Trust-Region-Free Policy Optimization for Stochastic Policies

Mingfei Sun
University of Oxford
mingfei.sun@cs.ox.ac.uk

Benjamin Ellis
University of Oxford
benjamin.ellis@keble.ox.ac.uk

Anuj Mahajan
University of Oxford
anuj.mahajan@cs.ox.ac.uk

Sam Devlin
Microsoft Research
sam.devlin@microsoft.com

Katja Hofmann
Microsoft Research
katja.hofmann@microsoft.com

Shimon Whiteson
University of Oxford
shimon.whiteson@cs.ox.ac.uk

Abstract

Trust Region Policy Optimization (TRPO) is an iterative method that simultaneously maximizes a *surrogate objective* and enforces a trust region constraint over consecutive policies in each iteration. The combination of the surrogate objective maximization and the trust region enforcement has been shown to be crucial to guarantee a monotonic policy improvement. However, solving a trust-region-constrained optimization problem can be computationally intensive as it requires many steps of conjugate gradient and a large number of on-policy samples. In this paper, we show that the trust region constraint over policies can be safely substituted by a *trust-region-free* constraint without compromising the underlying monotonic improvement guarantee. The key idea is to generalize the surrogate objective used in TRPO in a way that a monotonic improvement guarantee still emerges as a result of constraining the maximum advantage-weighted ratio between policies. This new constraint outlines a conservative mechanism for iterative policy optimization and sheds light on practical ways to optimize the generalized surrogate objective. We show that the new constraint can be effectively enforced by being conservative when optimizing the generalized objective function in practice. We call the resulting algorithm Trust-REgion-Free Policy Optimization (TREFree) as it is free of any explicit trust region constraints. Empirical results show that TREFree outperforms TRPO and Proximal Policy Optimization (PPO) in terms of policy performance and sample efficiency.

Keywords: policy optimization; deep reinforcement learning

1 Introduction

Trust Region Policy Optimization (TRPO) [9] is an iterative method that optimizes stochastic policies with a trust region constraint. One of the key ideas in TRPO is to simultaneously optimize a *surrogate objective* and enforce a trust region constraint over consecutive policies at each iteration. The use of surrogate objectives stems from the seminal work of [6], which modifies the policy gradient (PG) objective [12] by substituting the on-policy state distribution with a distribution induced by the policy from the preceding iteration. [9] show that, despite the mismatch between what the policy update should optimize, i.e. the PG objective, and what is optimized in practice, i.e. the surrogate objective, a monotonic improvement guarantee for policy performance can still emerge from constraining the policy update at each iteration. The resulting TRPO algorithm thus strictly enforces a Kullback-Leibler (KL) divergence constraint between consecutive policies, and seeks to solve a KL-constrained surrogate objective optimization at each iteration. This combination of the surrogate objective maximization and the trust region enforcement has also been shown to be crucial for policy improvement in practice [9, 2].

However, solving a KL-constrained optimization problem can be computationally intensive [10]. In particular, TRPO use a quadratic approximation of the KL that augments natural policy gradients [7] with a line-search step that critically ensures KL enforcement [9]. This procedure requires many steps of conjugate gradient and a large number of on-policy samples making it both computationally intensive and sample inefficient [13]. Many follow-up studies attempt to improve TRPO, for example by leveraging Kronecker-factored approximated curvature to approximate the trust region [13], solving the KL-regularized optimization analytically via Expectation-Maximization [1, 5], transforming TRPO into an unconstrained optimization by policy space projections [3] or integrating the constraint into differentiable layers [8].

In this paper, we propose simplifying policy optimization by completely removing the trust region constraint, without compromising the underlying monotonic improvement guarantee. Specifically, instead of the framework of surrogate objective optimization [6, 9], we generalize the surrogate objective used in TRPO in a way that a monotonic improvement guarantee still emerges as a result of constraining the maximum advantage-weighted ratio between policies. This new constraint is different from the trust region constraint in TRPO in that it does not seek to impose any divergence constraint over consecutive policies. Instead, it outlines a conservative mechanism to bound the maximum advantage-weighted ratios in each iteration, and sheds light on practical ways to directly optimize the generalized surrogate objective. We show that the new constraint can be simply enforced by being conservative when optimizing the generalized objective function in practice. Furthermore, we present Trust-REgion-Free Policy Optimization (TREFree), a practical policy optimization method for optimizing stochastic policies. Empirical results show that TREFree is effective in optimizing policies, outperforming TRPO and PPO in both performance and sample efficiency.

