# Achieving Zero Constraint Violation for Constrained Reinforcement Learning via Conservative Natural Policy Gradient Primal-Dual Algorithm

**Qinbo Bai**[1], **Armit Singh Bedi**[2], **Vaneet Aggarwal**[1]

[1] Purdue University
[2] University of Maryland
bai113@purdue.edu, University of Maryland, vaneet@purdue.edu

## Abstract

We consider the problem of constrained Markov decision process (CMDP) in continuous state actions spaces where the goal is to maximize the expected cumulative reward subject to some constraints. We propose a novel Conservative Natural Policy Gradient Primal Dual Algorithm (CNPGPD) to achieve zero constraint violation while achieving state of the art convergence results for the objective value function. For general policy parametrization, we prove convergence of value function to global optimal upto an approximation error due to restricted policy class. We improve the sample complexity of existing constrained NPGPD algorithm. To the best of our knowledge, this is the first work to establish zero constraint violation with Natural policy gradient style algorithms for infinite horizon discounted CMDPs. We demonstrate the merits of proposed algorithm via experimental evaluations.

## Introduction

Reinforcement learning problem is formulated as a Markov Decision Process (MDP) and can be solved using different algorithms in the literature (Sutton 1988). To deal with the scalability issue to the the large state and action spaces, policy parametrization is widely used (Ding et al. 2020; Xu, Liang, and Lan 2021; Agarwal et al. 2020). The problem becomes challenging when we have constraints and is called constrained MDPs (CMDPs). The problem is popular across various application domains such as robotics, communications, computer vision, autonomous driving, etc. (Arulkumaran et al. 2017; Kiran et al. 2021). Mathematically, the problem is sequential in nature, agent observes the state, takes an action, and then transitions to next state. Further, an agent also needs to satisfy a set of constraints as well such as safety constraints, power constraints and maneuver constraints. CMDPs are challenging to solve specially in large state action spaces (Ding et al. 2020; Xu, Liang, and Lan 2021) which is the focus of our work.

The constraint violations could be catastrophic in applications such as in power systems (Vu et al. 2020) or autonomous vehicle control (Wen et al. 2020). In the literature, various algorithms are proposed to solve CMDP in large actions spaces in a model free manner (See Table 1 for comparisons). The main performance metric here is the sample complexity, which is the number of samples required to achieve $\epsilon$-optimal objective and $\epsilon$-constraint violation. However, there doesn't exist literature which gives zero violation gurantee on large state and action space. Hence, we ask this question: "*Is it possible to achieve zero constraint violations for CMDP problems in large state action spaces while solving in a model free manner?*"

We answer this question in an affirmative sense in this work. We proposed a novel Conservative Natural Policy Gradient Primal Dual Algorithm (C-NPG-PDA) in this paper. We utilize a novel idea of conservative constraints to policy gradient algorithms and establish convergence guarantees of global optima for general policy parametrization. Our contributions are summarized as follows.

- We propose a Natural Policy Gradient algorithm which achieves **zero constraint violation** for constrained MDPs in large state and action space. The proposed algorithm also converges to the neighborhood of the global optima with general parametrization. It is challenging to establish the zero violation result with general parametrization due to the lack of strong duality, and hence we perform a novel analysis to establish a bound between the conservative and original problems.
- We show that even if we don't utilize the conservative idea (proposed in this work), we are able to improve the sample complexity from $\mathcal{O}\left(\frac{1}{\epsilon^6}\right)$ (Ding et al. 2020)[Theorem 3] to $\mathcal{O}\left(\frac{1}{\epsilon^4}\right)$. To achieve this, we utilize the first order stationary result from (Liu et al. 2020) to bound the NPG update direction. However, due to the introduction of the constraint and the Lagrange function, the update of the Lagrange multiplier needs to be considered in the analysis.
- We perform initial proof of concepts experiments of the proposed algorithm with a random CMDP model and validate the convergence of the objective, with zero constraint violations.

[1] The detailed dependence on $(1-\gamma)$ is not shown in the original paper.

[2] In (Xu, Liang, and Lan 2021), the authors used a two layer

| Parametrization | Algorithm | Sample Complexity | Constraint violation | Generative Model |
|---|---|---|---|---|
| Softmax | PMD-PD (Liu et al. 2021) | $\mathcal{O}\left(1/\epsilon^3\right)$ [1] | Zero | No |
|  | PD-NAC (Zeng, Doan, and Romberg 2022) | $\mathcal{O}\left(1/\epsilon^6\right)$ [1] | Zero | No |
|  | NPG-PD (Ding et al. 2020) | $\mathcal{O}\left(1/(1-\gamma)^5\epsilon^2\right)$ | $\tilde{O}(\epsilon)$ | Yes |
|  | CRPO (Xu, Liang, and Lan 2021) | $\mathcal{O}\left(1/(1-\gamma)^7\epsilon^4\right)$ | $\tilde{O}(\epsilon)$ | Yes |
| General | NPG-PD (Ding et al. 2020) | $\mathcal{O}\left(1/(1-\gamma)^8\epsilon^6\right)$ | $\tilde{O}(\epsilon)$ | Yes |
|  | CRPO (Xu, Liang, and Lan 2021) | $\mathcal{O}\left(1/(1-\gamma)^{13}\epsilon^6\right)$ [2] | $\mathcal{O}(\epsilon)$ | Yes |
|  | **C-NPG-PDA (This work**, Theorem 1) | $\tilde{O}\left(1/(1-\gamma)^6\epsilon^4\right)$ | **Zero** | **No** |
| Lower bound | (Vaswani, Yang, and Szepesvári 2022) | $\tilde{\Omega}\left(1/(1-\gamma)^5\epsilon^2\right)$ | Zero | N/A |

Table 1: This table summarizes the different state of the art policy-based algorithms available in the literature with softmax or general Parametrization for CMDPs. We note that the proposed algorithm in this work is able to achieve the best sample complexity among them all while achieving zero constraint violation as well.

