# Refined Policy Improvement Bounds for MDPs

**J. G. Dai** [1 2] **Mark Gluzman** [3]

## Abstract

The policy improvement bound on the difference of the discounted returns plays a crucial role in the theoretical justification of the trust-region policy optimization (TRPO) algorithm. The existing bound leads to a degenerate bound when the discount factor approaches one, making the applicability of TRPO and related algorithms questionable when the discount factor is close to one. We refine the results in (Schulman et al., 2015; Achiam et al., 2017) and propose a novel bound that is "continuous" in the discount factor. In particular, our bound is applicable for MDPs with the long-run average rewards as well.

## 1. Introduction

In (Kakade & Langford, 2002) the authors developed a conservative policy iteration algorithm for Markov decision processes (MDPs) that can avoid catastrophic large policy updates; each iteration generates a new policy as a mixture of the old policy and a greedy policy. They proved that the updated policy is guaranteed to improve when the greedy policy is properly chosen and the updated policy is sufficiently close to the old one. In (Schulman et al., 2015) the authors generalized the proof of (Kakade & Langford, 2002) to a policy improvement bound for two *arbitrary* randomized policies. This policy improvement bound allows one to find an updated policy that guarantees to improve by solving an unconstrained optimization problem. (Schulman et al., 2015) also proposed a practical algorithm, called trust region policy optimization (TRPO), that approximates the theoretically-justified update scheme by solving a constrained optimization problem in each iteration. In recent years, several modifications of TRPO have been proposed (Schulman et al., 2016; 2017; Achiam et al., 2017; Abdolmaleki et al., 2018). These studies continued to exploit the policy improvement bound to theoretically motivate their algorithms.

The policy improvement bounds in (Schulman et al., 2015; Achiam et al., 2017) are lower bounds on the difference of the expected *discounted returns* under two policies. Unfortunately, the use of these policy improvement bounds becomes questionable and inconclusive when the discount factor is close to one. These policy improvement bounds degenerate as discount factor converges to one. That is, the lower bounds on the difference of discounted returns converge to negative infinity as the discount factor goes to one, although the difference of discounted returns converges to the difference of (finite) average rewards. Nevertheless, numerical experiments demonstrate that the TRPO algorithm and its variations perform best when the discount factor $\gamma$ is close to one, a region that the existing bounds do not justify; e.g. (Schulman et al., 2015; 2016; 2017) used $\gamma = 0.99$, and (Schulman et al., 2016; Achiam et al., 2017) used $\gamma = 0.995$ in their experiments.

Recent studies (Dai & Gluzman, 2021; Zhang & Ross, 2021) proposed policy improvement bounds for average rewards, showing that a family of TRPO algorithms can be used for continuing problems with long-run average reward objectives. Still it remains unclear how the large values of the discount factor can be justified and why the policy improvement bounds in (Schulman et al., 2015; Achiam et al., 2017) for the discounted rewards do not converge to one of the bounds provided in (Dai & Gluzman, 2021; Zhang & Ross, 2021).

In this study, we provide a unified derivation of policy improvement bounds for both discounted and average reward MDPs. Our bounds depend on the discount factor *continuously*. When the discount factor converges to 1, the corresponding bound for discounted returns converges to a policy improvement bound for average rewards. We achieve these results by two innovative observations. First, we embed the discounted future state distribution under a fixed policy as the stationary distribution of a modified Markov chain. Second, we introduce an *ergodicity coefficient* from Markov chain perturbation theory to bound the one-norm of the difference of discounted future state distributions, and

[1]School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China [2]School of Operations Research and Information Engineering, Cornell University, Ithaca, USA [3]Center for Applied Mathematics, Cornell University, Ithaca, USA. Correspondence to: Mark Gluzman <mg2289@cornell.edu>.

*Workshop on Reinforcement Learning Theory, the $38^{th}$ International Conference on Machine Learning*, 2021. 

prove that this bound is optimal in a certain sense. Our results justify the use of a large discount factor in TRPO algorithm and its variations.

## 2. Preliminaries

We consider an MDP defined by the tuple $(\mathcal{X}, \mathcal{A}, P, r, \mu)$, where $\mathcal{X}$ is a finite state space; $\mathcal{A}$ is a finite action space; $P$ is the transition probability function, $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the reward function; $\mu$ is the probability distribution of the initial state $x_0$.

