tasks from OpenAI gym [8], as [60] does. The Mujoco benchmark is a set of challenging robot control tasks in simulated environments designed for controller optimization in reinforcement learning [49]. In this domain, the simulated robots are expected to achieve consistent performances while avoiding failures that lead to dangerous results.

Experiment Setup. We conduct our experiments in an online learning setting and include several recent risk-averse RL methods as baselines: the mean-variance policy optimization (MVP) [55], mean-variance policy iteration (MVPI) [60], and variance-constrained actor-critic (VARAC) [62]. We set $\lambda = 1$ and run each algorithm for 10^6 steps and evaluate the algorithm every 10^4 steps for 20 episodes. All curves are averaged over 10 independent runs and use shaded areas to indicate standard errors. The experiment's details are provided in Appendix C. All experiments' parameters are tuned through rigorous grid search.

We report the learning curves of TOPS with NPG and PPO, respectively, in Fig. 2 and 3. For MVP, since the algorithm uses coordinate gradient at each step, it does not have a PPO or NPG version, and therefore we only report its learning curve in Fig. 2. For MVPI, we use its on-policy version for the experiment. As it works with any off-the-shelf policy search method, we implement NPG and

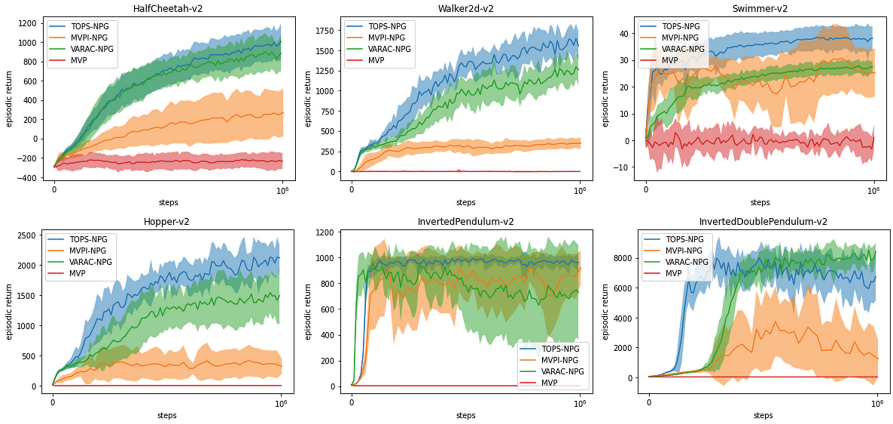

Fig. 2. Training progress of TOPS-NPG and baseline algorithms.

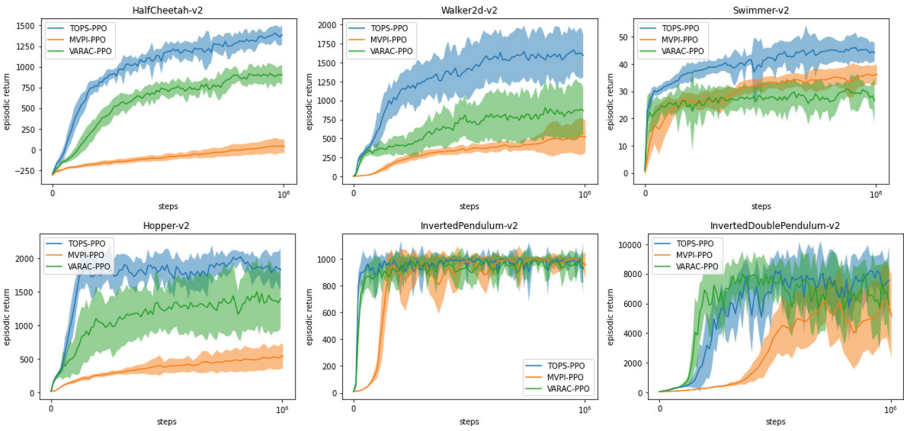

Fig. 3. Training progress of TOPS-PPO and baseline algorithms.

PPO with two-layer over-parameterized neural networks as its policy search component.

There are several interesting findings of our experiments. *First*, the results show that TOPS outperforms other baselines in most of the testbeds with respect to initial learning speed, the variance during the learning process, and the steady-state mean. In particular, TOPS outperforms other methods with a large margin with respect to the mean and variance of the learning curve on `Walker2d-v2`, `Hopper-v2` and `InvertedPendulum-v2`, as seen from Fig. 2 and Fig. 3. *Second*, the partial order of performance level tends to be consistent across different tasks, regardless of the base method (NPG or PPO). The order of performance level (from the best to the worst) in most subfigures of Fig. 2 is TOPS-NPG, VARAC-NPG, and MVPI-NPG. Similar results are observed in the majority

of the subfigures in Fig. 3, where the order is TOPS-PPO, VARAC-PPO, and MVPI-PPO. Compared to other methods, MVP performs poorly in all tested domains, which indicates it may not suit the tasks. Note that on domains such as `Walker2d-v2` and `Hopper-v2`, MVP's curves show zero variance with zero mean. This is due to the fact that these domains have a sparse reward nature (i.e., the reward is 0 for most states), and zero mean and zero reward indicate that MVP simply learns nothing useful. Overall, these results demonstrate that our TOPS algorithm can achieve state-of-the-art risk-averse performance on the challenging robot simulator testbeds.

6 Related Work

In risk-averse RL, variance is a more popular risk measure among its many peers [14,26,33,45,55], with the related approaches usually referred to as mean-variance RL. Variance stands out due to its advantage in interpretability and computation [29,34]. Most mean-variance RL methods consider the variance of the total reward [14,26,55]. In contrast, [7] and [60] propose a reward-volatility risk measure using the variance of a per-step reward. They show that the reward-volatility method is better at capturing the short-term risk and easier to compute.

In particular, we distinguish our method with [7,55], and [60]. [55] utilize the variance of the cumulative rewards. However, this method's theoretical analysis is limited to the sample complexity of episodic average-reward MDP, and the choice of solvers is restricted. Both [7] and [60] use the variance of the per-step reward, which introduces a policy-dependent-reward issue. [7] solve this directly, but it is much more difficult than normal MDP due to the lack of tools, and this approach also requires double-sampling. [60] avoid these issues by proposing an augmented MDP in a flexible framework that can apply any off-the-shelf policy evaluation and control method and name it MVPI. Our paper is inspired by [60], but our approach differs from MVPI. At any iteration, MVPI needs to keep updating π until it maximizes the objective function defined in Eq. (8) with a fixed y, while TOPS is only required to update π once. This key difference enables TOPS to train faster, as shown in Sect. 5. We present an algorithm comparison between TOPS and MVPI in Algorithm 2 of the Appendix B. Furthermore, we can also provide a theoretical analysis of global convergence.

Second, we compare our proof technique with those of [31,51,57,62]. These papers share a similar method in the first part of the analysis, but the second parts are different because each works on a different policy gradient method. Our paper adopts the methods of [51] and [31] for neural NPG and neural PPO, respectively, and develops a method for the Q-function value based on the two. Compared with [51] and [31], our results of convergence are in probability rather than expectation because we utilize different techniques when we characterize certain error bounds. Specifically, we develop a high probability bound for the critic part while using an expectation bound in the actor part. The closest work to ours is [62], which utilizes the variance of the cumulative

rewards and therefore needs to learn from consecutive trajectories instead of non-consecutive transitions. A corresponding disadvantage is that it requires two critics to represent value functions associated with the original reward and the squared reward. On the contrary, TOPS only needs one due to the deployment of the augmented MDP. Additionally, while they present theoretical proofs, they do not provide empirical results. Moreover, in their proof, Eq. (4.15) does not hold, which leads to an invalid core conclusion regarding their Eq. (4.17), an essential part of the proof. Therefore, the validity of the theoretical analysis starting from their Eq. (4.15) remains unclear. The details are mentioned in Appendix D.2. The most recent work along this research line is [57]. Unlike other works, they solve safe reinforcement learning problems in the primal space, termed CRPO. In addition, the theoretical analysis of CRPO is restricted to a simplified version of NPG. In contrast, we present theoretical proof for neural NPG and PPO as the policy search algorithms. Other related work includes [56,63], and [6]. A more detailed discussion is provided in Appendix E.

7 Conclusion

This paper aims to answer the following question: *can risk-averse algorithms have a global convergence guarantee and learn from short trajectories?* Theoretical analysis for both neural NPG and neural PPO with two-layer over-parameterized neural networks are presented to show that TOPS can find the global optimality at an $\mathcal{O}(1/\sqrt{K})$ converge rate. We also demonstrate the empirical success of TOPS in Mujoco robot simulation domains.

Acknowledgment. BL's research is funded by the National Science Foundation (NSF) under grant NSF IIS1910794, Amazon Research Award, and Adobe gift fund.

A Notation Systems

- $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, the transition kernel $\mathcal{P}$, the reward function r, the initial state S_0 and its distribution μ_0, and the discounted factor γ.
- $r_{\max} > 0$ is a constant as the upper bound of the reward.
- State value function $V_\pi(s)$ and state-action value function $Q_\pi(s, a)$.
- The normalized state and state action occupancy measure of policy π is denoted by $\nu_\pi(s)$ and $\sigma_\pi(s, a)$
- T is the length of a *trajectory*.
- The return is defined as G. $J(\pi)$ is the expectation of G.
- Policy π_θ is parameterized by the parameter θ.
- τ is the temperature parameter in the softmax parameterization of the policy.
- $F(\theta)$ is the Fisher information matrix.
- $\eta_T D$ is the learning rate of TD update. Similarly, $\eta_N PG$ is the learning rate of NPG update. $\eta_P PO$ is the learning rate of PPO update.
- β is the penalty factor of KL difference in PPO update.

- $f\big((s,a);\theta\big)$ is the two-layer over-parameterized neural network, with m as its width.
- ϕ_θ is the feature mapping of the neural network.
- $\mathcal{D}$ is the parameter space for θ, with Υ as its radius.
- $M > 0$ is a constant as the initialization upper bound on θ.
- $J_\lambda^G(\pi)$ is the mean-variance objective function.
- $J_\lambda(\pi)$ is the reward-volatility objective function, with λ as the penalty factor.
- $J_\lambda^y(\pi)$ is the transformed reward-volatility objective function, with y as the auxiliary variable.
- $\tilde{r}$ is the reward for the augmented MDP. Similarly, $\tilde{V}_\pi(s)$ and $\tilde{Q}_\pi(s,a)$ are state value function and state-action value function of the augmented MDP, respectively. $\tilde{J}(\pi)$ is the risk-neural objective of the augmented MDP.
- $\hat{y}_k$ is an estimator of y at k-th iteration.
- ω is the parameter of critic network.
- $\delta_k = \text{argmin}_{\delta\in\mathcal{D}}\|\hat{F}(\theta_k)\delta - \tau_k\hat{\nabla}_\theta J(\pi_{\theta_k})\|_2$.
- $\xi_k(\delta) = \hat{F}(\theta_k)\delta - \tau_k\hat{\nabla}_\theta\tilde{J}(\pi_{\theta_k}) - \mathbb{E}[\hat{F}(\theta_k)\delta - \tau_k\hat{\nabla}_\theta\tilde{J}(\pi_{\theta_k})]$.
- σ_ξ is a constant associated with the upper bound of the gradient variance.
- $\varphi_k, \psi_k, \varphi_k', \psi_k'$ are the concentability coefficients, upper bounded by a constant $c_0 > 0$.
- $\varphi_k^* = \mathbb{E}_{(s,a)\sim\sigma_\pi}\left[\left(\frac{d\pi^*}{d\pi_0} - \frac{d\pi_{\theta_k}}{d\pi_0}\right)^2\right]^{1/2}$.
- $\psi_k^* = \mathbb{E}_{(s,a)\sim\sigma_\pi}\left[\left(\frac{d\sigma_{\pi^*}}{d\sigma_\pi} - \frac{d\nu_{\pi^*}}{d\nu_\pi}\right)^2\right]^{1/2}$.
- K is the total number of iterations. Similarly, K_{TD} is the total number of TD iterations.
- $c_3 > 0$ is a constant as to quantify the difference in risk-neutral objective between optimal policy and any policy.

Algorithm 2: A comparison between TOPS and MVPI

```
for k = 1, ..., K do
    Step 1: y_k := (1 − γ)J(π_k);
    Step 2: J̃(π_{θ_k}) := E_{(s,a)∼σ_{π_θ}}(r_{s,a} − λr²_{s,a} + 2λr_{s,a}y_k);
        if MVPI: then
            θ_k := arg max_θ(J̃(π_{θ_k}));
            // This is achieved by line 9 to 15 in Algorithm 3
        else if TOPS: then
            if select NPG update then
                update θ_k according to Eq. (11) ;
            else if select PPO update then
                update θ_k according to Eq. (12) ;
end
Output: π_{θ_K};
```

B Algorithm Details

We provide a comparison between MVPI and TOPS. Note that neither NPG nor PPO solve $\theta_k := \arg\max_\theta(\tilde{J}(\pi_{\theta_k}))$ directly, but instead solve an approximation optimization problem at each iteration. We provide pseudo-code for the implementation of MVPI and VARAC in Algorithm 3 and 4.

C Experimental Details

Note that although the mean-volatility method can be adapted to off-policy methods [60], in this paper, for the ease of the theoretical analysis, our proposed method is an on-policy actor-critic algorithm.

C.1 Testbeds

We use six Mujoco tasks from Open AI gym [8] as testbeds. They are `Half Cheetah-v2`, `Hopper-V2`, `Swimmer-V2`, `Walker2d-V2`, `InvertedPendulum-v2`, and `InvertedDoublePendulum-v2`.

C.2 Hyper-parameter Settings

In the experiment we set $\lambda = 1$. We then tune learning rate for different algorithms. For MVP, we use the same setting as [60]. For MVPI, TOPS and VARAC with neural NPG, we tune the learning rate of the actor network from $\{0.1, 1\times10^{-2}, 1\times10^{-3}, 7\times10^{-4}\}$ and the learning rate of the critic network from $\{1\times10^{-2}, 1\times10^{-3}, 7\times10^{-4}\}$. For MVPI, TOPS and VARAC with neural PPO, we tune the learning rate of the actor network from $\{3\times10^{-3}, 3\times10^{-4}, 3\times10^{-5}\}$ and the learning rate of the critic network from $\{1\times10^{-2}, 1\times10^{-3}, 1\times10^{-4}\}$.