## Related Work

**Policy Gradient for Reinforcement Learning:** Reinforcement Learning algorithms can be divided into policy-based or value-based algorithm. Thanks to the Policy Gradient Theorem (Sutton et al. 2000), it is possible to obtain the gradient ascent direction for the standard reinforcement learning with the policy parameterization. However, in general, the objective in the reinforcement learning is non-convex with respective to the parameters (Agarwal et al. 2020), which makes the theory of global convergence difficult to derive and previous works (Papini et al. 2018; Xu, Gao, and Gu 2020a,b) are focused on the first order convergence. Recently, there is a line of interest on the global convergence result for reinforcement learning. The authors in (Zhang et al. 2020) apply the idea of escaping saddle points to the policy gradient and prove the convergence to the local optima. Further, authors in (Agarwal et al. 2020) provide provable global convergence result for direct parameterization and softmax parameterization with convergence rate $\mathcal{O}(1/\sqrt{T})$ and sample complexity $\mathcal{O}(1/\epsilon^6)$ in the tabular setting. For the restrictive policy parameterization setting, they propose a variant of NPG, Q-NPG and analyze the global convergence result with the function approximation error for both NPG and Q-NPG. (Mei et al. 2020) improves the convergence rate for policy gradient with softmax parameterization from $\mathcal{O}(1/\sqrt{t})$ to $\mathcal{O}(1/t)$ and shows a significantly faster linear convergence rate $\mathcal{O}(\exp(-t))$ for the entropy regularized policy gradient. However, no sample complexity result is achievable because policy evaluation has not been considered. With actor-critic method (Konda and Tsitsiklis 2000), (Wang et al. 2019) establishes the global optimal result for neural policy gradient method. (Liu et al. 2020) proposes a general framework of the analysis for policy gradient type of algorithms and gives the sample complexity for PG, NPG and the variance reduced version of them.

**Policy Gradient for Constrained Reinforcement Learning:** Although there exists quite a few studies for the un-constrained reinforcement learning problems, the research for the constrained setting is in its infancy and summarized in Table 1. The most famous method for the constrained problem is to use a primal-dual based algorithm. With the softmax-parametrization, (Liu et al. 2021) proposed policy mirror descent-primal dual (PMD-PD) algorithm to achieve zero constraint violation and achieve $\mathcal{O}(1/\epsilon^3)$ sample complexity. (Zeng, Doan, and Romberg 2022) proposed an Online Primal-Dual Natural Actor-Critic Algorithm and achieves zero constraint violation with $\mathcal{O}(1/\epsilon^6)$ sample complexity without the generative model. (Ding et al. 2020) proposed a primal-dual Natural Policy Gradient algorithm for both the softmax parametrization and general parametrization. However, the sample complexity for general case in their paper is $\mathcal{O}(1/\epsilon^6)$ which is quite high. (Xu, Liang, and Lan 2021) propose a primal approach policy-based algorithm for both the softmax parametrization and function approximation case. However, none of them achieve the zero constraint violation for the general parametrization case. As seen in Table 1 we achieve the best result for sample complexity in CMDP with general parametrization while also achieving zero constraint violation.

## Problem Formulation

We consider an infinite-horizon discounted Markov Decision Process $\mathcal{M}$ defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, r, g, \gamma, \rho)$,

neural network with $m$ as the width of the neural network. Larger width gives improved function approximation while increasing sample complexity. In its Theorem 2, if we choose $m = \mathcal{O}(T^4)$, then it gives $\epsilon$-convergence to the global optima and the sample complexity is $T \cdot K_{in} = \mathcal{O}(1/(1-\gamma)^1 3\epsilon^6)$. We note that this is the best choice (for sample complexity) that gives the error as $\epsilon$.

where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action space, respectively. In this paper, we focus on large station and action space, which means that the policy parametrization may not be fully sufficient. $\mathbb{P}:\mathcal{S}\times\mathcal{A}\to[0,1]$ denotes the transition probability distribution from a state-action pair to another state. $r:\mathcal{S}\times\mathcal{A}\to\Delta^S$ denotes the reward for the agent and $g^i:\mathcal{S}\times\mathcal{A}\to[-1,1], i\in[I]$ defines the $i^{th}$ constraint function for the agent. $\gamma\in(0,1)$ is the discounted factor and $\rho:\mathcal{S}\to[0,1]$ is the initial state distribution.

Define a joint stationary policy $\pi:\mathcal{S}\to\Delta^{\mathcal{A}}$ that maps a state $s\in\mathcal{S}$ to a probability distribution of actions defined as $\Delta^{\mathcal{A}}$ with a probability assigned to each action $a\in\mathcal{A}$. At the beginning of the MDP, an initial state $s_0\sim\rho$ is given and agent makes a decision $a_0\sim\pi(\cdot|s_0)$. The agent receives its reward $r(s_0,a_0)$ and constraints $g_i(s_0,a_0), i\in[I]$. Then it moves to a new state $s_1\sim\mathbb{P}(\cdot|s_0,a_0)$. We define the reward value function $J_r(\pi)$ and constraint value function $J_{g^i}(\pi), i\in[I]$ for the agent following policy $\pi$ as a discounted sum of reward and constraints over infinite horizon

$$\begin{aligned}V_r^\pi(s)&=\mathbf{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\Bigg|s_0=s\Bigg],\\ V_{g^i}^\pi(s)&=\mathbf{E}\Bigg[\sum_{t=0}^{\infty}\gamma^t g^i(s_t,a_t)\Bigg|s_0=s\Bigg].\end{aligned}\tag{1}$$

where $a_t\sim\pi(\cdot|s_t)$ and $s_{t+1}\sim\mathbb{P}(\cdot|s_t,a_t)$. Denote $J_r^\pi$ and $J_{g^i}^\pi$ as the expected value function w.r.t. the initial distribution such as

$$\begin{aligned}J_r(\pi)&=\mathbf{E}_{s_0\sim\rho}[V_r^\pi(s_0)],\\ \text{and}\ \ J_{g^i}(\pi)&=\mathbf{E}_{s_0\sim\rho}[V_{g^i}^\pi(s_0)].\end{aligned}\tag{2}$$

The agent aims to maximize the reward value function and satisfies constraints simultaneously. Formally, the problem can be formulated as

$$\begin{aligned}\max_{\pi}\quad & J_r(\pi)\\ s.t.\quad & J_{g^i}(\pi)\geq 0,\forall i\in[I].\end{aligned}\tag{3}$$

Define $\pi^*$ as the optimal-policy for the above problem. Here, we introduce the Slater Condition, which means the above problem is strictly feasible.