We let $\pi$ denote a stationary randomized policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the probability simplex over $\mathcal{A}$. Under policy $\pi$, the corresponding Markov chain has a transition matrix $P^\pi$ given by $P^\pi(x, y) := \sum_{a\in\mathcal{A}} \pi(a|x)P(y|x, a),\ x, y \in \mathcal{X}$. We assume that MDPs we consider are unichain, meaning that for any stationary policy $\pi$ the corresponding Markov chain with transition matrix $P^\pi$ contains only one recurrent class (Puterman, 2005). We use $d^\pi$ to denote a unique stationary distribution of Markov chain with transition matrix $P^\pi$.

For a vector $a$ and a matrix $A$, $a^T$ and $A^T$ denote their transposes. For a vector $a$, we use the following vector norm: $\|a\|_1 := \sum_{x\in\mathcal{X}} |a(x)|$. For a matrix $A$, we define the following induced norm: $\|A\|_1 := \max_{y\in\mathcal{X}} \sum_{x\in\mathcal{X}} |A(x, y)|$.

### 2.1. MDPs with infinite horizon discounted returns

We let $\gamma \in [0, 1)$ be a discount factor. We define the value function for a given policy $\pi$ as

$$V_\gamma^\pi(x) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t, a_t) \Big| \pi, x_0 = x\right],$$

where $x_t$, $a_t$ are random variables for the state and action at time $t$ upon executing the policy $\pi$ from the initial state $x$. For policy $\pi$ we define the state-action value function as $Q_\gamma^\pi(x, a) := r(x, a) + \gamma\mathbb{E}_{y\sim P^\pi(\cdot|x,a)}\left[V_\gamma^\pi(y)\right]$, and the advantage function as $A_\gamma^\pi(x, a) := Q_\gamma^\pi(x, a) - V_\gamma^\pi(x)$.

We define the discounted future state distribution of policy $\pi$ as

$$d_\gamma^\pi(x) := (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\Big[x_t = x | x_0 \sim \mu; x_1, x_2, \ldots \sim \pi\Big].$$

We measure the performance of policy $\pi$ by its expected discounted return from the initial state distribution $\mu$:

$$\eta_\gamma^\pi(\mu) := (1-\gamma)\mathbb{E}_{x\sim\mu}[V_\gamma^\pi(x)] = \mathbb{E}_{x\sim d_\gamma^\pi, a\sim\pi(\cdot|x)}[r(x, a)].$$

In the following lemma we give an alternative definition of the discounted future state distribution as a stationary distribution of a modified transition matrix.

**Lemma 1.** *For a stationary policy $\pi$, we define a discounted transition matrix for policy $\pi$ as*

$$P_\gamma^\pi := \gamma P^\pi + (1-\gamma)e\mu^T, \tag{1}$$

*where $e := (1, 1, .., 1)^T$ is a vector of ones, $e\mu^T$ is the matrix which rows are equal to $\mu^T$.*

*Then the discounted future state distribution of policy $\pi$, $d_\gamma^\pi$, is the stationary distribution of transition matrix $P_\gamma^\pi$.*

***Proof of Lemma 1.*** We need to show that $(d_\gamma^\pi)^T P_\gamma^\pi = (d_\gamma^\pi)^T$. Indeed, we get

$$\begin{aligned}
(d_\gamma^\pi)^T P_\gamma^\pi &= (1-\gamma)\mu^T \sum_{t=0}^{\infty} (\gamma P^\pi)^t P_\gamma^\pi \\
&= (1-\gamma)\mu^T \sum_{t=0}^{\infty} (\gamma P^\pi)^t \Big(\gamma P^\pi + (1-\gamma)e\mu^T\Big) \\
&= (1-\gamma)\mu^T \sum_{t=0}^{\infty} (\gamma P^\pi)^{t+1} + (1-\gamma)^2\mu^T \sum_{t=0}^{\infty} \gamma^n e\mu^T \\
&= (1-\gamma)\mu^T \sum_{t=0}^{\infty} (\gamma P^\pi)^{t+1} + (1-\gamma)\mu^T \\
&= (1-\gamma)\mu^T \Big(\sum_{t=0}^{\infty} (\gamma P^\pi)^{t+1} + I\Big) \\
&= (1-\gamma)\mu^T \sum_{t=0}^{\infty} (\gamma P^\pi)^t \\
&= (d_\gamma^\pi)^T.
\end{aligned}$$