Algorithm 3: MVPI with over-parameterized networks

Input: number of iteration K, learning rate for natural policy gradient (resp. PPO) TD η_{NPG} (resp. η_{PPO}), temperature parameters $\{\tau_k\}_{k=1}^K$;
Initialization: **Initialization**: Initialize policy network $f((s,a);\theta,b)$ as defined in Eq. (5). Set $\tau_1 = 1$. Initialize Q-network with (b,ω_1) similarly;
for $k = 1, \cdots, K$ **do**
  Sample a batch of transitions $\{s_t, a_t, r_t, s_t'\}_{t=1}^T$ following current policy with size of T;
  $y = \frac{1}{T}\sum_{t=1}^T r_t$;
  for $t = 1, \cdots, T$ **do**
    $\tilde{r}_t = r_t - \lambda r_t^2 + 2\lambda r_t y$, $a_t' \sim \pi(a|s_t')$;
  end
  repeat
    Q-value update: update ω_k according to Eq. (13);
    if *select NPG update* **then**
      update θ_k according to Eq. (11);
    else if *select PPO update* **then**
      update θ_k according to Eq. (12);
  until *CONVERGE*;
end
Output: π_{θ_K};

C.3 Computing Infrastructure

We conducted our experiments on a GPU GTX 970 and GPU GTX 1080Ti.

D Theoretical Analysis Details

In this section, we discuss the theoretical analysis in detail. We first present the overview in Sect. D.1. Then we provide additional assumptions in Sect. D.2. In the rest of the section, we present all the supporting lemmas and the proof for Theorem 1 and 2.

D.1 Overview

We provide Fig. 5 to illustrate the structure of the theoretical analysis. First, under Assumption 3 and 4, as well as Lemma 13. We can obtain Lemma 14, 15 and 16. These are the building blocks of Lemma 2, which is a shared component in the analysis of both NPG and PPO. The shared components also include Lemma 3, as well as Lemma 4 obtained under Assumption 5. For PPO analysis, under Assumption 2 and 4, we obtain Lemma 7 and 8 from Lemma 2 and 6, Then combined with Lemma 3, 4 and 9, we obtain Theorem 1, the major result of PPO analysis. Likely for NPG analysis, we first obtain Lemma 11 and 12 under Assumption 1, 2 and 4. Then together with Lemma 2, 3, 4 and 10, we obtain Theorem 2, the major result of NPG analysis.

Algorithm 4: VARAC

Input: number of iteration K, learning rate for natural policy gradient (resp. PPO) TD η_{NPG} (resp. η_{PPO}), temperature parameters $\{\tau_k\}_{k=1}^K$;
Initialization: Initialize policy network $f((s,a);\theta,b)$ as defined in Eq. (5). Set $\tau_1 = 1$. Initialize Q-network with (b,ω_1) similarly;
for $k = 1, \cdots, K$ **do**
    Sample a batch of transitions $\{s_t, a_t, r_t, s_t'\}_{t=1}^T$ following current policy with size of T;
    $y = \frac{1}{T}\sum_{t=1}^T r_t$;
    Q-value update: update both networks' ω_k according to Eq. (13);
    Output Q_k and W_k;
    update θ_k with NPG or PPO;
end
Output: π_{θ_K};

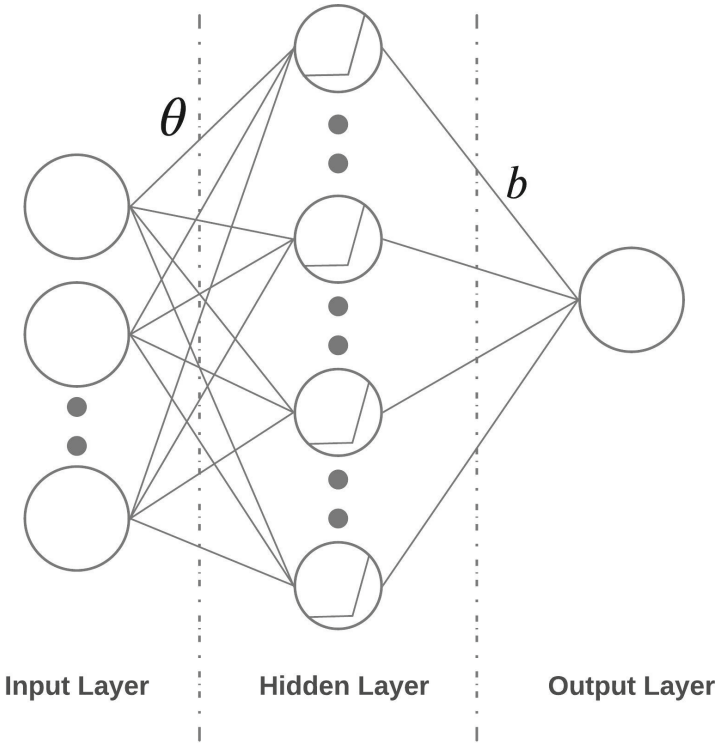

Fig. 4. A block diagram of over-parameterized neural network

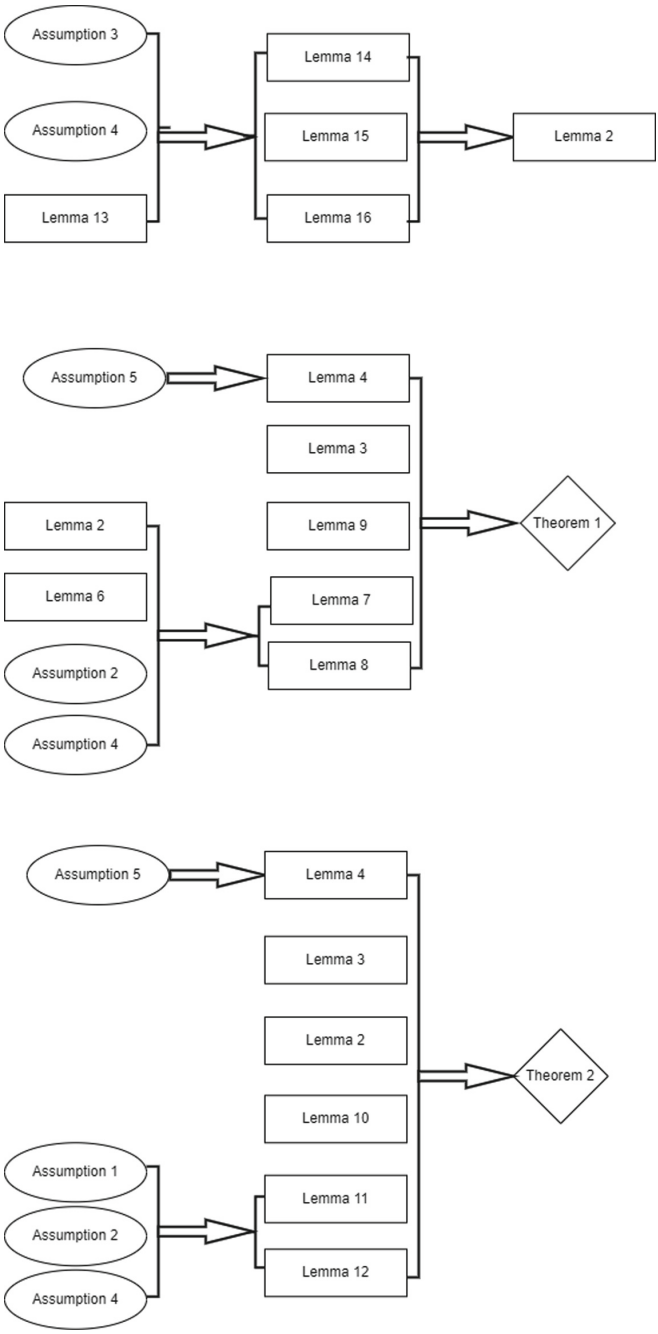

Fig. 5. A flow chart of the theoretical analysis

D.2 Additional Assumptions

Assumption 3 *(Action-value function class). We define*

$$\mathcal{F}_{\Upsilon,\infty} := \Bigg\{ f(s,a;\theta) = f_0(s,a)) + \int \mathbb{1}\{\theta^\top (s,a) > 0\}(s,a)^\top \iota(\theta) d\mu(w) : \|\iota(\theta)\|_\infty \leq \Upsilon/\sqrt{d} \Bigg\}$$

where $\mu : \mathbb{R}^d \to [0,1]$ is a probability density function of $\mathcal{N}(0, I_d/d)$. $f_0(s,a)$ is the two-layer neural network corresponding to the initial parameter Θ_{init}, and $\iota : \mathbb{R}^d \to \mathbb{R}^d$ is a weighted function. We assume that $\tilde{Q}_\pi \in \mathcal{F}_{\Upsilon,\infty}$ for all π.

Assumption 4 *(Regularity of stationary distribution). For any policy π, and $\forall x \in \mathbb{R}^d, \forall \|x\|_2 = 1$, and $\forall u > 0$, we assume that there exists a constant $c > 0$ such that $\mathbb{E}_{(s,a)\sim\sigma_\pi}\left[\mathbb{1}\{|x^\top (s,a)| \leq u\}\right] \leq cu$.*

Assumption 3 is a mild regularity condition on Q_π, as $\mathcal{F}_{\Upsilon,\infty}$ is a sufficiently rich function class and approximates a subset of the reproducing kernel Hilbert space (RKHS) [40]. Similar assumptions are widely imposed [4,16,38,51,58]. Assumption 4 is a regularity condition on the transition kernel $\mathcal{P}$. Such regularity holds so long as σ_π has an upper bound density, satisfying most Markov chains.

In [62] Lemma 4.15, they make a mistake in the proof. They accidentally flip a sign in $y^* - \bar{y}$ when transitioning from the first equation in the proof to Eq. (4.15). This invalidates the conclusion in Eq. (4.17), an essential part of the proof. We tackle this issue by proposing the next assumption.

Assumption 5 *(Convergence Rate of $J(\pi)$). We assume π^* (the optimal policy to the risk-averse objective function $J_\lambda(\pi)$) converges to the risk-neutral objective $J(\pi)$ for both NPG and PPO with the over-parameterized neural network to be $\mathcal{O}(1/\sqrt{k})$. Specifically, there exists a constant $c_3 > 0$ such that,*

$$J(\pi^*) - J(\pi_k) \leq \frac{c_3}{\sqrt{k}}$$

It was proved [31,51] that the optimal policy w.r.t the risk-neutral objective $J(\pi)$ obtained by NPG and PPO method with the over-parameterized two-layer neural network converges to the globally optimal policy at a rate of $\mathcal{O}(1/\sqrt{K})$, where K is the number of iteration. Since our method uses similar settings, we assume the convergence rates of risk-neutral objective $J(\pi)$ in our paper follow their results.

In the following subsections, we study TOPS's convergence of global optimality and provide a proof sketch.

D.3 Proof of Theorem 1

We first present the analysis of policy evaluation error, which is induced by TD update in Line 9 of Algorithm 1. We characterize the policy evaluation error in the following lemma:

Lemma 2 *(Policy Evaluation Error). We set learning rate of TD* $\eta_{TD} = \min\{(1-\gamma)/3(1+\gamma)^2, 1/\sqrt{K_{\mathrm{TD}}}\}$. *Under Assumption 3 and 4, it holds that, with probability of* $1-\delta$,

$$\begin{aligned}&\|\tilde{Q}_{\omega_k} - \tilde{Q}_{\pi_k}\|^2_{\nu_{\pi_k}}\\&= \mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta) + \Upsilon^{5/2} m^{-1/4}\sqrt{\log(1/\delta)}\\&+ \Upsilon r^2_{\max} m^{-1/4} + \Upsilon^2 K_{\mathrm{TD}}^{-1/2} + \Upsilon),\end{aligned} \tag{14}$$

where $\tilde{Q}_{\pi_k}$ is the Q-value function of the augmented MDP, and $\tilde{Q}_{\omega_k}$ is its estimator at the k-th iteration. We provide the proof and its supporting lemmas in Appendix D.6. In the following, we establish the error induced by the policy update. Equation (8) can be re-expressed as

$$J^y_\lambda(\pi) = \sum_{s,a} \sigma_\pi \big(r_{s,a} - \lambda r^2_{s,a} + 2\lambda r_{s,a} y_{k+1}\big) - \lambda y^2_{k+1} \tag{15}$$

It can be shown that $\forall \pi, \max_y J^y_\lambda(\pi) = J_\lambda(\pi)$ [55,60]. We denote the optimal policy to the augmented MDP associated with y^* by $\pi^*(y^*)$. By definition, it is obvious that π^* and $\pi^*(y^*)$ are equivalent. For simplicity, we will use the unified term π^* in the rest of the paper. We present Lemma 3 and 4.

Lemma 3 *(Policy's Performance Difference). For mean-volatility objective w.r.t. auxiliary variable* y *as* $J^y_\lambda(\pi)$ *defined in Eq. (15). For any policy* π *and* π', *we have the following,*

$$\begin{aligned}J^y_\lambda(\pi') - J^y_\lambda(\pi) = (1-\gamma)^{-1}\mathbb{E}_{s\sim\nu_{\pi'}}\big[&\mathbb{E}_{a\sim\pi'}[\tilde{Q}_{\pi,y}]\\&- \mathbb{E}_{a\sim\pi}[\tilde{Q}_{\pi,y}]\big],\end{aligned}$$

where $\tilde{Q}_{\pi,y}$ *is the state-action value function of the augmented MDP, and its rewards are associated with* y.

Proof. When y is fixed,

$$\begin{aligned}&J^y_\lambda(\pi') - J^y_\lambda(\pi)\\&= \sum_{s,a}\sigma_{\pi'}\tilde{r}_{s,a} - \sum_{s,a}\sigma_\pi \tilde{r}_{s,a} = \tilde{J}(\pi')) - \tilde{J}(\pi)\end{aligned} \tag{16}$$

We then follow Lemma 6.1 in [21]:

$$\tilde{J}(\pi') - \tilde{J}(\pi) = (1-\gamma)^{-1}\mathbb{E}_{(s,a)\sim\sigma_{\pi'}}\left[\tilde{A}_\pi\right] \tag{17}$$

where $\tilde{A}_\pi = \tilde{Q}_\pi - \tilde{V}_\pi$ is the advantage function of policy π. Meanwhile,

$$\mathbb{E}_{a\sim\pi'}[\tilde{A}_\pi] = \mathbb{E}_{a\sim\pi'}[\tilde{Q}_\pi] - \tilde{V}_\pi = \mathbb{E}_{a\sim\pi'}[\tilde{Q}_\pi] - \mathbb{E}_{a\sim\pi}[\tilde{Q}_\pi] \tag{18}$$

From Eq. (16), Eq. (17) and Eq. (18), we complete the proof.