2 Preliminaries

Markov decision process (MDP). Single-agent RL can be modelled as an infinite-horizon discounted Markov decision process (MDP) $\{\mathcal{S}, \mathcal{A}, P, r, d_0, \gamma\}$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability distribution, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $d_0 : \mathcal{S} \to \mathbb{R}$ is the initial state distribution and $\gamma \in [0, 1)$ is the discount factor. Let π denote a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, the performance for a stochastic policy $\pi(a|s)$ is defined as: $J(\pi) \triangleq \mathbb{E}_{s_0 \sim d_0, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$ The action-value function Q_π and value function V_π are defined as: $Q_\pi(s_t, a_t) \triangleq \mathbb{E}_t \left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l}, a_{t+l}) \right]$, $V_\pi(s_t) \triangleq \mathbb{E}_{a_t \sim \pi(\cdot|s_t)} \left[Q_\pi(s_t, a_t) \right]$. Accordingly, the advantage function is defined as $A_\pi(s, a) \triangleq Q_\pi(s, a) - V_\pi(s)$.

TRPO. Define the discounted state distribution as: $d_\pi(s) \triangleq \sum_{t=0}^{\infty} \gamma^t P(s_t = s | \pi, d_0)$. The following equation is useful [6]:

$$J(\tilde{\pi}) = J(\pi) + \sum_s d_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) A_\pi(s, a). \tag{1}$$

The complex dependency of $d_{\tilde{\pi}}(s)$ on $\tilde{\pi}$ makes the right hand side (RHS) difficult to optimize directly. [9] proposed to consider the following surrogate objective:

$$L_\pi(\tilde{\pi}) \triangleq J(\pi) + \sum_s d_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s, a), \tag{2}$$

where $d_{\tilde{\pi}}$ is replaced with d_π. TRPO introduces the idea of bounding the distribution change via the policy divergence. Specifically, define $D_{\text{TV}}^{\max}(\pi, \tilde{\pi}) \triangleq \max_s D_{\text{TV}}\big(\pi(\cdot|s), \tilde{\pi}(\cdot|s)\big)$, where D_{TV} is the total variation (TV) divergence.

Theorem 2.1. *([9]) Let $\alpha \triangleq D_{\text{TV}}^{\max}(\pi, \tilde{\pi})$, then the following bound holds: $J(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)^2} \alpha^2$, where $\epsilon = \max_{s,a} |A_\pi(s, a)|$.*

Since the TV divergence and the Kullback-Leibler (KL) divergence are related as follows: $D_{\text{TV}}^2(\pi, \tilde{\pi}) \leq \frac{1}{2} D_{\text{KL}}(\pi, \tilde{\pi})$, we then have the following $J(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - \frac{2\epsilon\gamma}{(1-\gamma)^2} D_{\text{KL}}^{\max}(\pi, \tilde{\pi})$, where $D_{\text{KL}}^{\max}(\pi, \tilde{\pi}) \triangleq \max_s D_{\text{KL}}(\pi, \tilde{\pi})$. This forms the

foundation of many policy optimization methods, including TRPO [9] and Proximal Policy Optimization (PPO) [10]. The KL divergence imposed over the consecutive policies, π and $\tilde{\pi}$, is also called the trust region. In practice, TRPO adopts a robust way to take large update steps by using a constraint (rather than a penalty) on the KL divergence, and also considers using the expected KL divergence, instead of the maximum over all states:

$$\max_{\tilde{\pi}} \quad \mathbb{E}_{(s,a)\sim d_\pi}\left[\frac{\tilde{\pi}(a|s)}{\pi(a|s)}A_\pi(s,a)\right], \quad \text{s.t.} \quad \mathbb{E}_s\left[D_{\mathrm{KL}}(\pi(\cdot|s), \tilde{\pi}(\cdot|s))\right] \leq \delta, \tag{3}$$