□

### 2.2. MDPs with long-run average rewards

The long-run average reward of policy $\pi$ is defined as

$$\begin{aligned}
\eta^\pi :&= \lim_{N\to\infty} \frac{1}{N}\mathbb{E}\left[\sum_{t=0}^{N-1} r(x_t, a_t) \mid \pi, x_0 \sim \mu\right] \\
&= \mathbb{E}_{x\sim d^\pi, a\sim\pi(\cdot|x)}\left[r(x, a)\right].
\end{aligned}$$

For an MDP with a long-run average reward objective we define the relative value function

$$V^\pi(x) := \lim_{N\to\infty} \mathbb{E}\left[\sum_{t=0}^{N-1} (r(x_t, a_t) - \eta^\pi) \mid \pi, x_0 = x\right],$$

the relative state-action value function $Q^\pi(x, a) := r(x, a) - \eta^\pi + \mathbb{E}_{y\sim P^\pi(\cdot|x,a)}\left[V^\pi(y)\right]$, and the relative advantage function $A^\pi(x, a) := Q^\pi(x, a) - V^\pi(x)$. The following relations hold for value, state-action value, and advantage functions.

**Lemma 2.** *We let $\pi$ be a stationary policy, $\gamma$ be the discount factor, and $\mu$ be the initial state distribution. Then the following limits hold for each $x \in \mathcal{X}$, $a \in \mathcal{A}$:*

$\eta^\pi = \lim_{\gamma\to 1} \eta^\pi_\gamma(\mu)$, $V^\pi(x) = \lim_{\gamma\to 1} \left(V^\pi_\gamma(x) - (1-\gamma)^{-1}\eta^\pi\right)$, $Q^\pi(x,a) = \lim_{\gamma\to 1} \left(Q^\pi_\gamma(x,a) - (1-\gamma)^{-1}\eta^\pi\right)$, *and* $A^\pi(x,a) = \lim_{\gamma\to 1} A^\pi_\gamma(x,a)$.

The proofs of identities for the average rewards and value functions can be found in Section 8 in (Puterman, 2005). The rest results follow directly.

## 3. Novel Policy Improvement Bounds

The policy improvement bound in (Schulman et al., 2015; Achiam et al., 2017) for the discounted returns serves to theoretically justify the TRPO algorithm and its variations. The following lemma is a reproduction of Corollary 1 in (Achiam et al., 2017).

**Lemma 3.** *For any two policies $\pi$ and $\tilde\pi$ the following bound holds:*

$$\eta^{\tilde\pi}_\gamma(\mu) - \eta^\pi_\gamma(\mu) \geq \mathop{\mathbb{E}}_{x\sim d^\pi_\gamma, a\sim\tilde\pi(\cdot|x)} \left[A^\pi_\gamma(x,a)\right] - \frac{2\gamma\epsilon^{\tilde\pi}_\gamma}{1-\gamma} \mathop{\mathbb{E}}_{x\sim d^\pi_\gamma} \left[TV\Big(\tilde\pi(\cdot|x) \,||\, \pi(\cdot|x)\Big)\right], \tag{2}$$

*where* $TV\Big(\tilde\pi(\cdot|x)||\pi(\cdot|x)\Big) := \frac{1}{2}\sum_{a\in\mathcal{A}} |\tilde\pi(a|x) - \pi(a|x)|$, *and* $\epsilon^{\tilde\pi}_\gamma := \max_{x\in\mathcal{X}} \Big|\mathop{\mathbb{E}}_{a\sim\tilde\pi(\cdot|x)} [A^\pi_\gamma(x,a)]\Big|$.

The left-hand side of (2) converges to the difference of average rewards as $\gamma \to 1$. Unfortunately, the right-hand side of (2) converges to the negative infinity because of $(1-\gamma)^{-1}$ factor in the second term. Our goal is to get a new policy improvement bound for discounted returns that does not degenerate.

The group inverse $D$ of a matrix $A$ is the unique matrix such that $ADA = A$, $DAD = D$, and $DA = AD$. From (Meyer, 1975), we know that if stochastic matrix $P$ is aperiodic and irreducible then the group inverse matrix of $I - P$ is well-defined and equals to $D = \sum_{t=0}^{\infty}(P^t - ed^T)$, where $d$ is the stationary distribution of $P$.