Lemma 3 is inspired by [21] and adopted by most work on global convergence [1, 31,57]. Next, we derive an upper bound for the error of the critic update in Line 5 of Algorithm 1:

Lemma 4 *(y Update Error). We characterize the error induced by the estimation of auxiliary variable y w.r.t the optimal value y^* at k-th iteration as, $J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi^*) = \frac{2c_3 r_{\max}(1-\gamma)\lambda}{\sqrt{k}}$, where $r_{\max}$ is the bound of the original reward, and c_3 is a constant error term.*

Proof. We start from the subproblem objective defined in Eq. (15) with y^* and $\hat{y}_k$:

$$
\begin{aligned}
&J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi^*) \\
&= \left(\sum_{s,a} \sigma_{\pi^*} \left(r_{s,a} - \lambda r_{s,a}^2 + 2\lambda r_{s,a} y^* \right) - \lambda y^{*2} \right) \\
&- \left(\sum_{s,a} \sigma_{\pi^*} \left(r_{s,a} - \lambda r_{s,a}^2 + 2\lambda r_{s,a} \hat{y}_k \right) - \lambda \hat{y}_k^2 \right) \\
&= 2\lambda \big(\sum_{s,a} \sigma_{\pi^*} r_{s,a} \big)(y^* - \hat{y}_k) - \lambda(y^{*2} - \hat{y}_k^2) \\
&= \lambda \langle y^* - \hat{y}_k, 2(1-\gamma) J(\pi^*) - y^* - \hat{y}_k \rangle \\
&= (1-\gamma)\lambda \langle y^* - \hat{y}_k, J(\pi^*) - \hat{J}(\pi_k) \rangle
\end{aligned}
$$

where we obtain the final two equalities by the definition of J_π and y. Because $r_{s,a}$ is upper-bounded by a constant $r_{\max}$, we have $|y^* - \hat{y}_k| \leq 2r_{\max}$. Under Assumption 5 we have,

$$
J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi^*) = \frac{2c_3 r_{\max}(1-\gamma)\lambda}{\sqrt{k}}
$$

Thus we finish the proof.

From Lemma 3 and 4, we can also obtain the following Lemma.

Lemma 5 *(Performance Difference on π and y). For mean-volatility objective w.r.t. auxiliary variable y as $J_\lambda^y(\pi)$ defined in Eq. (15). For any π, y and the optimal $\pi*, y*$, we have the following,*

$$
\begin{aligned}
J_\lambda^{y^*}(\pi^*) - J_\lambda^y(\pi) = (1-\gamma)^{-1} \mathbb{E}_{s\sim\nu_{\pi^*}} \big[&\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi,y}] \\
&- \mathbb{E}_{a\sim\pi}[\tilde{Q}_{\pi,y}] \big] + \frac{2c_3 r_{\max}(1-\gamma)\lambda}{\sqrt{k}}.
\end{aligned}
$$

where $\tilde{Q}_{\pi,y}$ is the state-action value function of the augmented MDP, and its rewards are associated with y.

Proof. It is easy to see that $J_\lambda^{y^*}(\pi^*) - J_\lambda^y(\pi) = J_\lambda^{y^*}(\pi^*) - J_\lambda^y(\pi^*) + J_\lambda^y(\pi^*) - J_\lambda^y(\pi)$. Then replace $J_\lambda^y(\pi^*) - J_\lambda^y(\pi)$ with Lemma 3 and $J_\lambda^{y^*}(\pi^*) - J_\lambda^y(\pi^*)$ with Lemma 4, we finish the proof.

Lemma 5 quantifies the performance difference of $J_\lambda^y(\pi)$ between any pair π, y and the optimal $\pi*, y*$, while Lemma 3 only quantifies the performance difference of $J_\lambda^y(\pi)$ between π and π' when y is fixed.

We now study the global convergence of TOPS with neural PPO as the policy update component. First, we define the neural PPO update rule.

Lemma 6 [31]. *Let $\pi_{\theta_k} \propto \exp\{\tau_k^{-1} f_{\theta_k}\}$ be an energy-based policy. We define the update*

$$\hat{\pi}_{k+1} = \arg\max_{\pi} \mathbb{E}_{s\sim\nu_k}[\mathbb{E}_\pi[Q_{\omega_k}] - \beta_k KL(\pi_\theta \| \pi_{\theta_k})],$$

where Q_{ω_k} is the estimator of the exact action-value function $Q^{\pi_{\theta_k}}$. We have

$$\hat{\pi}_{k+1} \propto \exp\{\beta_k^{-1} Q_{\omega_k} + \tau_k^{-1} f_{\theta_k}\}$$

And to represent $\hat{\pi}_{k+1}$ with $\pi_{\theta_{k+1}} \propto \exp\{\tau_{k+1}^{-1} f_{\theta_{k+1}}\}$, we solve the following sub-problem,

$$\theta_{k+1} = \arg\min_{\theta\in\mathbb{D}} \mathbb{E}_{(s,a)\sim\sigma_k}[(f_\theta(s,a) - \tau_{k+1}(\beta_k^{-1} Q_{\omega_k}(s,a) + \tau_k^{-1} f_{\theta_k}(s,a)))^2]$$

We analyze the policy improvement error in Line 13 of Algorithm 1. [31] proves that the policy improvement error can be characterized similarly to the policy evaluation error as in Eq. (14). Recall $\tilde{Q}_{\omega_k}$ is the estimator of Q-value, f_{θ_k} the energy function for policy, and $f_{\hat{\theta}}$ its estimator. We characterize the policy improvement error as follows: Under Assumptions 3 and 4, we set the learning rate of PPO $\eta_{\text{PPO}} = \min\{(1-\gamma)/3(1+\gamma)^2 1/\sqrt{K_{\text{TD}}}\}$, and with a probability of $1-\delta$:

$$\begin{aligned} &\|(f_{\hat{\theta}} - \tau_{k+1}(\beta^{-1}\tilde{Q}_{\omega_k} + \tau_k^{-1} f_{\theta_k})\|^2 \\ &= \mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta) + \Upsilon^{5/2} m^{-1/4}\sqrt{\log(1/\delta)} \\ &+ \Upsilon r_{\max}^2 m^{-1/4} + \Upsilon^2 K_{\text{TD}}^{-1/2} + \Upsilon). \end{aligned} \tag{19}$$

We quantify how the errors propagate in neural PPO [31] in the following.

Lemma 7 [31]. *(Error Propagation) We have,*

$$\begin{aligned} &\big|\mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathbb{E}_{a\sim\pi^*}[\log(\pi_{\theta_{k+1}}/\pi_{k+1}] - \mathbb{E}_{a\sim\pi_{\theta_k}} \\ &[\log(\pi_{\theta_{k+1}}/\pi_{k+1}]]\big| \le \tau_{k+1}^{-1}\varepsilon_k''\varphi_{k+1}^* + \beta^{-1}\varepsilon_k''\psi_k^* \end{aligned} \tag{20}$$

ε_k'' are defined in Eq. (14) as well as Eq. (19). $\varphi_k^* = \mathbb{E}_{(s,a)\sim\sigma_\pi}\left[\left(\frac{d\pi^*}{d\pi_0} - \frac{d\pi_{\theta_k}}{d\pi_0}\right)^2\right]^{1/2}$, $\psi_k^* = \mathbb{E}_{(s,a)\sim\sigma_\pi}\left[\left(\frac{d\sigma_{\pi^*}}{d\sigma_\pi} - \frac{d\nu_{\pi^*}}{d\nu_\pi}\right)^2\right]^{1/2}$. $\frac{d\pi^*}{d\pi_0}, \frac{d\pi_{\theta_k}}{d\pi_0}, \frac{d\sigma_{\pi^*}}{d\sigma_\pi}, \frac{d\nu_{\pi^*}}{d\nu_\pi}$ are the Radon-Nikodym derivatives [23]. We denote RHS in Eq. (20) by $\varepsilon_k = \tau_{k+1}^{-1}\varepsilon_k''\varphi_{k+1}^* + \beta^{-1}\varepsilon_k''\psi_k^*$. Lemma 7 essentially quantifies the error from which we use the two-layer neural network to approximate the action-value function

and policy instead of having access to the exact ones. Please refer to [31] for complete proofs of Lemma 6 and 7.

$$\begin{aligned}&\big|\mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathbb{E}_{a\sim\pi^*}[\log(\pi_{\theta_{k+1}}/\pi_{k+1}]\\&-\mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\pi_{\theta_{k+1}}/\pi_{k+1}]]\big|\leq\tau_{k+1}^{-1}\varepsilon_k''\varphi_{k+1}^*+\beta^{-1}\varepsilon_k''\psi_k^*\end{aligned}$$

We then characterize the difference between energy functions in each step [31]. Under the optimal policy $\pi*$,

Lemma 8 [31]. *(Stepwise Energy Function difference) Under the same condition of Lemma 7, we have*

$$\mathbb{E}_{s\sim\nu_{\pi^*}}[\|\tau_{k+1}^{-1}f_{\theta_{k+1}}-\tau_k^{-1}f_{\theta_k}\|_\infty^2]\leq 2\varepsilon_k'+2\beta_k^{-2}U, \tag{21}$$

where $\varepsilon_k'=|\mathcal{A}|\tau_{k+1}^{-2}\epsilon_{k+1}^2$
and $U=2\mathbb{E}_{s\sim\nu_{\pi^*}}[\max_{a\in\mathcal{A}}(\tilde{Q}_{\omega_0})^2]+2\Upsilon^2$.

Proof. By the triangle inequality, we get the following,

$$\begin{aligned}&\|\tau_{k+1}^{-1}f_{\theta_{k+1}}-\tau_k^{-1}f_{\theta_k}\|_\infty^2\\&\leq 2\big(\|\tau_{k+1}^{-1}f_{\theta_{k+1}}-\tau_k^{-1}f_{\theta_k}-\beta^{-1}\tilde{Q}_{\omega_k}\|_\infty^2+\|\beta^{-1}\tilde{Q}_{\omega_k}\|_\infty^2\big)\end{aligned} \tag{22}$$

We take the expectation of both sides of Eq. (22) with respect to $s\sim\nu_{\pi^*}$. With the 1-Lipshitz continuity of $\tilde{Q}_{\omega_k}$ in ω and $\|\omega_k-\Theta_{\text{init}}\|_2\leq\Upsilon$, we have,

$$\begin{aligned}&\mathbb{E}_{\nu_{\pi^*}}\big[\|\tau_{k+1}^{-1}f_{\theta_{k+1}}-\tau_k^{-1}f_{\theta_k}\|_\infty^2\big]\\&\leq 2(|\mathcal{A}|\tau_{k+1}^{-2}\epsilon_{k+1}^2+\mathbb{E}_{s\sim\nu_{\pi^*}}[\max_{a\in\mathcal{A}}(\tilde{Q}_{\omega_0})^2]+\Upsilon^2)\end{aligned}$$

Thus complete the proof.

We then derive a difference term associated with π_{k+1} and π_{θ_k}, where at the k-th iteration π_{k+1} is the solution for the following subproblem,

$$\pi_{k+1}=\arg\max_\pi\Big(\mathbb{E}_{s\sim\nu_{\pi_k}}\big[\mathbb{E}_{a\sim\pi}[\tilde{Q}_{\pi_k,\hat{y}_k}]-\beta\text{KL}(\pi\|\pi_{\theta_k})\big]\Big)$$

and π_{θ_k} is the policy parameterized by the two-layered over-parameterized neural network. The following lemma establishes the one-step descent of the KL-divergence in the policy space:

Lemma 9 *(One-step difference of π). For π_{k+1} and π_{θ_k}, we have*

$$\begin{aligned}&KL(\pi^*\|\pi_{\theta_k})-KL(\pi^*\|\pi_{\theta_{k+1}})\\&\geq\big(\mathbb{E}_{a\sim\pi^*}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{k+1}})]-\mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{k+1}})]\big)\\&\quad+\beta^{-1}\big(\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}]-\mathbb{E}_{a\sim\pi_{\theta_k}}[\tilde{Q}_{\pi_k,\hat{y}_k}]\big)\\&\quad+\frac{1}{2}\|\pi_{\theta_{k+1}}-\pi_{\theta_k}\|_1^2+\big(\mathbb{E}_{a\sim\pi_{\theta_k}}[\tau_{k+1}^{-1}f_{\theta_{k+1}}-\tau_k^{-1}f_{\theta_k}]\\&\quad-\mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\tau_{k+1}^{-1}f_{\theta_{k+1}}-\tau_k^{-1}f_{\theta_k}]\big)\end{aligned} \tag{23}$$

Proof. We start from

$$\mathrm{KL}(\pi^*\|\pi_{\theta_k}) - \mathrm{KL}(\pi^*\|\pi_{\theta_{k+1}}) = \mathbb{E}_{a\sim\pi^*}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})]$$

$$(\text{By definition, } \mathrm{KL}(\pi_{\theta_{k+1}}\|\pi_{\theta_k}) = \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})]))$$

$$= \big(\mathbb{E}_{a\sim\pi^*}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})] - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})]\big) +$$

$$\mathrm{KL}(\pi_{\theta_{k+1}}\|\pi_{\theta_k})$$

We then add and subtract terms,

$$= \mathbb{E}_{a\sim\pi^*}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})] - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})] + \mathrm{KL}$$

$$(\pi_{\theta_{k+1}}\|\pi_{\theta_k}) + \beta^{-1}\big(\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\tilde{Q}_{\pi_k,\hat{y}_k}]\big)$$

$$- \beta^{-1}\big(\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\tilde{Q}_{\pi_k,\hat{y}_k}]\big)$$

$$+ \mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\pi_{\theta_{k+1}}\pi_{\theta_k})]$$

Rearrange the terms and we get,

$$= \big(\mathbb{E}_{a\sim\pi^*}[\log(\pi_{\theta_{k+1}}) - \log(\pi_{\theta_k}) - \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k}]$$

$$- \mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\pi_{\theta_{k+1}}) - \log(\pi_{\theta_k}) - \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k}]\big)$$

$$+ \beta^{-1}\big(\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\tilde{Q}_{\pi_k,\hat{y}_k}]\big) + \mathrm{KL}$$

$$(\pi_{\theta_{k+1}}\|\pi_{\theta_k}) + \big(\mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})] - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}$$

$$[\log(\pi_{\theta_{k+1}}\pi_{\theta_k})]\big) \tag{24}$$

Recall that $\pi_{k+1} \propto \exp\{\tau_k^{-1} f_{\theta_k} + \beta^{-1}\tilde{Q}^y_{\pi_k}\}$. We define the two normalization factors associated with ideal improved policy π_{k+1} and the current parameterized policy π_{θ_k} as,

$$Z_{k+1}(s) := \sum_{a'\in\mathcal{A}} \exp\{\tau_k^{-1} f_{\theta_k}(s,a') + \beta^{-1}\tilde{Q}^y_{\pi_k}(s,a')\}$$

$$Z_{\theta_{k+1}}(s) := \sum_{a'\in\mathcal{A}} \exp\{\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a')\}$$