where δ is a hyperparameter to specify the trust region. PPO with ratio clipping further simplifies such trust region constraint by leverage ratio clipping and considers the following optimization problem:

$$\max_{\tilde{\pi}} \mathbb{E}_{d_\pi}\left[\min\left(\frac{\tilde{\pi}(a|s)}{\pi(a|s)}A_\pi, \mathrm{clip}(\frac{\tilde{\pi}(a|s)}{\pi(a|s)}, 1-\epsilon, 1+\epsilon)A_\pi\right)\right], \quad \text{where } \epsilon \text{ is the clipping hyperparameter.} \tag{4}$$

3 Conservative policy optimization

We show in this section that the trust region constraint over policies can be safely substituted by a *trust-region-free* constraint when we consider a generalized form of the surrogate objective function. We also present a monotonic improvement guarantee for stochastic policies with the generalized surrogate objective and the trust-region-free constraint.

3.1 Optimization of stochastic policies

Definition 3.1. Define state-action function: $A(s,a) \triangleq r(s,a) + \mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s')] - f(s)$, where f is a function $f : \mathcal{S} \to \mathbb{R}$.

Consider updating a stochastic policy from π to $\tilde{\pi}$ via policy gradients. The following proposition from [2] is useful.

Proposition 3.2. *For any stochastic policies $\tilde{\pi}$, π, and the state-action function defined above,*

$$J(\tilde{\pi}) - J(\pi) = \mathbb{E}_{s\sim d_{\tilde{\pi}}(s), a\sim\tilde{\pi}}[A(s,a)] - \mathbb{E}_{s\sim d_\pi(s), a\sim\pi}[A(s,a)]. \tag{5}$$

This proposition generalizes (1) to a broader family of functions $A(s,a)$. One can easily verify that the advantage function of π, i.e., $A_\pi(s,a)$, satisfies the definition with f function as the value function. In this case, (5) is equivalent to (1). Furthermore, this proposition implies that the performance difference between any two policies can be described by their state-action distribution shift, i.e., $d_{\tilde{\pi}}(s)\tilde{\pi}(a) - d_\pi(s)\pi(a)$, weighted by a function $A(s,a)$. In practice, it would be very unlikely to have access to $d_{\tilde{\pi}}(s)$. We thus leverage the same trick used in TRPO to substitute $d_{\tilde{\pi}}(s)$ with the state distribution induced by policy π, i.e., $d_\pi(s)$, and consider the following objective:

$$G_\pi(\tilde{\pi}) \triangleq \mathbb{E}_{s\sim d_\pi(s), a\sim\pi(\cdot|s)}\left[\left(\frac{\tilde{\pi}(a|s)}{\pi(a|s)} - 1\right)A(s,a)\right]. \tag{6}$$

This new objective generalizes the surrogate objective in (3) to any function defined in 3.1. We have the following bound,

Theorem 3.3. *For any two stochastic policies $\tilde{\pi}$ and π, the following bound holds:*

$$J(\tilde{\pi}) - J(\pi) \geq G_\pi(\tilde{\pi}) - \frac{2\gamma}{1-\gamma}(\delta + \epsilon), \quad \text{where } \delta = \max_{s,a}\left|\left(\frac{\tilde{\pi}(a|s)}{\pi(a|s)} - 1\right)A(s,a)\right| \text{ and } \epsilon = \left|\sum_a \pi(a|s)A(s,a)\right|.$$

To simplify further analysis, we call $\left(\frac{\tilde{\pi}(a|s)}{\pi(a|s)} - 1\right)$ the *ratio deviation*. This theorem states that the policy improvement gap can be effectively bounded by the maximum product of the ratio deviation $\left(\frac{\tilde{\pi}(a|s)}{\pi(a|s)} - 1\right)$ and the state-action function $A(s,a)$. Moreover, as this product also appears in the definition of $G_\pi(\tilde{\pi})$ in (6), one can thus consider constraining it when optimizing $G_\pi(\tilde{\pi})$. This is what we call the *conservative policy optimization*. We discuss how this conservative policy update can be implemented in practice in the next section.