We let $D^\pi_\gamma$ be the group inverse of matrix $I - P^\pi_\gamma$, where $P^\pi_\gamma$ is defined by (1). Following (Seneta, 1991), we define a one-norm ergodicity coefficient for a matrix $A$ as

$$\tau_1[A] := \max_{\substack{\|x\|_1=1 \\ x^T e=0}} \|A^T x\|_1. \tag{3}$$

The one-norm ergodicity coefficient has two important properties. First,

$$\|A^T x\|_1 \leq \tau_1[A]\|x\|_1, \tag{4}$$

for any matrix $A$ and vector $x$ such that $x^T e = 0$. Second, $\tau_1[A] = \tau_1[A + ec^T]$, for any vector $c$. By Lemma 4 below, $\tau_1\left[D^\pi_\gamma\right] = \tau_1\left[(I - \gamma P^\pi)^{-1}\right]$, for $\gamma < 1$.

**Lemma 4.** *We let $\pi$ be an arbitrary policy. Then*

$$D^\pi_\gamma = (I - \gamma P^\pi)^{-1} + e(d^\pi_\gamma)^T(I - (I - \gamma P^\pi)^{-1}) - e(d^\pi)^T.$$

We are ready to state the main result of our study.

**Theorem 1.** *The following bound on the difference of discounted returns of two policies $\pi$ and $\tilde\pi$ holds:*

$$\eta^{\tilde\pi}_\gamma(\mu) - \eta^\pi_\gamma(\mu) \geq \mathop{\mathbb{E}}_{x\sim d^\pi_\gamma, a\sim\tilde\pi(\cdot|x)} \left[A^\pi_\gamma(x,a)\right] - 2\gamma\epsilon^{\tilde\pi}_\gamma \tau_1\left[D^{\tilde\pi}_\gamma\right] \mathop{\mathbb{E}}_{x\sim d^\pi_\gamma} \left[TV\Big(\tilde\pi(\cdot|x) \,||\, \pi(\cdot|x)\Big)\right]. \tag{5}$$

We provide a sketch of the proof of Theorem 1.

***Proof of Theorem 1.*** We closely follow the first steps in the proof of Lemma 2 in (Achiam et al., 2017) and start with

$$\eta^{\tilde\pi}_\gamma(\mu) - \eta^\pi_\gamma(\mu) \geq \mathop{\mathbb{E}}_{x\sim d^\pi_\gamma, a\sim\tilde\pi(\cdot|x)} \left[A^\pi_\gamma(x,a)\right] - \max_{x\in\mathcal{X}} \Big|\mathop{\mathbb{E}}_{a\sim\tilde\pi(\cdot|x)} [A^\pi_\gamma(x,a)]\Big| \|d^\pi_\gamma - d^{\tilde\pi}_\gamma\|_1.$$

Next, unlike (Achiam et al., 2017), we obtain an upper bound on $\|d^{\tilde\pi}_\gamma - d^\pi_\gamma\|_1$ that does not degenerate as $\gamma \to 1$. We use the following perturbation identity:

$$(d^{\tilde\pi}_\gamma)^T - (d^\pi_\gamma)^T = \gamma(d^\pi_\gamma)^T(P^\pi - P^{\tilde\pi})D^{\tilde\pi}_\gamma. \tag{6}$$

Identity (6) follows from the perturbation identity for stationary distributions, see equation (4.1) in (Meyer, 1980), and the fact that $d^{\tilde\pi}_\gamma$ and $d^\pi_\gamma$ are the stationary distributions of the discounted transition matrices $P^{\tilde\pi}_\gamma$ and $P^\pi_\gamma$, respectively. We make use of the ergodicity coefficient (3) to get a new perturbation bound:

$$\begin{aligned} \|d^{\tilde\pi}_\gamma - d^\pi_\gamma\|_1 &= \gamma \left\|\left(D^{\tilde\pi}_\gamma\right)^T (P^\pi - P^{\tilde\pi})^T d^\pi_\gamma\right\|_1 \\ &\leq \gamma\tau_1\left[D^{\tilde\pi}_\gamma\right] \left\|(P^\pi - P^{\tilde\pi})^T d^\pi_\gamma\right\|_1 \\ &\leq 2\gamma\tau_1\left[D^{\tilde\pi}_\gamma\right] \mathop{\mathbb{E}}_{x\sim d^\pi_\gamma} \left[\mathrm{TV}\Big(\tilde\pi(\cdot|x) \,||\, \pi(\cdot|x)\Big)\right], \end{aligned} \tag{7}$$

where inequality (7) holds due to (4) and equality $(P^\pi - P^{\tilde\pi})e = 0$.