We then have,

$$\pi_{k+1}(a|s) = \frac{\exp\{\tau_k^{-1} f_{\theta_k}(s,a) + \beta^{-1}\tilde{Q}^y_{\pi_k}(s,a)\}}{Z_{k+1}(s)}, \tag{25}$$

$$\pi_{\theta_{k+1}}(a|s) = \frac{\exp\{\tau_{k+1}^{-1} f_{\theta_{k+1}}(s,a)\}}{Z_{\theta_{k+1}}(s)} \tag{26}$$

For any π, π' and k, we have,

$$\mathbb{E}_{a\sim\pi}[\log Z_{\theta_{k+1}}] - \mathbb{E}_{a\sim\pi'}[\log Z_{\theta_{k+1}}] = 0 \tag{27}$$

$$\mathbb{E}_{a\sim\pi}[\log Z_{k+1}] - \mathbb{E}_{a\sim\pi'}[\log Z_{k+1}] = 0 \tag{28}$$

Now we look back at a few terms on RHS from Eq. (24):

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi^*}\left[\log(\pi_{\theta_k}) + \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k}\right] \\
&- \mathbb{E}_{a\sim\pi_{\theta_k}}\left[\log(\pi_{\theta_k}) + \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k}\right] \\
&= \left(\mathbb{E}_{a\sim\pi^*}[\tau_k^{-1} f_{\theta_k} + \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k} - \log Z_{\theta_{k+1}}]\right. \\
&\left. - \mathbb{E}_{a\sim\pi_{\theta_k}}[\tau_k^{-1} f_{\theta_k} + \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k} - \log Z_{\theta_{k+1}}]\right) \\
&= \mathbb{E}_{a\sim\pi^*}\left[\log\frac{\exp\{\tau_k^{-1} f_{\theta_k} + \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k}\}}{Z_{k+1}}\right] \\
&\quad - \mathbb{E}_{a\sim\pi_{\theta_k}}\left[\log\frac{\exp\{\tau_k^{-1} f_{\theta_k} + \beta^{-1}\tilde{Q}_{\pi_k,\hat{y}_k}\}}{Z_{k+1}}\right] \\
&= \mathbb{E}_{a\sim\pi^*}[\log\pi_{k+1}] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\log\pi_{k+1}]
\end{aligned} \tag{29}$$

For Eq. (29), we obtain the first equality by Eq. (26). Then, by swapping Eq. (27) with Eq. (28), we obtain the second equality. We achieve the concluding step with the definition in Eq. (25). Following a similar logic, we have,

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}}) - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\log(\frac{\pi_{\theta_{k+1}}}{\pi_{\theta_k}})] \\
&= \mathbb{E}_{a\sim\pi_{\theta_k}}[\tau_{k+1}^{-1} f_{\theta_{k+1}} - \log Z_{\theta_{k+1}} - \tau_k^{-1} f_{\theta_k} + \log Z_{\theta_k}] - \\
&\quad \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\tau_{k+1}^{-1} f_{\theta_{k+1}} - \log Z_{\theta_{k+1}} - \tau_k^{-1} f_{\theta_k} + \log Z_{\theta_k}] \\
&= \mathbb{E}_{a\sim\pi_{\theta_k}}[\tau_{k+1}^{-1} f_{\theta_{k+1}} - \tau_k^{-1} f_{\theta_k}] - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\tau_{k+1}^{-1} f_{\theta_{k+1}} - \\
&\quad \tau_k^{-1} f_{\theta_k}]
\end{aligned} \tag{30}$$

Finally, by using the Pinsker's inequality [12], we have,

$$\mathrm{KL}(\pi_{\theta_{k+1}} \| \pi_{\theta_k}) \geq 1/2 \| \pi_{\theta_{k+1}} - \pi_{\theta_k} \|_1^2 \tag{31}$$

Plugging Eqs. (29), (30), and (31) into Eq. (24), we have

$$\begin{aligned}
&\mathrm{KL}(\pi^* \| \pi_{\theta_k}) - \mathrm{KL}(\pi^* \| \pi_{\theta_{k+1}}) \\
&\geq \left(\mathbb{E}_{a\sim\pi^*}[\log(\pi_{\theta_{k+1}}) - \log(\pi_{k+1})] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\log(\pi_{\theta_{k+1}})\right. \\
&\left. - \log(\pi_{k+1})]\right) + \beta^{-1}\left(\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\tilde{Q}_{\pi_k,\hat{y}_k}]\right) \\
&+ \frac{1}{2}\| \pi_{\theta_{k+1}} - \pi_{\theta_k} \|_1^2 + \left(\mathbb{E}_{a\sim\pi_{\theta_k}}[\tau_{k+1}^{-1} f_{\theta_{k+1}} - \tau_k^{-1} f_{\theta_k}]\right. \\
&\left. - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\tau_{k+1}^{-1} f_{\theta_{k+1}} - \tau_k^{-1} f_{\theta_k}]\right)
\end{aligned}$$

Rearranging the terms, we obtain Lemma 9.

Lemma 9 serves as an intermediate-term for the major result's proof. We obtain upper bounds by telescoping this term in Theorem 1. Now we are ready to present the proof for Theorem 1.

Proof. First we take expectation of both sides of Eq. (23) with respect to $s \sim \nu_{\pi^*}$ from Lemma 9 and insert Eq. (20) to obtain,

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_{k+1}})] - \mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_k})] \\
\leq& \varepsilon_k - \beta^{-1}\mathbb{E}_{s\sim\nu_{\pi^*}}\left[\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\tilde{Q}_{\pi_k,\hat{y}_k}]\right] \\
&- 1/2\mathbb{E}_{s\sim\nu_{\pi^*}}\left[\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|_1^2\right] - \mathbb{E}_{s\sim\nu_{\pi^*}}\left[\mathbb{E}_{a\sim\pi_{\theta_k}}\right. \\
&\left.[\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}] - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}[\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}]\right]
\end{aligned} \tag{32}$$

Then, by Lemma 3, we have,

$$\begin{aligned}
&\beta^{-1}\mathbb{E}_{s\sim\nu_{\pi^*}}\left[\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_{\theta_k}}[\tilde{Q}_{\pi_k,\hat{y}_k}]\right] \\
&= \beta^{-1}(1-\gamma)\left(J_\lambda^{\hat{y}_k}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi)\right)
\end{aligned} \tag{33}$$

And with Hölder's inequality, we have,

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu_{\pi^*}}\left[\mathbb{E}_{a\sim\pi_{\theta_k}}[\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}] - \mathbb{E}_{a\sim\pi_{\theta_{k+1}}}\right. \\
&\left.[\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}]\right] \\
=& \mathbb{E}_{s\sim\nu_{\pi^*}}\left[\left\langle\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}, \pi_{\theta_k} - \pi_{\theta_{k+1}}\right\rangle\right] \\
\leq& \mathbb{E}_{s\sim\nu_{\pi^*}}\left[\|\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}\|_\infty\|\pi_{\theta_k} - \pi_{\theta_{k+1}}\|_1\right]
\end{aligned} \tag{34}$$

Insert Eqs. (33) and (34) into Eq. (32), we have,

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_{k+1}})] - \mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_k})] \\
\leq& \varepsilon_k - (1-\gamma)\beta^{-1}\left(J_\lambda^{\hat{y}_k}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi)\right) - 1/2\mathbb{E}_{s\sim\nu_{\pi^*}} \\
&\left[\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|_1^2\right] + \mathbb{E}_{s\sim\nu_{\pi^*}}\left[\|\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}\|_\infty\right. \\
&\left.\|\pi_{\theta_k} - \pi_{\theta_{k+1}}\|_1\right] \\
\leq& \varepsilon_k - (1-\gamma)\beta^{-1}\left(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi) - J_\lambda^{y^*}(\pi^*)\right. \\
&\left. + J_\lambda^{\hat{y}_k}(\pi^*)\right) + 1/2\mathbb{E}_{s\sim\nu_{\pi^*}}\left[\|\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}\|_\infty^2\right] \\
\leq& \varepsilon_k - (1-\gamma)\beta^{-1}\left(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi)\right) \\
&+ (1-\gamma)\beta^{-1}\left(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi^*)\right) \\
&+ 1/2\mathbb{E}_{s\sim\nu_{\pi^*}}\left[\|\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}\|_\infty^2\right].
\end{aligned}$$

The second inequality holds by using the inequality $2AB - B^2 \leq A^2$, with a minor abuse of notations. Here, $A := \|\tau_{k+1}^{-1}f_{\theta_{k+1}} - \tau_k^{-1}f_{\theta_k}\|_\infty$ and $B := \|\pi_{\theta_k} - \pi_{\theta_{k+1}}\|_1$. Then, by plugging in Lemma 4 and Eq. (21) we end up with,

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_{k+1}})] - \mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_k})] \\
\leq& \varepsilon_k - (1-\gamma)\beta^{-1}\left(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\right) \\
&+ (1-\gamma)\beta^{-1}\left(\frac{2c_3M(1-\gamma)\lambda}{\sqrt{k}}\right) + (\varepsilon_k' + \beta_k^{-2}U)
\end{aligned} \tag{35}$$

Rearrange Eq. (35), we have

$$\begin{aligned}
&(1-\gamma)\beta^{-1}\big(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \\
\leq\ &\mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_k})] - \mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_{\theta_{k+1}})] \\
&+ \big(\frac{2c_3 M(1-\gamma)^2\lambda}{\beta\sqrt{k}}\big) + \varepsilon_k + \varepsilon'_k + \beta_k^{-2}U
\end{aligned} \tag{36}$$

And then telescoping Eq. (36) results in,

$$\begin{aligned}
&(1-\gamma)\sum_{k=1}^{K}\beta^{-1}\min_{k\in[K]}\big(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \\
\leq\ &(1-\gamma)\sum_{k=1}^{K}\beta^{-1}\big(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \\
\leq\ &\mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_0)] - \mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_K)] \\
&+ \lambda r_{\max}(1-\gamma)^2\sum_{k=1}^{K}\beta^{-1}\big(\frac{2c_3}{\sqrt{k}}\big) + U\sum_{k=1}^{K}\beta_k^{-2} \\
&+ \sum_{k=1}^{K}(\varepsilon_k + \varepsilon'_k)
\end{aligned} \tag{37}$$

We complete the final step in Eq. (37) by plugging in Lemma 4 and Eq. (20). Per the observation we make in the proof of Theorem 2,

1. $\mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi^*\|\pi_0)] \leq \log\mathcal{A}$ due to the uniform initialization of policy.
2. $\mathrm{KL}(\pi^*\|\pi_K)$ is a non-negative term.

We now have,

$$\begin{aligned}
&\min_{k\in[K]} J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k) \\
\leq\ &\frac{\log|\mathcal{A}| + UK\beta^{-2} + \sum_{k=1}^{K}(\varepsilon_k + \varepsilon'_k)}{(1-\gamma)K\beta^{-1}} \\
&+ \lambda r_{\max}(1-\gamma)\big(\frac{2c_3}{\sqrt{k}}\big)
\end{aligned}$$

Replacing β with $\beta_0\sqrt{K}$ finishes the proof.

D.4 Proof of Theorem 2

In the following part, we focus the convergence of neural NPG. We first define the following terms under neural NPG update rule.

Lemma 10 [51]. *For energy-based policy π_θ, we have policy gradient and Fisher information matrix,*

$$\nabla_\theta J(\pi_\theta) = \tau \mathbb{E}_{d_{\pi_\theta}(s,a)}[Q_{\pi_\theta}(s,a)(\phi_\theta(s,a) - \mathbb{E}_{\pi_\theta}[\phi_\theta(s,a')])]$$
$$F(\theta) = \tau^2 \mathbb{E}_{d_{\pi_\theta}(s,a)}[(\phi_\theta(s,a) - \mathbb{E}_{\pi_\theta}[\phi_\theta(s,a')])$$
$$(\phi_\theta(s,a) - \mathbb{E}_{\pi_\theta}[\phi_\theta(s,a')])^\top]$$

We then derive an upper bound for $J_\lambda^{y^*}(\pi^*) - J_\lambda^{y^*}(\pi_k)$ for the neural NPG method in the following lemma:

Lemma 11 *(One-step difference of π). It holds that, with probability of $1-\delta$,*

$$\begin{aligned}
&(1-\gamma)\big(J_\lambda^{\hat{y}_k}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \\
\le\ &\eta_{\mathrm{NPG}}^{-1} \mathbb{E}_{s\sim\nu_{\pi^*}}\big[KL(\pi^*\|\pi_k) - KL(\pi^*\|\pi_{k+1})\big] \\
&+ \eta_{\mathrm{NPG}}(9\Upsilon^2 + r_{\max}^2) + 2c_0\epsilon_k' + \eta_{\mathrm{NPG}}^{-1}\epsilon_k'',
\end{aligned}$$

where

$$\begin{aligned}
\epsilon_k' &= \mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta) + \Upsilon^{5/2} m^{-1/4}\sqrt{\log(1/\delta)} \\
&\quad + \Upsilon r_{\max}^2 m^{-1/4} + \Upsilon^2 K_{\mathrm{TD}}^{-1/2} + \Upsilon), \\
\epsilon_k'' &= 8\eta_{\mathrm{NPG}}\Upsilon^{1/2} c_0 \sigma_\xi^{1/2} T^{-1/4} \\
&\quad + \mathcal{O}((\tau_{k+1} + \eta_{\mathrm{NPG}})\Upsilon^{3/2} m^{-1/4} \\
&\quad + \eta_{\mathrm{NPG}}\Upsilon^{5/4} m^{-1/8}),
\end{aligned}$$

c_0 is defined in Assumption 2 and σ_ξ is defined in Assumption 1. Meanwhile, Υ is the radius of the parameter space, m is the width of the neural network, and T is the sample batch size.