**Assumption 1** (Slater Condition)**.** *There exists a $\varphi>0$ and $\bar{\pi}$ that $J_{g^i}(\bar{\pi})\geq\varphi,\forall i\in[I]$.*

## Proposed Approach

We consider a policy-based algorithm on this problem and parameterize the policy $\pi$ as $\pi_\theta$ for some parameter $\theta\in\Theta$ such as softmax parametrization or a deep neural network. In this section, we first give the form of the true gradient and introduce some properties of it. Then, we propose the Conservative Natural Policy Descent Primal-Dual Algorithm (C-NPG-PD), where the conservative idea is utilized to achieve zero constraint violation.

### Gradient of Value Function and Properties

For the analysis of the convergence for the proposed algorithm, it is necessary to establish the form of the true and its properties. Here, we utilize the Policy Gradient Theorem and write the gradient for the objective function as

$$\begin{aligned}&\nabla_\theta J_r(\pi_\theta)=\\ &\mathbf{E}_{\tau\sim p(\tau|\theta)}\Bigg[\sum_{t=0}^{\infty}\nabla_\theta\log(\pi_\theta(a_t|s_t))\Bigg(\sum_{h=t}^{\infty}\gamma^h r(s_h,a_h)\Bigg)\Bigg]\end{aligned}\tag{4}$$

The computation of the gradient is well known and the proof is removed to the Appendix for completeness. We note that the log-policy function $\log\pi_\theta(a|s)$ is also called log-likelihood function in statistics (Kay 1997) and we make the following assumption.

**Assumption 2.** *The log-likelihood function is $G$-Lipschitz and $M$-smooth. Formally,*

$$\begin{aligned}\|\nabla_\theta\log\pi_\theta(a|s)\|\leq G\quad\forall\theta\in\Theta,\forall(s,a)\in\mathcal{S}\times\mathcal{A},\\ \|\nabla_\theta\log\pi_{\theta_1}(a|s)-\nabla_\theta\log\pi_{\theta_2}(a|s)\|\leq M\|\theta_1-\theta_2\|\\ \forall\theta_1,\theta_2\in\Theta,\forall(s,a)\in\mathcal{S}\times\mathcal{A}.\end{aligned}\tag{5}$$

**Remark 1.** *The Lipschitz and smoothness properties for the log-likelihood are quite common in the field of policy gradient algorithm (Agarwal et al. 2020; Zhang et al. 2021; Liu et al. 2020). Such properties can also be verified for simple parametrization such as Gaussian policy.*

The following two lemmas give the property of the value functions and its gradient, which are useful in the convergence proof. The detailed proof can be found in Appendix.[1]

**Lemma 1.** *Under Assumption 2, both the objective function $\boldsymbol{J}_r^{\pi_\theta}$ and the constraint function $\boldsymbol{J}_{g^i}^{\pi_\theta}$ are $L_J$-smooth w.r.t. $\theta$. Formally,*

$$\|\nabla_\theta\boldsymbol{J}_r(\theta_1)-\nabla_\theta\boldsymbol{J}_r(\theta_2)\|_2\leq L_J\|\theta_1-\theta_2\|_2\quad\forall\theta_1,\theta_2\in\Theta\tag{6}$$

*where $L_J=\frac{M}{(1-\gamma)^2}+\frac{2G^2}{(1-\gamma)^3}$*

**Lemma 2.** *Under Assumption 2, both the gradient of objective function $\nabla_\theta\boldsymbol{J}_r^{\pi_\theta}$ and that of the constraint function $\nabla_\theta\boldsymbol{J}_{g^i}^{\pi_\theta}$ are bounded. Formally,*

$$\begin{aligned}\|\nabla_\theta J_r(\theta)\|_2&\leq\frac{G}{(1-\gamma)^2}\\ \|\nabla_\theta J_{g^i}(\theta)\|_2&\leq\frac{G}{(1-\gamma)^2}\quad\forall i\in[I].\end{aligned}$$

## Natural Policy Gradient Primal-Dual Method with Zero Constraint Violation

In order to achieve zero constraint violation, we consider the conservative stochastic optimization framework proposed in (Akhtar, Bedi, and Rajawat 2021) and define the conservative version of the original problem as

$$\begin{aligned}\max_{\pi}\quad & J_r(\pi)\\ s.t.\quad & J_{g^i}(\pi)\geq\kappa,\forall i\in[I]\end{aligned}\tag{7}$$

where $\kappa>0$ is the parameter to control the constraint violation which we will explicitly mention in Theorem 1 in

[1] The appendix is uploaded to https://arxiv.org/abs/2206.05850

Sec. . The idea here to achieve zero constraint violation is to consider the tighter problem to make it less possible to make violation for the original problem. Notice that it is obvious that $\kappa$ must be less than $\frac{1}{1-\gamma}$ to make the conservative problem still feasible. Combining this idea, we introduce the Natural Policy Gradient method. The NPG Method utilizes the Fisher information matrix defined as

$$F_\rho(\theta) = \mathbf{E}_{s\sim d_\rho^{\pi_\theta}} \mathbf{E}_{a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^T] \tag{8}$$

where $d_\rho^\pi$ is the state visitation measure defined as

$$d_\rho^\pi := (1-\gamma)\mathbf{E}_{s_0\sim\rho}\left[\sum_{t=0}^{\infty}\gamma^t \mathbf{Pr}^\pi(s_t = s|s_0)\right] \tag{9}$$

We define the Lagrange function as

$$J_L(\pi_\theta, \boldsymbol{\lambda}) = J_r(\pi_\theta) + \sum_{i\in[I]} \lambda^i(\pi_\theta) J_{g^i}(\pi_\theta) \tag{10}$$

For simplicity, we denote $J_r(\theta), J_{g^i}(\theta), J_L(\theta, \boldsymbol{\lambda})$ as the short for $J_r(\pi_\theta), J_{g^i}(\pi_\theta), J_L(\pi_\theta, \boldsymbol{\lambda})$ and the Natural Policy Gradient method is written as

$$\begin{aligned} \theta^{k+1} &= \theta^k + \eta_1 F_\rho(\theta^k)^\dagger \nabla_\theta J_L(\theta^t, \boldsymbol{\lambda}^k) \\ \lambda_i^{k+1} &= \mathcal{P}_{(0,\sigma_\lambda]}\Big(\lambda_i^k - \eta_2\big(J_{g^i}(\theta^k) - \kappa\big)\Big) \end{aligned} \tag{11}$$

For the projection of the Lagrange multiplier, we make the following assumption.