Theorem 3.3 differs from Theorem 2.1 in two respects. First, it presents a lower bound for the performance improvement with respect to a state-action function defined in Definition 3.1. This $A(s,a)$ does not necessarily need to be the advantage function. Second, instead of imposing the TV constraint over the policies as in Theorem 2.1, the above theorem considers the maximum product of the ratio deviation and the state-action function. According to [11], the TV constraint between any two policies can be equivalently translated into a constraint over ratio deviations. In this sense, TRPO is essentially a special case of Theorem 3.3 by leveraging the TV to bound the ratio deviations under the assumption that the advantage function should be small. Namely, TRPO optimizes the policy regardless of how the magnitude of the advantage might change throughout optimization. Consequently, TRPO may fail to optimize the policy when the advantage function is large in magnitude at some state-action sample even though the TV divergence is well bounded at one iteration.

3.2 Practical policy optimization methods

We now present the practical policy optimization methods for stochastic policies. We first offer intuitions to understand the underlying idea of conservative policy optimization in the above theorems.

Theorem 3.3 is closely related to some existing policy optimization methods. For example, optimizing $G_\pi(\tilde{\pi})$ without any conservative constraint is equivalent to the policy gradient method [12]. Also, there are two ways to impose such conservative constraints: the ratio-conservative and the objective-conservative, which refer to removing the incentive of increasing the ratio deviations (corresponding to PPO [10]) or the objective, respectively, when optimizing the objective function.

Non-conservative   With no consideration of the conservative policy update principle, one can directly optimize $G_\pi(\tilde{\pi})$ with the state-action function chosen as the advantage function of π. Such policy optimization is performed by *policy gradient methods* [12]: $\max_\theta \quad \mathbb{E}_{(s,a)\sim d_\pi}\left[\frac{\tilde{\pi}_\theta(s,a)}{\pi(s,a)}A_\pi(s,a)\right]$. Theorem 3.3 implies that optimizing the above objective with the same set of sampled data for multiple times (i.e., multi-epoch optimization as in [10]), could incur a significant degradation in policy performance, since the product between the ratio deviation and the advantage can be large. Thus, applying policy gradients for multiple epoch optimization does not guarantee policy improvement [6].

Ratio-conservative   One can take into account the conservative policy update rule by constraining the ratio deviations. Specifically, when optimizing $G_\pi(\tilde{\pi})$ with the state-action function as the advantage, one can clip the ratio deviations to remove the incentive of inducing unexpected large deviations (i.e., *ratio-conservative*), as follows $\max_\theta \quad \mathbb{E}_{(s,a)\sim d_\pi}\left[\text{clip}\left(\frac{\tilde{\pi}_\theta(s,a)}{\pi(s,a)}-1,-\lambda,\lambda\right)A_\pi(s,a)\right]$, where $\lambda > 0$ is a hyper-parameter for ratio deviation clipping. This new objective resembles the ratio clipping objective (4) used in *PPO* [10], which has been shown to be effective in practice, especially with a normalized advantage function. However, such ratio clipping scheme ignores the potential effect of the advantage function on the ratio deviation, and thus can fail to monotonically improve the policy performance when the advantage is large at some state-action point and dominates $\left(\frac{\tilde{\pi}_\theta(s,a)}{\pi(s,a)}-1\right)A_\pi(s,a)$ for a small ratio deviation. Furthermore, solely bounding the divergence between policies, as used in trust region methods for policy optimization [9, 2], may not be a good option in practice, as it is not sufficient for policy improvement when the advantage function fluctuates greatly across state-action samples.