Proof. We start from the following,

$$\begin{aligned}
&\mathrm{KL}(\pi^*\|\pi_k) - \mathrm{KL}(\pi^*\|\pi_{k+1}) - \mathrm{KL}(\pi_{k+1}\|\pi_k) \\
&= \mathbb{E}_{a\sim\pi^*}\big[\log(\frac{\pi_{k+1}}{\pi_k})\big] - \mathbb{E}_{a\sim\pi_{k+1}}\big[\log(\frac{\pi_{k+1}}{\pi_k})\big] \qquad (38) \\
&\text{(by KL's definition).}
\end{aligned}$$

We now show the building blocks of the proof. *First*, we add and subtract a few terms to RHS of Eq. (38) then take the expectation of both sides with respect to $s \sim \nu_{\pi^*}$. Rearrange these terms, we get,

$$\begin{aligned}
&\mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathrm{KL}(\pi^*\|\pi_k) - \mathrm{KL}(\pi^*\|\pi_{k+1}) - \mathrm{KL}(\pi_{k+1}\|\pi_k)\big] \\
&= \eta_{\mathrm{NPG}}\mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_k}[\tilde{Q}_{\pi_k,\hat{y}_k}]\big] \\
&+ H_k \qquad (39)
\end{aligned}$$

where H_k is denoted by,

$$\begin{aligned} H_k := & \mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathbb{E}_{a\sim\pi^*}[\log(\frac{\pi_{k+1}}{\pi_k}) - \eta_{\mathrm{NPG}}\tilde{Q}_{\omega_k}\big] \\ & - \mathbb{E}_{a\sim\pi_k}\big[\log(\frac{\pi_{k+1}}{\pi_k}) - \eta_{\mathrm{NPG}}\tilde{Q}_{\omega_k}]\big] \\ & + \eta_{\mathrm{NPG}}\mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\omega_k} - \tilde{Q}_{\pi_k,\hat{y}_k}] \\ & - \mathbb{E}_{a\sim\pi_k}[\tilde{Q}_{\omega_k} - \tilde{Q}_{\pi_k,\hat{y}_k}]\big] \\ & + \mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathbb{E}_{a\sim\pi_k}[\log(\frac{\pi_{k+1}}{\pi_k})] \\ & - \mathbb{E}_{a\sim\pi_{k+1}}[\log(\frac{\pi_{k+1}}{\pi_k})]\big] \end{aligned} \tag{40}$$

By Lemma 3, we have

$$\begin{aligned} & \eta_{\mathrm{NPG}}\mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathbb{E}_{a\sim\pi^*}[\tilde{Q}_{\pi_k,\hat{y}_k}] - \mathbb{E}_{a\sim\pi_k}[\tilde{Q}_{\pi_k,\hat{y}_k}]\big] \\ & = \eta_{\mathrm{NPG}}(1-\gamma)\big(J_\lambda^{\hat{y}_k}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \end{aligned} \tag{41}$$

Insert Eqs. (41) back to Eq. (39), we have,

$$\begin{aligned} & \eta_{\mathrm{NPG}}(1-\gamma)\big(J_\lambda^{\hat{y}_k}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \\ & = \mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathrm{KL}(\pi^*\|\pi_k) - \mathrm{KL}(\pi^*\|\pi_{k+1}) - \mathrm{KL}(\pi_{k+1}\|\pi_k)\big] \\ & - H_k \\ \leq & \mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathrm{KL}(\pi^*\|\pi_k) - \mathrm{KL}(\pi^*\|\pi_{k+1}) - \mathrm{KL}(\pi_{k+1}\|\pi_k)\big] \\ & + |H_k| \end{aligned} \tag{42}$$

We reach the final inequality of Eq. (42) by algebraic manipulation. *Second*, we follow Lemma 5.5 of [51] and obtain an upper bound for Eq. (40). Specifically, with probability of $1-\delta$,

$$\begin{aligned} & \mathbb{E}_{a\sim\mathrm{init}}\Big[|H_k| - \mathbb{E}_{s\sim\nu_{\pi^*}}[\mathrm{KL}(\pi_{k+1}\|\pi_k)]\Big] \\ \leq & \eta_{\mathrm{NPG}}^2(9\Upsilon^2 + r_{\max}^2) + 2\eta_{\mathrm{NPG}}c_0\epsilon_k' + \epsilon_k'' \end{aligned} \tag{43}$$

The expectation is taken over randomness. With these building blocks of Eqs. (42) and (43), we are now ready to reach the concluding inequality. Plugging Eqs. (43) back into Eq. (42), we end up with, with probability of $1-\delta$,

$$\begin{aligned} & \eta_{\mathrm{NPG}}(1-\gamma)\big(J_\lambda^{\hat{y}_k}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big) \\ \leq & \mathbb{E}_{s\sim\nu_{\pi^*}}\big[\mathrm{KL}(\pi^*\|\pi_k) - \mathrm{KL}(\pi^*\|\pi_{k+1})\big] \\ & + \eta_{\mathrm{NPG}}^2(9\Upsilon^2 + r_{\max}^2) + 2\eta_{\mathrm{NPG}}c_0\epsilon_k' + \epsilon_k'' \end{aligned} \tag{44}$$

Dividing both sides of Eq. (44) by η_{NPG} completes the proof. The details are included in the Appendix.

We have the following Lemma to bound the error terms H_k defined in Eq. (40) of Lemma 11.

Lemma 12 [51]. *Under Assumptions 4, we have*

$$\begin{aligned}&\mathbb{E}_{a\sim\text{init}}\Big[|H_k| - \mathbb{E}_{s\sim\nu_{\pi^*}}[KL(\pi_{k+1}\|\pi_k)]\Big]\\ &\leq \eta_{\text{NPG}}^2(9\Upsilon^2 + r_{\max}^2) + \eta_{\text{NPG}}(\varphi_k' + \psi_k')\epsilon_k' + \epsilon_k''\end{aligned}$$

Here the expectation is taken over all the randomness. We have $\epsilon_k' := \|Q_{\omega_k} - Q_{\pi_k}\|_{\nu_{\pi_k}}^2$ *and*

$$\begin{aligned}\epsilon_k'' = \sqrt{2}\Upsilon^{1/2}\eta_{\text{NPG}}(\varphi_k + \psi_k)\tau_k^{-1}\big\{&\mathbb{E}_{(s,a)\sim\sigma_{\pi_{\theta_k}}}[\|\xi_k(\delta_k)\|_2^2]\\ &+ \mathbb{E}_{(s,a)\sim\sigma_{\pi_{\omega_k}}}[\|\xi_k(\omega_k)\|_2^2]\big\}^{1/2}\\ &+ \mathcal{O}((\tau_{k+1} + \eta_{\text{NPG}})\Upsilon^{3/2}m^{-1/4} + \eta_{\text{NPG}}\Upsilon^{5/4}m^{-1/8}).\end{aligned}$$

Recall $\xi_k(\omega_k)$ *and* $\xi_k(\omega_k)$ *are defined in Assumption 1, while* φ_k,ψ_k, φ_k', *and* ψ_k *are defined in Assumption 2.*

Please refer to [51] for complete proof. Finally, we are ready to show the proof for Theorem 2.

Proof. First, we combine Lemma 4 and 11 to get the following:

$$\begin{aligned}&(1-\gamma)\big(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi^*) + J_\lambda^{\hat{y}_k}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big)\\ &\leq \eta_{\text{NPG}}^{-1}\mathbb{E}_{s\sim\nu_{\pi^*}}[\text{KL}(\pi^*\|\pi_k) - \text{KL}(\pi^*\|\pi_{k+1})]\\ &\quad + \eta_{\text{NPG}}(9\Upsilon^2 + r_{\max}^2) + 2c_0\epsilon_k' + \eta_{\text{NPG}}^{-1}\epsilon_k''\\ &\quad + \frac{2c_3M(1-\gamma)^2\lambda}{\sqrt{k}}\end{aligned} \tag{45}$$

We can then see this:

1. $\mathbb{E}_{s\sim\nu_{\pi^*}}[\text{KL}(\pi^*\|\pi_1)] \leq \log|\mathcal{A}|$ due to the uniform initialization of policy.
2. $\text{KL}(\pi^*\|\pi_{K+1})$ is a non-negative term.

And by setting $\eta_{\text{NPG}} = 1/\sqrt{K}$ and telescoping Eq. (45), we obtain,

$$\begin{aligned}&(1-\gamma)\min_{k\in[K]}\big(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k)\big)\\ &\leq (1-\gamma)\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}(J_\lambda^{y^*}(\pi^*) - J_\lambda^{\hat{y}_k}(\pi_k))\\ &\leq \frac{1}{\sqrt{K}}(\mathbb{E}_{s\sim\nu_{\pi^*}}[\text{KL}(\pi^*\|\pi_1)] + 9\Upsilon^2 + r_{\max}^2) + \frac{1}{K}\sum_{k=1}^{K}\\ &\quad (2\sqrt{K}c_0\epsilon_k' + \eta_{\text{NPG}}^{-1}\epsilon_k'' + \frac{2c_3M(1-\gamma)^2\lambda}{\sqrt{k}})\end{aligned} \tag{46}$$

plug ϵ_k' and ϵ_k'' defined in Lemma 11 into Eq. (46), and set ϵ_k as,

$$\begin{aligned}\epsilon_k &= \sqrt{8}c_0\Upsilon^{1/2}\sigma_\xi^{1/2}T^{-1/4} \\ &+ \mathcal{O}\big((\tau_{k+1}K^{1/2}+1)\Upsilon^{3/2}m^{-1/4}+\Upsilon^{5/4}m^{-1/8}\big) \\ &+ c_0\mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta)+\Upsilon^{5/2}m^{-1/4}\sqrt{\log(1/\delta)} \\ &+ \Upsilon r_{\max}^2 m^{-1/4}+\Upsilon^2 K_{\mathrm{TD}}^{-1/2}+\Upsilon)\end{aligned}$$

we complete the proof.

D.5 Proof of Lemma 1

Proof. First, we have $\mathbb{E}[G] = \frac{1}{1-\gamma}\mathbb{E}[R]$, i.e., the per-step reward R is an unbiased estimator of the cumulative reward G. Second, it is proved that $\mathbb{V}(G) \leq \frac{\mathbb{V}(R)}{(1-\gamma)^2}$ [7]. Given $\lambda \geq 0$, summing up the above equality and inequality, we have

$$\begin{aligned}\frac{1}{(1-\gamma)}J_{\frac{\lambda}{(1-\gamma)}}(\pi) &= \frac{1}{(1-\gamma)}\Big(\mathbb{E}[R]-\frac{\lambda}{(1-\gamma)}\mathbb{V}(R)\Big) \\ &\leq \mathbb{E}[G]-\lambda\mathbb{V}(G) = J_\lambda^G(\pi).\end{aligned}$$

It completes the proof.

D.6 Proof of Lemma 2

We first provide the supporting lemmas for Lemma 2. We define the local linearization of $f((s,a);\theta)$ defined in Eq. (4) at the initial point Θ_{init} as,

$$\hat{f}((s,a);\theta) = \frac{1}{\sqrt{m}}\sum_{v=1}^{m} b_v \mathbb{1}\{[\Theta_{\text{init}}]_v^\top(s,a)>0\}[\theta]_v^\top(s,a) \tag{47}$$

We then define the following function spaces,

$$\begin{aligned}\mathcal{F}_{\Upsilon,m} := \Bigg\{ &\frac{1}{\sqrt{m}}\sum_{v=1}^{m} b_v \mathbb{1}\big\{[\Theta_{\text{init}}]_v^\top(s,a)>0\big\}[\theta]_v^\top(s,a) : \\ &\|\theta-\Theta_{\text{init}}\|_2 \leq \Upsilon\Bigg\},\end{aligned}$$

and

$$\begin{aligned}\bar{\mathcal{F}}_{\Upsilon,m} := \Bigg\{ &\frac{1}{\sqrt{m}}\sum_{v=1}^{m} b_v \mathbb{1}\big\{[\Theta_{\text{init}}]_v^\top(s,a)>0\big\}[\theta]_v^\top(s,a) : \\ &\|[\theta]_v-[\Theta_{\text{init}}]_v\|_\infty \leq \Upsilon/\sqrt{md}\Bigg\}.\end{aligned}$$

$[\Theta_{\text{init}}]_r \sim \mathcal{N}(0, I_d/d)$ and $b_r \sim \text{Unif}(\{-1,1\})$ are the initial parameters. By the definition, $\bar{\mathcal{F}}_{\Upsilon,m}$ is a subset of $\mathcal{F}_{\Upsilon,m}$. The following lemma characterizes the deviation of $\bar{\mathcal{F}}_{\Upsilon,m}$ from $\mathcal{F}_{\Upsilon,\infty}$.

Lemma 13 *(Projection Error)* [40]. *Let $f \in \mathcal{F}_{\Upsilon,\infty}$, where $\mathcal{F}_{\Upsilon,\infty}$ is defined in Assumption 3. For any $\delta > 0$, it holds with probability at least $1-\delta$ that*

$$\|\Pi_{\bar{\mathcal{F}}_{\Upsilon,m}} f - f\|_{\varsigma} \leq \Upsilon m^{-1/2}[1+\sqrt{2\log(1/\delta)}]$$

where ς is any distribution over $S \times A$.

Please refer to [40] for a detail proof.

Lemma 14 *(Linearization Error). Under Assumption 4, for all $\theta \in \mathcal{D}$, where $\mathcal{D} = \{\xi \in \mathbb{R}^{md} : \|\xi - \Theta_{init}\|_2 \leq \Upsilon\}$, it holds that,*

$$\mathbb{E}_{\nu_\pi}\Big[\Big(f\big((s,a);\theta\big) - \hat{f}\big((s,a);\theta\big)\Big)^2\Big] \leq \frac{4c_1\Upsilon^3}{\sqrt{m}}$$

where $c_1 = c\sqrt{\mathbb{E}_{\mathcal{N}(0,I_d/d)}[1/\|(s,a)\|_2^2]}$, and c is defined in Assumption 4.