**Assumption 3.** *The Lagrange multiplier $\lambda^i$ is bounded. Formally, $\lambda^i \in (0, \Lambda], \forall i \in [I]$*

**Remark 2.** *For the direct parametrization or softmax parametrization, it can be proved that the Lagrange multiplier is bounded by utilizing the strong duality. However, strong duality doesn't hold for the general parametrization (see Assumption 6). To prove the global convergence the boundedness of the Lagrange multiplier is required in the convergence analysis of the global convergence (Zeng et al. 2021), and a similar assumption has been made in (Zeng et al. 2021).*

We note that the pseudo-inverse of the Fisher information matrix is difficult to calculate. However, the NPG update direction can be related to the compatible function approximation error defined as

$$\begin{aligned} &L_{d_\rho^\pi,\pi}(\omega, \theta, \boldsymbol{\lambda}) = \mathbf{E}_{s\sim d_\rho^\pi}\mathbf{E}_{a\sim\pi(\cdot|s)} \\ &\left[\left(\nabla_\theta \log \pi_\theta(a|s)\cdot(1-\gamma)\omega - A_{L,\boldsymbol{\lambda}}^{\pi_\theta}(s,a)\right)^2\right] \end{aligned} \tag{12}$$

Given a fixed $\boldsymbol{\lambda}^k$ and $\theta^k$, it can be proved that the minimizer $\omega_*^k$ of $L_{d_\rho^\pi,\pi}(\omega, \theta^k, \boldsymbol{\lambda}^k)$ is exact the NPG update direction. Thus, it is possible to utilize the Stochastic Gradient Descent (SGD) algorithm to achieve the minimizer $\omega_*^k$. The gradient of $L_{d_\rho^\pi,\pi}(\omega, \theta^k, \boldsymbol{\lambda}^k)$ can be computed as

$$\begin{aligned} &\nabla_\omega L_{d_\rho^\pi,\pi}(\omega, \theta^k, \boldsymbol{\lambda}^k) = 2(1-\gamma)\nabla_\theta \log \pi_\theta^k(a|s)\cdot\mathbf{E}_{s\sim d_\rho^\pi} \\ &\quad \mathbf{E}_{a\sim\pi(\cdot|s)}\left[\nabla_\theta \log \pi_\theta^k(a|s)\cdot(1-\gamma)\omega - A_{L,\boldsymbol{\lambda}^k}^{\pi_\theta^k}(s,a)\right] \end{aligned} \tag{13}$$

Where $A_{L,\boldsymbol{\lambda}^k}^{\pi_\theta^k}(s,a)$ is the advantage function for the Lagrange function and is defined as

$$\begin{aligned} A_{L,\boldsymbol{\lambda}^k}^{\pi_\theta^k}(s,a) &= \left[Q_r^{\pi_\theta^k}(s,a) - V_r^{\pi_\theta^k}(s)\right] \\ &\quad + \sum_{i\in[I]}\lambda_k^i\left[Q_{g^i}^{\pi_\theta^k}(s,a) - V_{g^i}^{\pi_\theta^k}(s)\right] \end{aligned} \tag{14}$$

However, it is challenging to achieve the exact value of the advantage function and thus we estimate it as $\hat{A}_{L,\boldsymbol{\lambda}^k}^{\pi_\theta^k}(s,a)$ using the following procedure. The stochastic version of gradient can be written as

$$\begin{aligned} &\hat{\nabla}_\omega L_{d_\rho^\pi,\pi}(\omega, \theta^k, \boldsymbol{\lambda}^k) = 2(1-\gamma)\nabla_\theta \log \pi_\theta^k(a|s)\cdot \\ &\quad \left[\nabla_\theta \log \pi_\theta^k(a|s)\cdot(1-\gamma)\omega - \hat{A}_{L,\boldsymbol{\lambda}^k}^{\pi_\theta^k}(s,a)\right] \end{aligned} \tag{15}$$

Based on the stochastic version of the gradient mentioned above, we propose the Natural Gradient Descent Primal Dual with Zero Violation in Algorithm 1. In line 1, we initialize the parameter $\theta$ and Lagrange multiplier $\boldsymbol{\lambda}$. From Line 3 to Line 10, we use SGD to compute the Natural Policy gradient. From Line 11 to Line 15, we estimate an unbiased value function for constraint. Finally, in Line 16, we perform the conservative primal-dual update.

## Convergence Rate Results

Before stating the convergence result for the policy gradient algorithm, we describe the following assumptions which will be needed for the main result.

**Assumption 4.** *For all $\theta \in \mathbb{R}^d$, the Fisher information matrix induced by policy $\pi_\theta$ and initial state distribution $\rho$ satisfies*

$$\begin{aligned} F_\rho(\theta) &= \mathbf{E}_{s\sim d_\rho^{\pi_\theta}}\mathbf{E}_{a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s)\nabla_\theta \log \pi_\theta(a|s)^T] \\ &\succeq \mu_F\cdot\mathbf{I}_d \end{aligned} \tag{18}$$

*for some constant $\mu_F > 0$*

**Remark 3.** *The positive definiteness assumption is standard in the field of policy gradient based algorithms (Kakade 2001; Peters and Schaal 2008; Liu et al. 2020; Zhang et al. 2020). A common example which satisfies such assumption is Gaussian policy with mean parameterized linearly (See Appendix B.2 in (Liu et al. 2020)).*

**Assumption 5.** *Define the transferred function approximation error as below*

$$\begin{aligned} &L_{d_\rho^{\pi^*},\pi^*}(\omega, \theta, \boldsymbol{\lambda}) = \mathbf{E}_{s\sim d_\rho^{\pi^*}}\mathbf{E}_{a\sim\pi^*} \\ &\left[\left(\nabla_\theta \log \pi_\theta(a|s)\cdot(1-\gamma)\omega - A_{L,\boldsymbol{\lambda}}^{\pi_\theta}(s,a)\right)^2\right] \end{aligned} \tag{19}$$