□

The novel policy improvement bound (5) converges to a meaningful bound on the difference of average rewards as $\gamma$ goes to 1. Corollary 1 follows from Theorem 1, Lemma 2 and the fact that $\tau_1\left[D^{\tilde\pi}_\gamma\right] \to \tau_1\left[D^{\tilde\pi}\right]$ as $\gamma \to 1$.

**Corollary 1.** *The following bound on the difference of long-run average rewards of two policies $\pi$ and $\tilde{\pi}$ holds:*

$$\eta^{\tilde{\pi}} - \eta^{\pi} \geq \mathop{\mathbb{E}}_{x\sim d^{\pi}, a\sim\tilde{\pi}(\cdot|x)} [A^{\pi}(x,a)] - 2\epsilon^{\tilde{\pi}} \tau_1 \left[D^{\tilde{\pi}}\right] \mathop{\mathbb{E}}_{x\sim d^{\pi}} \left[TV\Big(\tilde{\pi}(\cdot|x) \;||\; \pi(\cdot|x)\Big)\right], \tag{8}$$

*where $D^{\tilde{\pi}}$ is the group inverse of matrix $I - P^{\tilde{\pi}}$, $\epsilon^{\tilde{\pi}} := \max_{x\in\mathcal{X}} \left| \mathop{\mathbb{E}}_{a\sim\tilde{\pi}(\cdot|x)} [A^{\pi}(x,a)] \right|$.*

Lemma 5 demonstrates that we use the best (smallest) norm-wise bound on the difference of stationary distributions in the proof of Theorem 1. Lemma 5 is based on (Kirkland et al., 2008).

**Lemma 5.** *We consider two irreducible and aperiodic transition matrices $P$ and $\tilde{P}$ with stationary distributions $d$ and $\tilde{d}$, respectively. We say that $\tau[\tilde{P}]$ is a condition number of matrix $\tilde{P}$ if inequality*

$$\|d - \tilde{d}\|_1 \leq \tau[\tilde{P}] \|(P - \tilde{P})^T d\|_1, \tag{9}$$

*holds for any transition matrix $P$. We let $\tilde{D}$ be a group inverse matrix of $I - \tilde{P}$.*

*Then $\tau_1[\tilde{D}]$ is the smallest condition number: $\tau_1[\tilde{D}] \leq \tau[\tilde{P}]$ holds for any condition number $\tau(\tilde{P})$ satisfying (9).*

Lemma 5 shows that inequality (7) in the proof of Theorem 1 is a key to the improvement of the policy improvement bounds in (Schulman et al., 2015; Achiam et al., 2017). Moreover, it follows from Lemma 5 that Corollary 1 provides a better policy improvement bound for the average reward criterion than (Dai & Gluzman, 2021; Zhang & Ross, 2021).

## 4. Interpretation of $\tau_1[D^{\pi}_{\gamma}]$

We provide several bounds on $\tau_1[D^{\pi}_{\gamma}]$ to reveal its dependency on the discount factor $\gamma$ and policy $\pi$. First, we show how the magnitude of $\tau_1[D^{\pi}_{\gamma}]$ is governed by the subdominant eigenvalues of the Markov chain. We let $P$ be an irreducible Markov chain and let $D$ be the group inverse matrix of $I - P$. We define the spectrum of transition matrix $P$ as $\{1, \lambda_2, \lambda_3, ..., \lambda_{|\mathcal{X}|}\}$, where $|\mathcal{X}|$ is a cardinality of the state space $\mathcal{X}$. Then, the ergodicity coefficient can be bounded as

$$\tau_1[D] \leq \sum_{i=2}^{|\mathcal{X}|} \frac{1}{1-\lambda_i} = \text{trace}(D),$$

see (Seneta, 1993). Matrix $P_{\gamma}$ defined by (1) is called the Google matrix, and if the spectrum of transition matrix $P$ is $\{1, \lambda_2, \lambda_3, ..., \lambda_{|\mathcal{X}|}\}$, then the spectrum of the Google matrix $P_{\gamma}$ is $\{1, \gamma\lambda_2, \gamma\lambda_3, ..., \gamma\lambda_{|\mathcal{X}|}\}$, see (Haveliwala & Kamvar, 2003; Langville & Meyer, 2003). Hence, the discounting decreases the subdominant eigenvalue of the transition matrix that leads to the following bound.