Proof. We start from the definitions in Eq. (4) and Eq. (47),

$$\begin{aligned}
&\mathbb{E}_{\nu_\pi}\Big[\Big(f\big((s,a);\theta\big) - \hat{f}\big((s,a);\theta\big)\Big)^2\Big] \\
&= \mathbb{E}_{\nu_\pi}\Big[\Big(\frac{1}{\sqrt{m}}\Big|\sum_{v=1}^{m}\Big(\big(\mathbb{1}\{[\theta]_v^\top(s,a) > 0\} - \mathbb{1}\{[\Theta_{\text{init}}]_v^\top(s,a) \\
&> 0\}\big)b_v[\theta]_v^\top(s,a)\Big)\Big|\Big)^2\Big] \\
&\leq \frac{1}{m}\mathbb{E}_{\nu_\pi}\Big[\Big(\sum_{v=1}^{m}\Big(\Big|\mathbb{1}\{[\theta]_v^\top(s,a) > 0\} - \mathbb{1}\{[\Theta_{\text{init}}]_v^\top(s,a) \\
&> 0\}\Big|\Big|b_v\Big|\Big|[\theta]_v^\top(s,a)\Big|\Big)\Big)^2\Big] \qquad (48)
\end{aligned}$$

The above inequality holds because the fact that $|\sum W| \leq \sum|W|$, where $W = \big((\mathbb{1}\{[\theta]_v^\top(s,a) > 0\} - \mathbb{1}\{[\Theta_{\text{init}}]_v^\top(s,a) > 0\})b_v[\theta]_v^\top(s,a)\big)$. Θ_{init} is defined in Eq. (5). Next, since $\mathbb{1}\{[\Theta_{\text{init}}]_v^\top(s,a) > 0\} \neq \mathbb{1}\{[\theta]_v^\top(s,a) > 0\}$, we have,

$$\begin{aligned}
|[\Theta_{\text{init}}]_v^\top(s,a)| &\leq |[[\theta]_v^\top(s,a) - \Theta_{\text{init}}]_v^\top(s,a)| \\
&\leq \|[\theta]_v - [\Theta_{\text{init}}]_v\|_2, \qquad (49)
\end{aligned}$$

where we obtain the last inequality from the Cauchy-Schwartz inequality. We also assume that $\|(s,a)\|_2 \leq 1$ without loss of generality [31,51]. Equation (49) further implies that,

$$\begin{aligned}
&|\mathbb{1}\{[\theta]_v^\top(s,a) > 0\} - \mathbb{1}\{[\Theta_{\text{init}}]_v^\top(s,a) > 0\}| \\
&\leq \mathbb{1}\{|[\Theta_{\text{init}}]_v^\top(s,a)| \leq \|[\theta]_v - [\Theta_{\text{init}}]_v\|_2\} \qquad (50)
\end{aligned}$$

Then plug Eq. (50) and the fact that $|b_v| \leq 1$ back to Eq. (48), we have the following,

$$
\begin{aligned}
&\mathbb{E}_{\nu_\pi}\Big[\Big(f\big((s,a);\theta\big)-\hat{f}\big((s,a);\theta\big)\Big)^2\Big]\\
&\leq \frac{1}{m}\mathbb{E}_{\nu_\pi}\Bigg[\Bigg(\sum_{v=1}^{m}\mathbb{1}\Big\{\Big|[\Theta_{\text{init}}]_v^\top(s,a)\Big|\leq\Big\|[\theta]_v-[\Theta_{\text{init}}]_v\Big\|_2\Big\}\\
&\quad\Big|[\theta]_v^\top(s,a)\Big|\Bigg)^2\Bigg]\\
&\leq \frac{1}{m}\mathbb{E}_{\nu_\pi}\Bigg[\Bigg(\sum_{v=1}^{m}\mathbb{1}\Big\{\Big|[\Theta_{\text{init}}]_v^\top(s,a)\Big|\leq\Big\|[\theta]_v-[\Theta_{\text{init}}]_v\Big\|_2\Big\}\\
&\quad\Big(\Big|\big([\theta]_v-[\Theta_{\text{init}}]_v\big)^\top(s,a)\Big|+\Big|[\Theta_{\text{init}}]_v^\top(s,a)\Big|\Big)\Bigg)^2\Bigg]\\
&\leq \frac{1}{m}\mathbb{E}_{\nu_\pi}\Bigg[\Bigg(\sum_{v=1}^{m}\mathbb{1}\Big\{\Big|[\Theta_{\text{init}}]_v^\top(s,a)\Big|\leq\Big\|[\theta]_v-[\Theta_{\text{init}}]_v\Big\|_2\Big\}\\
&\quad\Big(\Big\|[\theta]_v-[\Theta_{\text{init}}]_v\Big\|_2+\Big|[\Theta_{\text{init}}]_v^\top(s,a)\Big|\Big)\Bigg)^2\Bigg]\\
&\leq \frac{1}{m}\mathbb{E}_{\nu_\pi}\Bigg[\Bigg(\sum_{v=1}^{m}\mathbb{1}\Big\{\Big|[\Theta_{\text{init}}]_v^\top(s,a)\Big|\leq\Big\|[\theta]_v-[\Theta_{\text{init}}]_v\Big\|_2\Big\}\\
&\quad 2\Big\|[\theta]_v-[\Theta_{\text{init}}]_v\Big\|_2\Bigg)^2\Bigg]
\end{aligned}
\tag{51}
$$

We obtain the second inequality by the fact that $|A|\leq|A-B|+|B|$. Then follow the Cauchy-Schwartz inequality and $\|(s,a)\|_2\leq 1$ we have the third equality. By inserting Eq. (49) we achieve the fourth inequality. We continue Eq. (51) by following the Cauchy-Schwartz inequality and plugging $\big\|[\theta]-[\Theta_{\text{init}}]\big\|_2\leq\Upsilon$,

$$
\begin{aligned}
&\mathbb{E}_{\nu_\pi}\Big[\Big(f\big((s,a);\theta\big)-\hat{f}\big((s,a);\theta\big)\Big)^2\Big]\\
&\leq \frac{4\Upsilon^2}{m}\mathbb{E}_{\nu_\pi}\Big[\sum_{v=1}^{m}\mathbb{1}\{|[\Theta_{\text{init}}]_v^\top(s,a)|\leq\|[\theta]_v-[\Theta_{\text{init}}]_v\|_2\}\Big]\\
&= \frac{4\Upsilon^2}{m}\sum_{v=1}^{m}P_{\nu_\pi}|[\Theta_{\text{init}}]_v^\top(s,a)|\leq\|[\theta]_v-[\Theta_{\text{init}}]_v\|_2)\\
&\leq \frac{4c\Upsilon^2}{m}\sum_{v=1}^{m}\frac{\|[\theta]_v-[\Theta_{\text{init}}]_v\|_2}{\|\Theta_{\text{init}}]_v\|_2}\\
&\leq \frac{4c\Upsilon^2}{m}\Big(\sum_{v=1}^{m}\|[\theta]_v-[\Theta_{\text{init}}]_v\|_2^2\Big)^{-1/2}\Big(\sum_{v=1}^{m}\frac{1}{\|\Theta_{\text{init}}]_v\|_2^2}\Big)^{-1/2}\\
&\leq \frac{4c_1\Upsilon^3}{\sqrt{m}}
\end{aligned}
\tag{52}
$$

We obtain the second inequality by imposing Assumption 4 and the third by following the Cauchy-Schwartz inequality. Finally, we set $c_1 := c\sqrt{\mathbb{E}_{\mathcal{N}(0,I_d/d)}[1/\|(s,a)\|_2^2]}$. Thus, we complete the proof.

In the t-th iterations of TD iteration, we denote the temporal difference terms w.r.t $\hat{f}((s,a);\theta_t)$ and $f((s,a);\theta_t)$ as

$$\begin{aligned}\delta_t^0((s,a),(s,a)';\theta_t) &= \hat{f}((s,a)';\theta_t) - \gamma\hat{f}((s,a);\theta_t) \\ &\quad - r_{s,a}, \\ \delta_t^\theta((s,a),(s,a)';\theta_t) &= f((s,a)';\theta_t) - \gamma f((s,a);\theta_t) \\ &\quad - r_{s,a}.\end{aligned}$$

For notation simplicity in the sequel we write $\delta_t^0((s,a),(s,a)';\theta_t)$ and $\delta_t^\theta((s,a),(s,a)';\theta_t)$ as δ_t^0 and δ_t^θ. We further define the stochastic semi-gradient $g_t(\theta_t) := \delta_t^\theta \nabla_\theta f((s,a);\theta_t)$, its population mean $\bar{g}_t(\theta_t) := \mathbb{E}_{\nu_\pi}[g_t(\theta_t)]$. The local linearization of $\bar{g}_t(\theta_t)$ is $\hat{g}_t(\theta_t) := \mathbb{E}_{\nu_\pi}[\delta_t^0 \nabla_\theta \hat{f}((s,a);\theta_t)]$. We denote them as $g_t, \bar{g}_t, \hat{g}_t$ respectively for simplicity.

Lemma 15. *Under Assumption 4, for all $\theta_t \in \mathcal{D}$, where $\mathcal{D} = \{\xi \in \mathbb{R}^{md} : \|\xi - \Theta_{init}\|_2 \le \Upsilon\}$, it holds with probability of $1-\delta$ that,*

$$\begin{aligned}&\|\bar{g}_t - \hat{g}_t\|_2 \\ &= \mathcal{O}\Big(\Upsilon^{3/2} m^{-1/4}\big(1 + (m\log\frac{1}{\delta})^{-1/2}\big) + \Upsilon^{1/2} r_{\max} m^{-1/4}\Big)\end{aligned}$$

Proof. By the definition of $\bar{g}_t$ and $\hat{g}_t$, we have

$$\begin{aligned}&\big\|\bar{g}_t - \hat{g}_t\big\|_2^2 \\ &= \big\|\mathbb{E}_{\nu_\pi}[\delta_t^\theta \nabla_\theta f((s,a);\theta_t) - \delta_t^0 \nabla_\theta \hat{f}((s,a);\theta_t)]\big\|_2^2 \\ &= \big\|\mathbb{E}_{\nu_\pi}[(\delta_t^\theta - \delta_t^0)\nabla_\theta f((s,a);\theta_t) + \delta_t^0(\nabla_\theta f((s,a);\theta_t) - \\ &\nabla_\theta \hat{f}((s,a);\theta_t))]\big\|_2^2 \\ &\le 2\mathbb{E}_{\nu_\pi}\big[(\delta_t^\theta - \delta_t^0)^2 \|\nabla_\theta f((s,a);\theta_t)\|_2^2\big] + \\ &\quad 2\mathbb{E}_{\nu_\pi}\big[\big(|\delta_t^0| \|\nabla_\theta f((s,a);\theta_t) - \nabla_\theta \hat{f}((s,a);\theta_t))\|_2\big)^2\big] \qquad (53)\end{aligned}$$

We obtain the inequality because $(A+B)^2 \le 2A^2 + 2B^2$. We first upper bound $\mathbb{E}_{\nu_\pi}\big[(\delta_t^\theta - \delta_t^0)^2 \|\nabla_\theta f((s,a);\theta_t)\|_2^2\big]$ in Eq. (53). Since $\|(s,a)\|_2 \le 1$, we have $\|\nabla_\theta f((s,a);\theta_t)\|_2 \le 1$. Then by definition, we have the following first inequality,

$$\begin{aligned}&\mathbb{E}_{\nu_\pi}\Big[\Big(\delta_t^\theta - \delta_t^0\Big)^2 \Big\|\nabla_\theta f((s,a);\theta_t)\Big\|_2^2\Big] \\ &\le \mathbb{E}_{\nu_\pi}\Big[\Big(f\big((s,a);\theta_t\big) - \hat{f}\big((s,a);\theta_t\big) - \gamma\Big(f\big((s',a');\theta_t\big) \\ &\quad - \hat{f}\big((s',a');\theta_t\big)\Big)\Big)^2\Big] \\ &\le \mathbb{E}_{\nu_\pi}\Big[\Big(\Big|f\big((s,a);\theta_t\big) - \hat{f}\big((s,a);\theta_t\big)\Big| + \Big|f\big((s',a');\theta_t\big) \\ &\quad - \hat{f}\big((s',a');\theta_t\big)\Big|\Big)^2\Big]\end{aligned}$$

$$
\begin{aligned}
&\leq 2\mathbb{E}_{\nu_\pi}\left[\left(f\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_t\big)\right)^2\right]+2\mathbb{E}_{\nu_\pi}\\
&\quad\left[\left(f\big((s',a');\theta_t\big)-\hat{f}\big((s',a');\theta_t\big)\right)^2\right]\\
&\leq 4\mathbb{E}_{\nu_\pi}\left[\left(f\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_t\big)\right)^2\right]\leq\frac{16c_1\Upsilon^3}{\sqrt{m}}
\end{aligned}
\tag{54}
$$

We obtain the second inequality by $|\gamma|\leq 1$, *then obtain the third inequality by the fact that* $(A+B)^2\leq 2A^2+2B^2$. We reach the final step by inserting Lemma 14. We then proceed to upper bound $\mathbb{E}_{\nu_\pi}\left[|\delta_t^0|\|\nabla_\theta f((s,a);\theta_t)-\nabla_\theta\hat{f}((s,a);\theta_t))\|_2\right]$. From Hölder's inequality, we have,

$$
\begin{aligned}
&\mathbb{E}_{\nu_\pi}\left[\left(|\delta_t^0|\|\nabla_\theta f((s,a);\theta_t)-\nabla_\theta\hat{f}((s,a);\theta_t))\|_2\right)^2\right]\\
&\leq\mathbb{E}_{\nu_\pi}\left[(\delta_t^0)^2\right]\mathbb{E}_{\nu_\pi}\left[\|\nabla_\theta f((s,a);\theta_t)-\nabla_\theta\hat{f}((s,a);\theta_t))\|_2^2\right]
\end{aligned}
\tag{55}
$$

We first derive an upper bound for first term in Eq. (55), starting from its definition,

$$
\begin{aligned}
&\mathbb{E}_{\nu_\pi}\left[(\delta_t^0)^2\right]\\
&=\mathbb{E}_{\nu_\pi}\left[\left[\hat{f}\big((s',a');\theta_t\big)-\gamma\hat{f}\big((s,a);\theta_t\big)-r_{s,a}\right]^2\right]\\
&\leq 3\mathbb{E}_{\nu_\pi}\left[\left(\hat{f}\big((s',a');\theta_t\big)\right)^2\right]+3\mathbb{E}_{\nu_\pi}\left[\left(\gamma\hat{f}\big((s,a);\theta_t\big)\right)^2\right]\\
&\quad+3\mathbb{E}_{\nu_\pi}\left[r_{s,a}^2\right]\\
&\leq 6\mathbb{E}_{\nu_\pi}\left[\left(\hat{f}\big((s,a);\theta_t\big)\right)^2\right]+3r_{\max}^2\\
&=6\mathbb{E}_{\nu_\pi}\left[\left(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)+\hat{f}\big((s,a);\theta_{\pi^*}\big)\right.\right.\\
&\quad\left.\left.-Q_\pi+Q_\pi\right)^2\right]+3r_{\max}^2\\
&\leq 18\mathbb{E}_{\nu_\pi}\left[\left(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)^2\right]+18\mathbb{E}_{\nu_\pi}\\
&\quad\left[\left(\hat{f}\big((s,a);\theta_{\pi^*}\big)-Q_\pi\right)^2\right]+18\mathbb{E}_{\nu_\pi}\left[\left(Q_\pi\right)^2\right]+3r_{\max}^2\\
&\leq 72\Upsilon^2+18\mathbb{E}_{\nu_\pi}\left[\left(\hat{f}\big((s,a);\theta_{\pi^*}\big)-Q_\pi\right)^2\right]\\
&\quad+21(1-\gamma)^{-2}r_{\max}^2
\end{aligned}
\tag{56}
$$