*We assume that this error satisfies $L_{d_\rho^{\pi^*},\pi^*}(\omega_*^{\theta,\boldsymbol{\lambda}}, \theta, \boldsymbol{\lambda}) \leq \epsilon_{bias}$ for any $\theta \in \Theta, \boldsymbol{\lambda} \in \Lambda$, where $\omega_*^{\theta,\boldsymbol{\lambda}}$ is given as*

$$\begin{aligned} &\omega_*^{\theta,\boldsymbol{\lambda}} = \arg\min_\omega L_{d_\rho^{\pi^\theta},\pi^\theta}(\omega, \theta, \boldsymbol{\lambda}) = \arg\min_\omega \mathbf{E}_{s\sim d_\rho^{\pi_\theta}}\mathbf{E}_{a\sim\pi_\theta} \\ &\left[\left(\nabla_\theta \log \pi_\theta(a|s)\cdot(1-\gamma)\omega - A_{L,\boldsymbol{\lambda}}^{\pi_\theta}(s,a)\right)^2\right] \end{aligned} \tag{20}$$

Algorithm 1: **C**onservative **N**atural **G**radient **D**escent **P**rimal-**D**ual **A**lgorithm (C-NPG-PDA)

**Input**: Sample size K, SGD learning iteration $N$, Initial distribution $\boldsymbol{\rho}$. Discounted factor $\gamma$.
**Parameter**: Step-size $\eta_1$, $\eta_2$, SGD learning rate $\alpha$, Slater variable $\varphi$, Conservative variable $\kappa$
**Output**: $\theta^k$, $k \in [0, K-1]$

```
Initialize $\boldsymbol{\lambda}^1 = \mathbf{0}$, $\theta^1 = 0, \omega_0 = 0$
for $k = 1, 2, ..., K$ do
    for $n = 1, 2, ..., N$ do
        Sample $s \sim d_\rho^{\pi_{\theta^k}}$ and $a \sim \pi_{\theta^k}(\cdot|s)$
        Sample $Q^{\pi_\theta^k}(s,a)$ and $V^{\pi_\theta^k}(s)$ for reward function
        and constraint functions following Algorithm 2
        Estimate the Advantage Function $\hat{A}^{\pi_\theta^k}_{L,\boldsymbol{\lambda}^k}(s,a)$ following Eq. (14)
        Estimate SGD gradient $\hat{\nabla}_\omega L_{d_\rho^\pi,\pi}(\omega_n, \theta^k, \boldsymbol{\lambda}^k)$ following Eq. (15)
        SGD update $\omega_{n+1} = \omega_n - \alpha \cdot \hat{\nabla}_\omega L_{d_\rho^\pi,\pi}(\omega_n, \theta^k, \boldsymbol{\lambda}^k)$
    end for
    Compute NPG update direction as $\omega = \frac{1}{N}\sum_{n=1}^N \omega_n$
    for $n = 1, 2, ..., N$ do
        Sample $s \sim \rho$ and $a \sim \pi_{\theta^k}(\cdot|s)$
        Sample constraint value functions $V^{\pi_{\theta^k}}_{g^i,n}(s)$ following Algorithm 2
    end for
    Estimate expected constraint value function
    $\hat{J}_{g^i}(\pi_\theta^k) = \frac{1}{N}\sum_{n=1}^N V^{\pi_{\theta^k}}_{g^i,n}, \forall i \in [I]$
    Update the primal and dual variable as
        $\theta^{k+1} = \theta^k + \eta_1 \omega$    (16)
        $\lambda_i^{k+1} = \mathcal{P}_{[0,\sigma_\lambda]}\Big(\lambda_i^k - \eta_2(\hat{J}_{g^i}(\pi_{\theta^k}) - \kappa)\Big), \forall i \in [I]$    (17)
end for
```

*It can be shown that $\omega_*^\theta$ is the exact Natural Policy Gradient (NPG) update direction.*

**Remark 4.** *By Eq.* (19) *and* (20)*, the transferred function approximation error expresses an approximation error with distribution shifted to $(d_\rho^{\pi^*}, \pi^*)$. With the softmax parameterization or linear MDP structure (Jin et al. 2020), it has been shown that $\epsilon_{bias} = 0$ (Agarwal et al. 2020). When parameterized by the restricted policy class, $\epsilon > 0$ due to $\pi_\theta$ not containing all policies. However, for a rich neural network parameterization, the $\epsilon_{bias}$ is small (Wang et al. 2019). Similar assumption has been adopted in (Liu et al. 2020) and (Agarwal et al. 2020).*

## Global Convergence For NPG-PD Method

To analyze the global convergence of the proposed algorithm, we firstly demonstrate the convergence of Lagrange function for the conservative problem, which is shown in the following Lemmas.

Algorithm 2: Estimate Value Function for objective or constraint function

**Input**: starting state and action $s, a$, reward function $r$ or constraint function $g^i$ (Here we denote as function $h$ for simplicity), policy $\pi$, discounted factor $\gamma$.
**Output**: state action value function $\hat{Q_H}(s,a)$ or state value function $\hat{V_h}(s)$

```
Estimate state action value function as $\hat{Q_h}(s,a) = \sum_{t=0}^{T-1} h(s_t, a_t)$, where $s_0 = s$,$a_0 = a$,$a_t \sim \pi(\cdot|s_t)$,
$s_{t+1} \sim P(\cdot|s_t, a_t)$,$T \sim Geo(1-\gamma)$
Estimate state value function as $\hat{V_h}(s) = \sum_{t=0}^{T-1} h(s_t, a_t)$, where $s_0 = s$,$a_t \sim \pi(\cdot|s_t)$,$s_{t+1} \sim$
$P(\cdot|s_t, a_t)$,$T \sim Geo(1-\gamma)$
```

**Lemma 3.** *Suppose a general primal-dual gradient ascent algorithm updates the parameter as*

$$\begin{aligned}\theta^{k+1} &= \theta^k + \eta\omega^k \\ \lambda_i^{k+1} &= \mathcal{P}_{(0,\Lambda]}\Big(\lambda_i^k - \eta_2\big(J_{g^i}(\theta^k) - \kappa\big)\Big)\end{aligned} \tag{21}$$