**Lemma 6.** *We let $D^{\pi}_{\gamma}$ be the group inverse matrix of $I - P^{\pi}_{\gamma}$. Then for any discount factor $\gamma \in (0, 1]$*

$$\tau_1[D^{\pi}_{\gamma}] \leq \sum_{i=2}^{|\mathcal{X}|} \frac{1}{1-\gamma\lambda_i} \leq \frac{|\mathcal{X}| - 1}{1 - \gamma|\lambda_2|},$$

*where $\lambda_2$ is an eigenvalue of $P^{\pi}$ with the second largest absolute value.*

In Lemma 7 below we derive an alternative upper bound on $\tau_1[D^{\pi}_{\gamma}]$. For a given policy $\pi$, we assume the transition matrix $P^{\pi}$ is aperiodic and irreducible. By Proposition 1.7 in (Levin & Peres, 2017), there exists an integer $\ell$ such that $(P^{\pi})^q(x,y) > 0$ for all $x, y \in \mathcal{X}$, and $q \geq \ell$. Then, there exists a sufficiently small constant $\delta^{\pi}_{\mu} > 0$, such that

$$(P^{\pi})^{\ell}(x,y) \geq \delta^{\pi}_{\mu}\mu(y), \qquad \text{for each } x, y \in \mathcal{X}, \tag{10}$$

where $\mu$ denotes the distribution of the initial state.

**Lemma 7.** *We let $D^{\pi}_{\gamma}$ be the group inverse matrix of $I - P^{\pi}_{\gamma}$. We let $\delta^{\pi}_{\mu}$ be a constant that satisfies (10) for $P^{\pi}$ and some integer $\ell$. Then*

$$\tau_1[D^{\pi}_{\gamma}] \leq \frac{2\ell}{1 - \gamma + \gamma^{\ell}\delta^{\pi}_{\mu}},$$

*where $\delta^{\pi}_{\mu}$ and $\ell$ are independent of $\gamma$.*

## References


Abdolmaleki, A., Springenberg, J. T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. Maximum a Posteriori Policy Optimisation. In *Proceedings of ICLR'18*, 2018.

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained Policy Optimization. *Proceedings of ICML'17*, 70:22–31, 2017.

Dai, J. G. and Gluzman, M. Queueing Network Controls via Deep Reinforcement Learning. 2021. URL http://arxiv.org/abs/2008.01644.

Haveliwala, T. H. and Kamvar, S. D. The Second Eigenvalue of the Google Matrix. Technical report, 2003. URL https://nlp.stanford.edu/pubs/secondeigenvalue.pdf.

Kakade, S. and Langford, J. Approximately Optimal Approximate Reinforcement Learning. In *Proceedings of ICML'02*, pp. 267–274, 2002.

Kirkland, S. J., Neumann, M., and Sze, N. S. On optimal condition numbers for markov chains. *Numerische Mathematik*, 110(4):521–537, 2008.

Langville, A. N. and Meyer, C. D. Deeper Inside PageRank. *Internet Mathematics*, 1(3):335–380, 2003.

Levin, D. A. and Peres, Y. *Markov Chains and Mixing Times*. American Mathematical Society, 2nd edition, 2017.

Meyer, C. D. The Role of the Group Generalized Inverse in the Theory of Finite Markov Chains. *SIAM Review*, 17 (3):443–464, 1975.

Meyer, C. D. The Condition of a Finite Markov Chain and Perturbation Bounds for the Limiting Probabilities. *SIAM Journal on Algebraic Discrete Methods*, 1(3):273–283, 1980.

Puterman, M. L. *Markov decision processes : discrete stochastic dynamic programming*. Wiley-Interscience, 2005.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. Trust Region Policy Optimization. In *Proceedings of ICML'15*, pp. 1889–1897, 2015.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *Procedings of ICLR'16*, 2016.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms. 2017. URL http://arxiv.org/abs/1707.06347.

Seneta, E. Sensitivity analysis, ergodicity coefficients, and rank-one updates for finite Markov chains. In Stewart, W. (ed.), *Numerical Solution of Markov Chains*, pp. 121–129. Marcel Dekker, New York, 1991.

Seneta, E. Sensitivity of finite Markov chains under perturbation. *Statistics & Probability Letters*, 17(2):163–168, 1993. doi: 10.1016/0167-7152(93)90011-7.

Zhang, Y. and Ross, K. W. On-Policy Deep Reinforcement Learning for the Average-Reward Criterion. *Proceedings of ICML'21*, 2021.