We obtain the first and the third inequality by the fact that $(A+B+C)^2\leq 3A^2+3B^2+3C^2$. Recall $r_{\max}$ is the boundary for reward function r, which leads to the second inequality. We obtain the last inequality in Eq. (56) following the fact that $|\hat{f}((s,a);\theta_t)-\hat{f}((s,a);\theta_{\pi^*})|\leq\|\theta_t-\theta_{\pi^*}\|\leq 2\Upsilon$ and $Q_\pi\leq(1-\gamma)^{-1}r_{\max}$. Since $\bar{\mathcal{F}}_{\Upsilon,m}\subset\mathcal{F}_{\Upsilon,m}$, by Lemma 13, we have,

$$
E_{\nu_\pi}\left[\left(\hat{f}\big((s,a);\theta_{\pi^*}\big)-Q_\pi\right)^2\right]\leq\frac{\Upsilon^2\left(1+\sqrt{2\log(1/\delta)}\right)^2}{m}
\tag{57}
$$

Combine Eq. (56) and Eq. (57), we have with probability of $1-\delta$,

$$\begin{aligned}&\mathbb{E}_{\nu_\pi}\left[(\delta_t^0)^2\right]\\ &\leq 72\Upsilon^2(1+\frac{\log(1/\delta)}{m})+21(1-\gamma)^{-2}r_{\max}^2\end{aligned} \tag{58}$$

Lastly we have

$$\begin{aligned}&\mathbb{E}_{\nu_\pi}\left[\|\nabla_\theta f((s,a);\theta_t)-\nabla_\theta \hat{f}((s,a);\theta_t))\|_2^2\right]\\ &=\mathbb{E}_{\nu_\pi}\Big[\Big(\frac{1}{m}\sum_{v=1}^{m}\big(\mathbb{1}\{[\theta]_v^\top(s,a)>0\}-\mathbb{1}\{[\Theta_{\text{init}}]_v^\top(s,a)\\ &>0\}\big)^2(b_v)^2\|(s,a)\|_2^2\Big)\Big]\\ &\leq\mathbb{E}_{\nu_\pi}\Big[\frac{1}{m}\sum_{v=1}^{m}\big(\mathbb{1}\{|[\Theta_{\text{init}}]_v^\top(s,a)|\leq\|[\theta]_v-[\Theta_{\text{init}}]_v\|_2\}\big)\Big]\\ &\leq\frac{c_1\Upsilon}{\sqrt{m}}\end{aligned} \tag{59}$$

We obtain the first inequality by following Eq. (50) and the fact that $|b_v|\leq 1$ and $\|(s,a)\|_2\leq 1$. Then for the rest, we follow the similar argument in Eq. (52). To finish the proof, we plug Eq. (54), Eq. (58) and Eq. (59) back to Eq. (53),

$$\begin{aligned}&\|\bar{g}_t-\hat{g}_t\|_2^2\\ &\leq 2\Big(\frac{16c_1\Upsilon^3}{\sqrt{m}}+\Big(72\Upsilon^2(1+\frac{\log(1/\delta)}{m})+21(1-\gamma)^{-2}r_{\max}^2\Big)\\ &\frac{c_1\Upsilon}{\sqrt{m}}\Big)\\ &=\frac{176c_1\Upsilon^3}{\sqrt{m}}+\frac{144c_1\Upsilon^3\log(1/\delta)}{m^{3/2}}+\frac{42c_1\Upsilon r_{\max}^2}{(1-\gamma)^{-2}\sqrt{m}}\end{aligned}$$

Then we have,

$$\begin{aligned}&\|\bar{g}_t-\hat{g}_t\|_2\\ &\leq\sqrt{\frac{176c_1\Upsilon^3}{\sqrt{m}}+\frac{144c_1\Upsilon^3\log(1/\delta)}{m^{3/2}}+\frac{42c_1\Upsilon r_{\max}^2}{(1-\gamma)^{-2}\sqrt{m}}}\\ &\leq\sqrt{\frac{176c_1\Upsilon^3}{\sqrt{m}}}+\sqrt{\frac{144c_1\Upsilon^3\log(1/\delta)}{m^{3/2}}}+\sqrt{\frac{42c_1\Upsilon r_{\max}^2}{(1-\gamma)^{-2}\sqrt{m}}}\\ &=\mathcal{O}\Big(\Upsilon^{3/2}m^{-1/4}\big(1+(m\log\frac{1}{\delta})^{-1/2}\big)+\Upsilon^{1/2}r_{\max}m^{-1/4}\Big)\end{aligned}$$

Next, we provide the following lemma to characterize the variance of g_t.

Lemma 16 *(Variance of the Stochastic Update Vector)* [31]. *There exists a constant $\xi_g^2=\mathcal{O}(\Upsilon^2)$ independent of t. Such that for any $t\leq T$, it holds that*

$$\mathbb{E}_{\nu_\pi}[\|g_t(\theta_t)-\bar{g}_t(\theta_t)\|_2^2]\leq\xi_g^2$$

A detailed proof can be found in [31]. Now we provide the proof for Lemma 2.

Proof.

$$\begin{aligned}
&\left\|\theta_{t+1}-\theta_{\pi^*}\right\|_2^2 \\
&=\left\|\Pi_{\mathcal{D}}(\theta_t-\eta g_t(\theta_t))-\Pi_{\mathcal{D}}(\theta_{\pi^*}-\eta\hat{g}_t(\theta_{\pi^*}))\right\|_2^2 \\
&\leq\left\|(\theta_t-\theta_{\pi^*})-\eta\big(g_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\big)\right\|_2^2 \\
&=\left\|\theta_t-\theta_{\pi^*}\right\|_2^2-2\eta\big(g_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\big)^\top\big(\theta_t-\theta_{\pi^*}\big) \\
&\quad+\eta^2\left\|g_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\right\|_2^2
\end{aligned} \tag{60}$$

The inequality holds due to the definition of $\Pi_{\mathcal{D}}$. We first upper bound $\left\|g_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\right\|_2^2$ in Eq. (60),

$$\begin{aligned}
&\left\|g_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\right\|_2^2 \\
&=\left\|g_t(\theta_t)-\bar{g}_t(\theta_t)+\bar{g}_t(\theta_t)-\hat{g}_t(\theta_t)+\hat{g}_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\right\|_2^2 \\
&\leq 3\Big(\left\|g_t(\theta_t)-\bar{g}_t(\theta_t)\right\|_2^2+\left\|\bar{g}_t(\theta_t)-\hat{g}_t(\theta_t)\right\|_2^2+ \\
&\quad\left\|\hat{g}_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\right\|_2^2\Big)
\end{aligned} \tag{61}$$

The inequality holds due to fact that $(A+B+C)^2\leq 3A^2+3B^2+3C^2$. Two of the terms on the right hand side of Eq. (61) are characterized in Lemma 15 and Lemma 16. We therefore characterize the remaining term,

$$\begin{aligned}
&\left\|\hat{g}_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\right\|_2^2 \\
&=\mathbb{E}_{\nu_\pi}\left[\big(\delta_t^0(\theta_t)-\delta_t^0(\theta_{\pi^*})\big)^2\left\|\nabla_\theta\hat{f}\big((s,a);\theta_t\big)\right\|_2^2\right] \\
&\leq\mathbb{E}_{\nu_\pi}\Bigg[\Bigg(\Big(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)-\gamma\Big(\hat{f}\big((s',a');\\
&\quad\theta_t\big)-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)\Bigg)^2\Bigg] \\
&\leq\mathbb{E}_{\nu_\pi}\left[\Big(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)^2\right]+2\gamma\mathbb{E}_{\nu_\pi} \\
&\quad\left[\Big(\hat{f}\big((s',a');\theta_t\big)-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)\Big(\hat{f}\big((s,a);\theta_t\big)\right. \\
&\quad\left.-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)\right] \\
&\quad+\gamma^2\mathbb{E}_{\nu_\pi}\left[\Big(\hat{f}\big((s',a');\theta_t\big)-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)^2\right]
\end{aligned} \tag{62}$$

We obtain the first inequality by the fact that $\|\nabla_\theta\hat{f}((s,a);\theta_t)\|_2\leq 1$. Then we use the fact that (s,a) and (s',a') have the same marginal distribution as well as $\gamma<1$ for the second inequality. Follow the Cauchy-Schwarz inequality and

the fact that (s,a) and (s',a') have the same marginal distribution, we have

$$\begin{aligned}
&\mathbb{E}_{\nu_\pi}\Big[\Big(\hat{f}\big((s',a');\theta_t\big)-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)\Big(\hat{f}\big((s,a);\theta_t\big)\\
&-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)\Big]\\
\le&\,\mathbb{E}_{\nu_\pi}\Big[\Big(\hat{f}\big((s',a');\theta_t\big)-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)\Big]\mathbb{E}_{\nu_\pi}\\
&\Big[\Big(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)\Big]\\
=&\,\mathbb{E}_{\nu_\pi}\Big[\Big(\hat{f}\big((s',a');\theta_t\big)-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)^2\Big] \qquad (63)
\end{aligned}$$

We plug Eq. (63) back to Eq. (62),

$$\begin{aligned}
&\big\|\hat{g}_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\big\|_2^2\\
\le&\,(1+\gamma)^2\mathbb{E}_{\nu_\pi}\Big[\Big(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)^2\Big]. \qquad (64)
\end{aligned}$$

Next, we upper bound $\big(g_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\big)^\top\big(\theta_t-\theta_{\pi^*}\big)$. We have,

$$\begin{aligned}
&\big(g_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\big)^\top\big(\theta_t-\theta_{\pi^*}\big)\\
=&\,\big(g_t(\theta_t)-\bar{g}_t(\theta_t))\big)^\top\big(\theta_t-\theta_{\pi^*}\big)+\big(\bar{g}_t(\theta_t)-\hat{g}_t(\theta_t)\big)^\top\\
&\big(\theta_t-\theta_{\pi^*}\big)+\big(\hat{g}_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\big)^\top\big(\theta_t-\theta_{\pi^*}\big) \qquad (65)
\end{aligned}$$

One term on the right hand side of Eq. (65) are characterized by Lemma 16. We continue to characterize the remaining terms. First, by Hölder's inequality, we have

$$\begin{aligned}
&\big(\bar{g}_t(\theta_t)-\hat{g}_t(\theta_t)\big)^\top\big(\theta_t-\theta_{\pi^*}\big)\\
\ge&-\big\|\bar{g}_t(\theta_t)-\hat{g}_t(\theta_t)\big\|_2\big\|\theta_t-\theta_{\pi^*}\big\|_2\\
\ge&-2\Upsilon\big\|\bar{g}_t(\theta_t)-\hat{g}_t(\theta_t)\big\|_2 \qquad (66)
\end{aligned}$$

We obtain the second inequality since $\big\|\theta_t-\theta_{\pi^*}\big\|_2\le 2\Upsilon$ by definition. For the last term,

$$\begin{aligned}
&\big(\hat{g}_t(\theta_t)-\hat{g}_t(\theta_{\pi^*})\big)^\top\big(\theta_t-\theta_{\pi^*}\big)\\
=&\,\mathbb{E}_{\nu_\pi}\Big[\Big(\Big(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)-\gamma\Big(\hat{f}\big((s',a');\theta_t\big)\\
&-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)\Big)\Big(\nabla_\theta\hat{f}\big((s,a);\theta_t\big)\Big)^\top\Big(\theta_t-\theta_{\pi^*}\Big)\Big]\\
=&\,\mathbb{E}_{\nu_\pi}\Big[\Big(\Big(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)-\gamma\Big(\hat{f}\big((s',a');\theta_t\big)\\
&-\hat{f}\big((s',a');\theta_{\pi^*}\big)\Big)\Big)\Big(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\Big)\Big]
\end{aligned}$$

$$
\begin{aligned}
&\geq \mathbb{E}_{\nu_\pi}\left[\left(\left(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)\right)^2\right] \\
&\quad-\gamma\mathbb{E}_{\nu_\pi}\left[\left(\left(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)\right)^2\right] \\
&=(1-\gamma)\mathbb{E}_{\nu_\pi}\left[\left(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)^2\right], \qquad (67)
\end{aligned}
$$

where the inequality follows from Eq. (63). Combine Eqs. (60), (61), (64), (65), (66) and (67), we have,

$$
\begin{aligned}
&\left\|\theta_{t+1}-\theta_{\pi^*}\right\|_2^2 \\
&\leq \left\|\theta_t-\theta_{\pi^*}\right\|_2^2-\left(2\eta(1-\gamma)-3\eta^2(1+\gamma)^2\right) \\
&\quad\mathbb{E}_{\nu_\pi}\left[\left(\hat{f}\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)^2\right] \\
&\quad+3\eta^2\|\bar{g}_t-\hat{g}_t\|_2^2+4\eta\Upsilon\|\bar{g}_t-\hat{g}_t\|_2+4\Upsilon\eta|\xi_g| \\
&\quad+3\eta^2\xi_g^2 \qquad (68)
\end{aligned}
$$

We then bound the error terms by rearrange Eq. (68). First, we have, with probability of $1-\delta$,

$$
\begin{aligned}
&\mathbb{E}_{\nu_\pi}\left[\left(f\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)^2\right] \\
&=\mathbb{E}_{\nu_\pi}\left[\left(f\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_t\big)+\hat{f}\big((s,a);\theta_t\big)\right.\right. \\
&\quad\left.\left.-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)^2\right] \\
&\leq 2\mathbb{E}_{\nu_\pi}\left[\left(f\big((s,a);\theta_t\big)-\hat{f}\big((s,a);\theta_t\big)\right)^2+\left(\hat{f}\big((s,a);\theta_t\big)\right.\right. \\
&\quad\left.\left.-\hat{f}\big((s,a);\theta_{\pi^*}\big)\right)^2\right] \\
&\leq\left(\eta(1-\gamma)-1.5\eta^2(1+\gamma)^2\right)^{-1}\left(\left\|\theta_t-\theta_{\pi^*}\right\|_2^2\right. \\
&\quad\left.-\left\|\theta_{t+1}-\theta_{\pi^*}\right\|_2^2+4\Upsilon\eta|\xi_g|+3\eta^2\xi_g^2\right)+\epsilon_g \qquad (69)
\end{aligned}
$$

where

$$
\begin{aligned}
\epsilon_g &= \mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta)+\Upsilon^{5/2}m^{-1/4}\sqrt{\log(1/\delta)} \\
&\quad+\Upsilon r_{\max}^2 m^{-1/4})
\end{aligned}
$$