*When Assumptions 2 and 5 hold, we have*

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\Big(J_L(\pi^*_{\theta,\kappa}, \boldsymbol{\lambda}^k) - J_L(\pi_\theta^k, \boldsymbol{\lambda}^k)\Big) \leq \\ &\frac{\sqrt{\epsilon_{bias}}}{1-\gamma} + \frac{M\eta_1}{2K}\sum_{k=0}^{K-1}\mathbf{E}\|\omega^k\|^2 + \frac{\log(|\mathcal{A}|)}{\eta_1 K} \\ &+ \frac{G}{K}\sum_{k=1}^{K}\mathbf{E}\|(\omega^k - \omega_*^k)\|_2\end{aligned} \tag{22}$$

*where $\omega_*^k := \omega_*^{\theta^k}$ and is defined in Eq.* (20)

To prove the above Lemma, we extend the result in (Liu et al. 2020)[Proposition 4.5] to our setting. The extended result is stated and proved in Appendix. Then, to prove the global convergence of the Lagrange function, it is sufficient to bound $\frac{G}{K}\sum_{k=1}^K \mathbf{E}\|(\omega^k - \omega_*^k)\|_2$ and $\frac{M\eta_1}{2K}\sum_{k=0}^{K-1}\mathbf{E}\|\omega^k\|^2$ in Lemma 3. The detailed proof of them can be found in Appendix. At a high level, the first term is the difference between the estimated and exact NPG update direction, which can be bounded using the convergence of SGD procedure. The second term is the bound of the norm of estimated gradient. To bound the second term, we need the following first-order convergence result.

**Lemma 4.** *In the NPG update process, if we take $\eta_1 = \frac{\mu_F^2}{4G^2 L_J}$, then we have the first order stationary as*

$$\begin{aligned}&\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}\|\nabla_\theta J_L(\theta^k, \boldsymbol{\lambda}^k)\|_2^2 \leq \frac{16(2+3I\Lambda)G^4 L_J}{\mu_F^2 K(1-\gamma)} \\ &+ \frac{8(\mu_F^2 + 2G^4)}{\mu_F N}\Big[2[\frac{G^2(1+I\Lambda)}{\mu_F(1-\gamma)^2} + \frac{2}{(1-\gamma)^2}]\sqrt{d} \\ &+ \frac{G^2(1+I\Lambda)}{\mu_F(1-\gamma)^2}\Big]^2\end{aligned} \tag{23}$$

*Moreover, for any given $\epsilon > 0$, if we take $K = \mathcal{O}\bigg(\frac{I\Lambda}{(1-\gamma)^4\epsilon}\bigg)$ and $N = \mathcal{O}\bigg(\frac{I^2\Lambda^2}{(1-\gamma)^4\epsilon}\bigg)$, we have*

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}\|\nabla_\theta J_L(\theta^k, \boldsymbol{\lambda}^k)\|_2^2 \leq \epsilon \tag{24}$$

**Remark 5.** *The basic idea of the proof for first-order stationary is from (Liu et al. 2020). However, due to the introduction of the constraints, we need to further consider the update of the dual variable. The detailed proof can be founded in Appendix.*

Given the above Lemmas, it is sufficient to achieve the final bound of Lagrage function for conservative problem as below.

**Lemma 5.** *Under the Assumption 2, 3, 4 and 5, the proposed algorithm achieve the global convergence of the Lagrange function, which can be formally written as*

$$\frac{1}{K}\sum_{k=1}^{K}\mathbf{E}\bigg(J_L(\pi^*_{\theta,\kappa}, \boldsymbol{\lambda}^k) - J_L(\pi^k_\theta, \boldsymbol{\lambda}^k)\bigg) \leq \frac{\sqrt{\epsilon_{bias}}}{1-\gamma} + \epsilon_{K,N} \tag{25}$$

*where*

$$\begin{aligned}\epsilon_{K,N} = \mathcal{O}\bigg(\frac{1}{(1-\gamma)^3K}\bigg) + \mathcal{O}\bigg(\frac{I^2\Lambda^2}{(1-\gamma)^2N}\bigg)\\ + \mathcal{O}\bigg(\frac{I\Lambda}{(1-\gamma)\sqrt{N}}\bigg) + \mathcal{O}\bigg(\frac{I\Lambda}{K(1-\gamma)}\bigg)\end{aligned} \tag{26}$$

Before we get the final result for the regret and constraint violation, we need to bound the gap between the optimal value function of the original problem and conservative problem. Such gap can be bounded in the dual domain. To do that, we recall the definition of state-action occupancy measure $d^\pi \in \mathbb{R}^{|S||A|}$ as

$$d^\pi(s,a) = (1-\gamma)\mathbb{P}\bigg(\sum_{t=0}^{\infty}\gamma^t \cdot 1_{s_t=s,a_t=a}|\pi, s_0 \sim \rho\bigg) \tag{27}$$

We note that the objective and constraint can be written as

$$\begin{aligned}J_r(\pi_\theta) &= \frac{1}{1-\gamma}\left\langle r, d^{\pi_\theta}\right\rangle\\ J_{g^i}(\pi_\theta) &= \frac{1}{1-\gamma}\left\langle g^i, d^{\pi_\theta}\right\rangle, \forall i \in [I]\end{aligned} \tag{28}$$

Define $\mathcal{D}$ to be the set of vector $\phi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ satisfying

$$\begin{cases}\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\phi(s',a)(\delta_s(s') - \gamma\mathbf{P}(s|s',a)) = (1-\gamma)\rho(s)\\ \phi(s,a) \geq 0, \forall(s,a) \in \mathcal{S}\times\mathcal{A}\end{cases} \tag{29}$$

By summing the first constraint over $s$, we have $\sum_{s,a}\phi(s,a) = 1$, which means that $\phi$ in the above set is occupancy measure. By Eq. (28) and (29), we define the following problem which can be found in the reference (Altman 1999)

$$\begin{aligned}\max_{\phi\in\mathcal{D}} \quad & \frac{1}{1-\gamma}\left\langle r, \phi\right\rangle\\ s.t. \quad & \frac{1}{1-\gamma}\left\langle g^i, \phi\right\rangle \geq 0, \forall i \in [I]\end{aligned} \tag{30}$$

For the full-parameterized policy, it can be shown that the above problem is equivalent to the original problem Eq. (3). However, the strong duality doesn't hold for general parameterization. Thus, we need the following assumption to bound the gap between them.

**Assumption 6.** *For any $\phi \in \mathcal{D}$, we define a stationary policy as*

$$\pi'(a|s) = \frac{\phi(s,a)}{\sum_a \phi(s,a)}. \tag{31}$$

*We assume that there always exists a $\theta \in \Theta$ such that $|\pi'(a|s) - \pi_\theta(a|s)| \leq \epsilon_{bias2}, \forall(s,a) \in \mathcal{S}\times\mathcal{A}$*

**Remark 6.** *The intuition behind the above assumption is that the parameterization is rich enough so that we can always find a certain parameter $\theta$ and $\pi_\theta$ is close to the above stationary policy. A special case is softmax parameterization, where $\epsilon_{bias2} = 0$.*

With such an assumption, we reveal the relationship between the optimal value of primal problem and dual problem as follows, whose proof can be found in Appendix.

**Lemma 6.** *Under Assumption 6, denote $\pi_{\theta^*}$ as the optimal policy of the original problem defined in Eq.(3) and $\phi^*$ as the optimal occupancy measure for the new problem defined in Eq. (30), we have*

$$\left\langle r, \phi^*\right\rangle - \epsilon_{bias2} \leq J_r^{\pi^*_\theta} \leq \left\langle r, \phi^*\right\rangle \tag{32}$$

Equipped with the above lemma, we bound the gap between original problem and conservative problem in the following lemma.

**Lemma 7.** *Under Assumption 6, Denote $\pi_{\theta^*_\kappa}$ as the optimal policy for the conservative problem, we have*

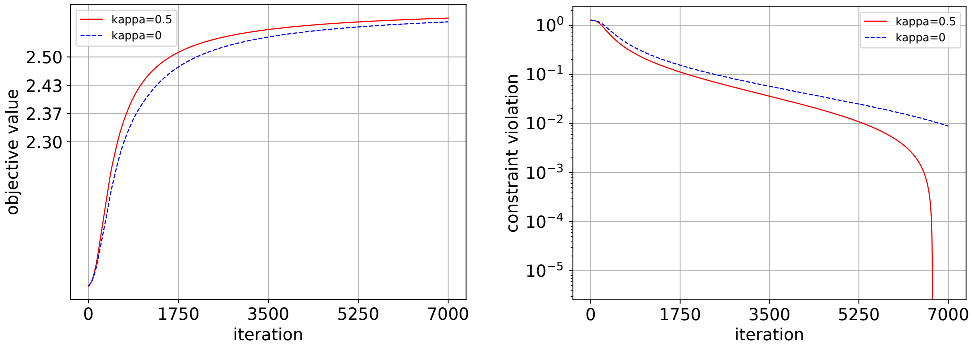
$$J_r^{\pi_{\theta^*}} - J_r^{\pi_{\theta^*_\kappa}} \leq \frac{\epsilon_{bias2}}{(1-\gamma)^2} + \frac{\kappa}{\varphi} \tag{33}$$

Equipped with Lemma 5 and 7, we provide the main result for the NPG-PD algorithm for the objective function and constrained violation. The detailed proof can be founded in Appendix.

**Theorem 1.** *For any $\epsilon > 0$, in the Natural Policy Gradient Algorithm 1, if step-size $\eta_1 = \frac{\mu_F^2}{4G^2L_J}$ and $\eta_2 = \frac{1}{\sqrt{K}}$, the number of iterations $K = \mathcal{O}\bigg(\frac{I^2\Lambda^2}{(1-\gamma)^4\epsilon^2}\bigg)$, the number of samples for per iteration $N = \mathcal{O}\bigg(\frac{I^2\Lambda^2}{(1-\gamma)^2\epsilon^2}\bigg)$ and take the conservative variable $\kappa$ as*

$$\sqrt{\frac{2}{\eta_2 K}\big(\frac{\sqrt{\epsilon_{bias}}}{1-\gamma} + \epsilon_{K,N} + \frac{2\sum_{i\in[I]}\lambda_i^* + 1}{1-\gamma}\big) + \frac{4}{K(1-\gamma)^2}}$$

*then we have $\epsilon$-optimal policy with zero constraint viola-*



(a) Comparison for the convergence of objective

(b) Comparison for the constriant violation

Figure 1: Comparison of objective and constraint violation between $\kappa = 0.5$ and $\kappa = 0$. For the constraint violation figure, we use the log axis to make zero constraint violation more obvious.

*tions. Formally,*

$$\frac{1}{K}\sum_{k=0}^{K-1}\left(J_r(\pi_\theta^*) - J_r(\pi_\theta^k)\right) \leq \mathcal{O}\left(\frac{\sqrt{\epsilon_{bias}}}{1-\gamma}\right) + \mathcal{O}\left(\frac{\epsilon_{bias2}}{(1-\gamma)^2}\right) + \mathcal{O}\left(\epsilon\right)$$
$$\frac{1}{K}\sum_{k=0}^{K-1} J_{g^i}(\pi_\theta^k) \geq 0 \tag{34}$$

*In other words, the NPG-PD algorithm needs $\mathcal{O}\left(\frac{I^4\Lambda^4}{(1-\gamma)^6\epsilon^4}\right)$ trajectories.*

**Remark 7.** *The proposed algorithm doesn't only achieve the zero constraint violation, but also achieves the state of art sample complexity over general parameterization policy-based algorithm. In Theorem 1, we can see that the algorithm converges to the neighbourhood of the global optimal and the bias is controlled by two parameters $\epsilon_{bias}$ and $\epsilon_{bias2}$ defined in Assumption 5 and 6, respectively. If the parameterization is sufficient enough, then $\epsilon_{bias} = \epsilon_{bias2} = 0$. However, whether there exists certain relationship between Assumption 5 and 6 is an interesting question for future work.*