We obtain the first inequality by the fact that $(A+B)^2\leq 2A^2+2B^2$. Then by Eq. (68), Lemma 14 and Lemma 15, we reach the final inequality. By telescoping Eq. (69) for $t=$ to T, we have, with probability of $1-\delta$,

$$
\begin{aligned}
&\left\|f\big((s,a);\theta_T\big) - \hat{f}\big((s,a);\theta_{\pi^*}\big)\right\|^2 \\
\leq &\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\nu_\pi}\left[\left(f\big((s,a);\theta_t\big) - \hat{f}\big((s,a);\theta_{\pi^*}\big)\right)^2\right] \\
\leq &T^{-1}\big(2\eta(1-\gamma) - 3\eta^2(1+\gamma)^2\big)^{-1}(\|\Theta_{\text{init}} - \theta_{\pi^*}\| + \\
&4\Upsilon T\eta|\xi_g| + 3T\eta^2\xi_g^2) + \epsilon_g
\end{aligned}
$$

Set $\eta = \min\{1/\sqrt{T}, (1-\gamma)/3(1+\gamma)^2\}$, which implies that $T^{-1/2}(2\eta(1-\gamma) - 3\eta^2(1+\gamma)^2)^{-1} \leq 1/(1-\gamma)^2$, then we have, with probability of $1-\delta$,

$$
\begin{aligned}
&\left\|f\big((s,a);\theta_T\big) - \hat{f}\big((s,a);\theta_{\pi^*}\big)\right\| \\
\leq &\frac{1}{(1-\gamma)^2\sqrt{T}}\big(\|\Theta_{\text{init}} - \theta_{\pi^*}\|_2^2 + 4\Upsilon\sqrt{T}|\xi_g| \\
&+ 3\xi_g^2\big) + \epsilon_g \\
\leq &\frac{\Upsilon^2 + 4\Upsilon\sqrt{T}|\xi_g| + 3\xi_g^2}{(1-\gamma)^2\sqrt{T}} + \epsilon_g \\
= &\mathcal{O}(\Upsilon^3 m^{-1/2}\log(1/\delta) + \Upsilon^{5/2}m^{-1/4}\sqrt{\log(1/\delta)} \\
&+ \Upsilon r_{\max}^2 m^{-1/4} + \Upsilon^2 T^{-1/2} + \Upsilon)
\end{aligned}
$$

We obtain the second inequality by the fact that $\|\Theta_{\text{init}} - \theta_{\pi^*}\|_2 \leq \Upsilon$. Then by definition we replace $\tilde{Q}_{\omega_k}$ and $\tilde{Q}_{\pi_k}$

E Additional Related Work

E.1 Global Optimality of Policy Search Methods

A major challenge of existing RL research is the lack of theoretical justification, such as sample complexity analysis, mainly because the objective function of policy search in RL is often nonconvex. It is challenging to determine if a policy search approach is guaranteed to reach the global optimal. Besides, the RL architecture components are usually parameterized by neural networks in practice. Its nonlinearity and complex nature render the analysis significantly difficult [62].

The theoretical understanding of policy gradient methods is also under tentative study. Work on this topic has been done mostly in tabular and linear parametrization settings for different variants of policy gradient. For example, [11] and [44] establish a non-asymptotic convergence guarantee for natural policy gradient (NPG, [22]) and trusted region policy optimization (TRPO, [42]), respectively. [35] show converge rate for softmax parametrization, while [1] analyze multiple variants of policy gradient. On the other side of the spectrum, [31,51] prove the global convergence and optimality of various policy gradient algorithms with over-parameterized neural networks. Furthermore, [62] apply the

global optimality analysis to variance-constrained actor-critic risk-averse control with cumulative average rewards, and proposed a corresponding variance-constrained actor-critic (VARAC) algorithm. However, the analysis procedure is complicated due to the risk constraints on cumulative rewards, and the algorithm's experimental performance remains unverified. Therefore, it remains interesting if there can be simplified global optimality analysis with verifiable experimental studies for risk-averse policy search methods.

E.2 Over-Parameterized Neural Networks in RL

Overparameterization, a technique of deploying more parameters than necessary, improves the performance of neural networks [59]. The learning ability and generalization of over-parameterized neural networks have been studied extensively [2,5,15]. Integration with over-parameterized neural networks can be found in multiple RL topics. One line of work is to prove the global optimality of RL algorithms in a non-linear approximation setting [31,51,62]. They use ReLU activation over-parameterized neural networks with policy gradient methods such as NPG and PPO. Our work also belongs to this category. Other works include [19], which also deploy a two-layered ReLU activation over-parameterized neural network on mean-field multi-agent reinforcement learning problem. Regularization with over-parameterized neural networks is also investigated recently [25,41].

References

1. Agarwal, A., Kakade, S.M., Lee, J.D., Mahajan, G.: On the theory of policy gradient methods: optimality, approximation, and distribution shift. J. Mach. Learn. Res. **22**(98), 1–76 (2021)
2. Allen-Zhu, Z., Li, Y., Liang, Y.: Learning and generalization in overparameterized neural networks, going beyond two layers. In: Advances in Neural Information Processing Systems 32 (2019)
3. Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization (2019)
4. Antos, A., Szepesvári, C., Munos, R.: Fitted Q-iteration in continuous action-space MDPs. In: Advances in Neural Information Processing Systems 20 (2007)
5. Arora, S., Du, S., Hu, W., Li, Z., Wang, R.: Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In: International Conference on Machine Learning, pp. 322–332. PMLR (2019)
6. Bhandari, J., Russo, D.: Global optimality guarantees for policy gradient methods. arXiv preprint arXiv:1906.01786 (2019)
7. Bisi, L., Sabbioni, L., Vittori, E., Papini, M., Restelli, M.: Risk-averse trust region optimization for reward-volatility reduction. In: Bessiere, C. (ed.) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-2020, pp. 4583–4589. International Joint Conferences on Artificial Intelligence Organization, July 2020. Special Track on AI in FinTech
8. Brockman, G., et al.: OpenAI gym. arXiv preprint arXiv:1606.01540 (2016)
9. Cai, Q., Yang, Z., Lee, J.D., Wang, Z.: Neural temporal-difference and Q-learning provably converge to global optima. arXiv preprint arXiv:1905.10027 (2019)

10. Cao, Y., Gu, Q.: Generalization error bounds of gradient descent for learning over-parameterized deep ReLU networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 3349–3356 (2020)
11. Cen, S., Cheng, C., Chen, Y., Wei, Y., Chi, Y.: Fast global convergence of natural policy gradient methods with entropy regularization. Oper. Res. **70**(4), 2563–2578 (2021)
12. Csiszár, I., Körner, J.: Information Theory: Coding Theorems for Discrete Memoryless Systems. Cambridge University Press, Cambridge (2011)
13. Dabney, W., et al.: A distributional code for value in dopamine-based reinforcement learning. Nature **577**(7792), 671–675 (2020)
14. Di Castro, D., Tamar, A., Mannor, S.: Policy gradients with variance related risk criteria. arXiv preprint arXiv:1206.6404 (2012)
15. Du, S.S., Zhai, X., Poczos, B., Singh, A.: Gradient descent provably optimizes over-parameterized neural networks. arXiv preprint arXiv:1810.02054 (2018)
16. Farahmand, A.M., Ghavamzadeh, M., Szepesvári, C., Mannor, S.: Regularized policy iteration with nonparametric function spaces. J. Mach. Learn. Res. **17**(1), 4809–4874 (2016)
17. Fu, Z., Yang, Z., Wang, Z.: Single-timescale actor-critic provably finds globally optimal policy. arXiv preprint arXiv:2008.00483 (2020)
18. Garcıa, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. J. Mach. Learn. Res. **16**(1), 1437–1480 (2015)
19. Gu, H., Guo, X., Wei, X., Xu, R.: Mean-field multi-agent reinforcement learning: a decentralized network approach. arXiv preprint arXiv:2108.02731 (2021)
20. Hans, A., Schneegaß, D., Schäfer, A.M., Udluft, S.: Safe exploration for reinforcement learning. In: ESANN, pp. 143–148. Citeseer (2008)
21. Kakade, S., Langford, J.: Approximately optimal approximate reinforcement learning. In: In Proceedings of the 19th International Conference on Machine Learning. Citeseer (2002)
22. Kakade, S.M.: A natural policy gradient. In: Advances in Neural Information Processing Systems 14 (2001)
23. Konstantopoulos, T., Zerakidze, Z., Sokhadze, G.: Radon-Nikodým theorem. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science, pp. 1161–1164. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-04898-2_468
24. Kovács, B.: Safe reinforcement learning in long-horizon partially observable environments (2020)
25. Kubo, M., Banno, R., Manabe, H., Minoji, M.: Implicit regularization in over-parameterized neural networks. arXiv preprint arXiv:1903.01997 (2019)
26. La, P., Ghavamzadeh, M.: Actor-critic algorithms for risk-sensitive MDPs. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 26. Curran Associates, Inc. (2013)
27. Lai, T.L., Xing, H., Chen, Z.: Mean-variance portfolio optimization when means and covariances are unknown. Ann. Appl. Stat. **5**(2A), June 2011. https://doi.org/10.1214/10-aoas422
28. Laroche, R., Tachet des Combes, R.: Dr Jekyll and Mr Hyde: the strange case of off-policy policy updates. In: Advances in Neural Information Processing Systems 34 (2021)
29. Li, D., Ng, W.L.: Optimal dynamic portfolio selection: multiperiod mean-variance formulation. Math. Financ. **10**(3), 387–406 (2000)

30. Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., Petrik, M.: Finite-sample analysis of proximal gradient TD algorithms. In: Proceedings of the Conference on Uncertainty in AI (UAI), pp. 504–513 (2015)
31. Liu, B., Cai, Q., Yang, Z., Wang, Z.: Neural trust region/proximal policy optimization attains globally optimal policy. In: Advances in Neural Information Processing Systems 32 (2019)
32. Majumdar, A., Pavone, M.: How should a robot assess risk? Towards an axiomatic theory of risk in robotics. In: Amato, N.M., Hager, G., Thomas, S., Torres-Torriti, M. (eds.) Robotics Research. SPAR, vol. 10, pp. 75–84. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-28619-4_10
33. Mannor, S., Tsitsiklis, J.: Mean-variance optimization in Markov decision processes. arXiv preprint arXiv:1104.5601 (2011)
34. Markowitz, H.M., Todd, G.P.: Mean-Variance Analysis in Portfolio Choice and Capital Markets, vol. 66. Wiley, New York (2000)
35. Mei, J., Xiao, C., Szepesvari, C., Schuurmans, D.: On the global convergence rates of softmax policy gradient methods. In: International Conference on Machine Learning, pp. 6820–6829. PMLR (2020)
36. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
37. Munos, R.: Performance bounds in Lp-norm for approximate value iteration. SIAM J. Control. Optim. **46**(2), 541–561 (2007)
38. Munos, R., Szepesvári, C.: Finite-time bounds for fitted value iteration. J. Mach. Learn. Res. **9**(5), 815–857 (2008)
39. Parker, D.: Managing risk in healthcare: understanding your safety culture using the Manchester patient safety framework (MaPSaF). J. Nurs. Manag. **17**(2), 218–222 (2009)
40. Rahimi, A., Recht, B.: Weighted sums of random kitchen sinks: replacing minimization with randomization in learning. In: Advances in Neural Information Processing Systems 21 (2008)
41. Satpathi, S., Gupta, H., Liang, S., Srikant, R.: The role of regularization in overparameterized neural networks. In: 2020 59th IEEE Conference on Decision and Control (CDC), pp. 4683–4688. IEEE (2020)
42. Schulman, J., Levine, S., Abbeel, P., Jordan, M., Moritz, P.: Trust region policy optimization. In: International Conference on Machine Learning, pp. 1889–1897. PMLR (2015)
43. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
44. Shani, L., Efroni, Y., Mannor, S.: Adaptive trust region policy optimization: global convergence and faster rates for regularized MDPs. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 5668–5675 (2020)
45. Sobel, M.J.: The variance of discounted Markov decision processes. J. Appl. Probab. **19**(4), 794–802 (1982)
46. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. A Bradford Book. MIT Press, Cambridge (2018)
47. Sutton, R.S., et al.: Fast gradient-descent methods for temporal-difference learning with linear function approximation. In: International Conference on Machine Learning, pp. 993–1000 (2009)
48. Thomas, G., Luo, Y., Ma, T.: Safe reinforcement learning by imagining the near future. In: Advances in Neural Information Processing Systems 34 (2021)

49. Todorov, E., Erez, T., Tassa, Y.: MuJoCo: a physics engine for model-based control. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5026–5033. IEEE (2012)
50. Vinyals, O., et al.: Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)
51. Wang, L., Cai, Q., Yang, Z., Wang, Z.: Neural policy gradient methods: global optimality and rates of convergence (2019)
52. Wang, M., Fang, E.X., Liu, H.: Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. Math. Program. **161**(1–2), 419–449 (2017)
53. Wang, W.Y., Li, J., He, X.: Deep reinforcement learning for NLP. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pp. 19–21 (2018)
54. Weng, J., Duburcq, A., You, K., Chen, H.: MuJoCo benchmark (2020). https://tianshou.readthedocs.io/en/master/tutorials/benchmark.html
55. Xie, T., et al.: A block coordinate ascent algorithm for mean-variance optimization. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper/2018/file/4e4b5fbbbb602b6d35bea8460aa8f8e5-Paper.pdf
56. Xu, P., Chen, J., Zou, D., Gu, Q.: Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In: Advances in Neural Information Processing Systems (2018)
57. Xu, T., Liang, Y., Lan, G.: CRPO: a new approach for safe reinforcement learning with convergence guarantee. In: International Conference on Machine Learning, pp. 11480–11491. PMLR (2021)
58. Yang, L., Wang, M.: Reinforcement learning in feature space: matrix bandit, kernels, and regret bound. In: International Conference on Machine Learning, pp. 10746–10756. PMLR (2020)
59. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Commun. ACM **64**(3), 107–115 (2021)
60. Zhang, S., Liu, B., Whiteson, S.: Mean-variance policy iteration for risk-averse reinforcement learning. In: AAAI Conference on Artificial Intelligence (AAAI) (2021)
61. Zhang, S., Tachet, R., Laroche, R.: Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. arXiv preprint arXiv:2111.02997 (2021)
62. Zhong, H., Fang, E.X., Yang, Z., Wang, Z.: Risk-sensitive deep RL: variance-constrained actor-critic provably finds globally optimal policy (2020)
63. Zou, D., Cao, Y., Zhou, D., Gu, Q.: Gradient descent optimizes over-parameterized deep ReLU networks. Mach. Learn. **109**(3), 467–492 (2020)