## Simulation[1]

In order to verify the performance of the proposed algorithm (Algorithm 1), we utilize the simulation code from (Ding et al. 2020) and compare the proposed algorithm with them. We establish a random CMDP, where the state space and action space are $|\mathcal{S}| = 10, |\mathcal{A}| = 5$. The transition matrix $P(s'|s,a)$ is chosen by generating each entry uniformly at random in $[0,1]$, followed by normalization. Similarly, the reward function $r(s,a) \sim U(0,1)$ and constraint function $g(s,a) \sim U(-0.71, 0.29)$. Only 1 constraint function is considered here. The initial state distribution is set to uniform and the discount factor is $\gamma = 0.8$. For the general parameterization, we use a feature map with dimension $d = 35$, and for each SGD procedure we use $N = 100$ number of samples. The learning rate for $\theta$ and $\lambda$ are set to 0.1. The more detailed information for the simulation setting can be found in Appendix. We run the algorithm for $K = 7000$ iterations and compare the proposed algorithm with $\kappa = 0.5$ and the NPG-PD algorithm (Ding et al. 2020) which doesn't consider the zero constraint violation case (equivalently $\kappa = 0$) in Figure 1.

From Fig. 1, we find that the convergence of the reward is similar and the proposed algorithm converges even faster than the non-zero constraint violation case. However, for the constraint violation, we find that when $\kappa = 0.5$, the log of constraint violation converges to negative infinity, which means that the constraint violation is below 0. In contrast, the constraint violation still exists when $\kappa = 0$. The comparison between $\kappa = 0.5$ and $\kappa = 0$ validates the result in Theorem 1.

## Conclusion

In this paper, we propose a novel algorithm for Constrained Markov Decision Process and the proposed algorithm achieves the state-of-the-art sample complexity over general parametrization policy-based algorithms. By revealing the relationship between the primal and dual problem, the gap between conservative problem and original problem is bounded, which finally leads to the analysis of zero constraint violation. The proposed algorithm converges to the neighbourhood of the global optimal and the gap is controlled by the richness of parametrization.

The key limitation of the work includes the assumptions used to prove the results. Simplifying or removing Assumptions 5 and 6 on the bias parameters is a valuable problem in the future work.

## References

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2020. Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes. In Abernethy,

[1] The code can be found at https://github.com/bqb3927586/NPG-zero-violation

J.; and Agarwal, S., eds., *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, 64–66. PMLR.

Akhtar, Z.; Bedi, A. S.; and Rajawat, K. 2021. Conservative Stochastic Optimization With Expectation Constraints. *IEEE Transactions on Signal Processing*, 69: 3190–3205.

Altman, E. 1999. *Constrained Markov decision processes: stochastic modeling*. Routledge.

Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; and Bharath, A. A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38.

Ding, D.; Zhang, K.; Basar, T.; and Jovanovic, M. 2020. Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 8378–8390. Curran Associates, Inc.

Jin, C.; Yang, Z.; Wang, Z.; and Jordan, M. I. 2020. Provably efficient reinforcement learning with linear function approximation. In Abernethy, J.; and Agarwal, S., eds., *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, 2137–2143. PMLR.

Kakade, S. 2001. A Natural Policy Gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, 1531–1538. Cambridge, MA, USA: MIT Press.

Kay, S. M. 1997. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*.

Konda, V. R.; and Tsitsiklis, J. N. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*, 1008–1014. Citeseer.

Liu, T.; Zhou, R.; Kalathil, D.; Kumar, P.; and Tian, C. 2021. Policy Optimization for Constrained MDPs with Provable Fast Global Convergence. *arXiv preprint arXiv:2111.00552*.

Liu, Y.; Zhang, K.; Basar, T.; and Yin, W. 2020. An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 7624–7636. Curran Associates, Inc.

Mei, J.; Xiao, C.; Szepesvari, C.; and Schuurmans, D. 2020. On the Global Convergence Rates of Softmax Policy Gradient Methods. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6820–6829. PMLR.

Papini, M.; Binaghi, D.; Canonaco, G.; Pirotta, M.; and Restelli, M. 2018. Stochastic Variance-Reduced Policy Gradient. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 4026–4035. PMLR.

Peters, J.; and Schaal, S. 2008. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4): 682–697. Robotics and Neuroscience.

Sutton, R. S. 1988. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1): 9–44.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In Solla, S.; Leen, T.; and Müller, K., eds., *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Vaswani, S.; Yang, L. F.; and Szepesvári, C. 2022. Near-optimal sample complexity bounds for constrained MDPs. *arXiv preprint arXiv:2206.06270.*

Vu, T. L.; Mukherjee, S.; Yin, T.; Huang, R.; Huang, Q.; et al. 2020. Safe reinforcement learning for emergency loadshedding of power systems. *arXiv preprint arXiv:2011.09664*.

Wang, L.; Cai, Q.; Yang, Z.; and Wang, Z. 2019. Neural Policy Gradient Methods: Global Optimality and Rates of Convergence. arXiv:1909.01150.

Wen, L.; Duan, J.; Li, S. E.; Xu, S.; and Peng, H. 2020. Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–7. IEEE.

Xu, P.; Gao, F.; and Gu, Q. 2020a. An Improved Convergence Analysis of Stochastic Variance-Reduced Policy Gradient. In Adams, R. P.; and Gogate, V., eds., *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, 541–551. PMLR.

Xu, P.; Gao, F.; and Gu, Q. 2020b. Sample Efficient Policy Gradient Methods with Recursive Variance Reduction. arXiv:1909.08610.

Xu, T.; Liang, Y.; and Lan, G. 2021. Crpo: A new approach for safe reinforcement learning with convergence guarantee. In *International Conference on Machine Learning*, 11480–11491. PMLR.

Zeng, S.; Anwar, M. A.; Doan, T. T.; Raychowdhury, A.; and Romberg, J. 2021. A decentralized policy gradient approach to multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, 1002–1012. PMLR.

Zeng, S.; Doan, T. T.; and Romberg, J. 2022. Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained Markov decision processes. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, 4028–4033. IEEE.

Zhang, J.; Ni, C.; Yu, Z.; Szepesvari, C.; and Wang, M. 2021. On the Convergence and Sample Efficiency of Variance-Reduced Policy Gradient Method. arXiv:2102.08607.

Zhang, K.; Koppel, A.; Zhu, H.; and Başar, T. 2020. Global Convergence of Policy Gradient Methods to (Almost) Locally Optimal Policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612.