Objective-conservative   We can instead apply the clipping scheme to the objective. Namely, we can clip the objective directly to achieve this conservative update principle, as follows:

$$\max_\theta \quad \mathbb{E}_{(s,a)\sim d_\pi}\left[\min\left(\left(\frac{\tilde{\pi}_\theta(s,a)}{\pi(s,a)}-1\right)A_\pi(s,a),\delta\right)\right],$$

where $\delta > 0$ is a hyper-parameter to control the conservativeness when optimizing the objective. The δ operates as a threshold beyond which the objective quantity have no contribution to the optimization. We call it *objective conservative*, and the resulting algorithm *Trust-REgion-Free Policy Optimization (TREFree)* as it is free of any explicit trust region constraint. TREFree is detailed in Algorithm 1.

Algorithm 1 TREFree algorithm

> **for** iterations $i = 1, 2, ...$ **do**
>   **for** actor $= 1, 2, ..., N$ **do**
>   Run policy π in environment
>   Compute advantage estimates $\hat{A}_\pi$
>   **end for**
>   **for** epoch $= 1, 2, ..., K$ **do**
>   Sample M samples $\{(s,a)\}$ from previous rollouts.
>   Compute $\mathcal{L}(\theta) \triangleq \frac{1}{M}\sum_{s,a}\min\left(\left(\frac{\tilde{\pi}_\theta(a|s)}{\pi(a|s)}-1\right)\hat{A}_\pi(s,a),\delta\right)$.
>   Maximize $\mathcal{L}(\theta)$ w.r.t θ via gradient descent.
>   **end for**
> $\pi \leftarrow \tilde{\pi}_\theta$.
> **end for**

---

4   Experiments

In this section, we compare TREFree with TRPO and PPO across the Mujoco continuous control tasks. We adopt the same strategy as in TRPO [9] and PPO [10] to normalize the observations, rewards, and advantages. Specifically, the observations, rewards and advantages are normalized to zero mean and unit variance using a running mean and standard deviation. Both observations and rewards are normalized using a running mean and standard deviation per-timestep over the whole training process, while the advantages are normalized only within a training batch. Moreover, we leverage the actor-critic framework by parameterizing the actor and critic with 2-layered perceptrons, each of which has 64 hidden units and is activated with $\tanh$. Also, the actor and critic share parameters by reusing the first layer of their neural networks, which has been reported to stabilize the training and improve performance [9, 10, 4]. The policy is modeled as a Gaussian distribution, with mean and variance parameterized by the actor neural network.

We now compare TREFree with TRPO and PPO across the Mujoco continuous control tasks with different control complexity [4], We used the publicly available and widely used repository (https://github.com/openai/baselines) as the baseline implementation. For TRPO and PPO, we use the default hyper-parameters given in [9, 10]. We also sweep over
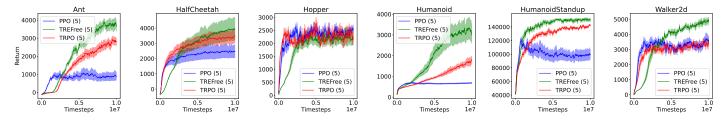
Figure 1: Contrasting TREFree with TRPO and PPO on Mujoco benchmark tasks; training with a margin size 0.01.

the clipping range for PPO and use the best performing value as the baseline. Furthermore, we heuristically decay the learning rate of TREFree and PPO linearly from 0.0003 to 0, as we found this annealing strategy stabilizes training for both PPO and TREFree. We heuristically set δ in TREFree to 0.01 as it is found to perform well across all tasks.

The performance comparison on the Mujoco continuous control tasks is presented in Figure 1. Overall, TREFree performs better than the baselines on all the tasks except Hopper. TREFree outperforms TRPO and PPO by a large margin in terms of the final policy performance on tasks Ant, HalfCheetah, Humanoid, HumanoidStandup and Walker2d. These five Mujoco environments are more complicated than Hopper. The training curves in Figure 1 also show how TRE-Free often outpaces other baselines in improving policy performance. Though TREFree is outperformed by PPO and TRPO on Hopper, the performance curves of all these methods in this specific environment fluctuates greatly over time, and overlap each other.

We also report the ratio ranges of different methods in Figure 2 to show the underlying differences between TREFree and the baseline methods. The probability ratios $\frac{\tilde{\pi}(a\,|\,s)}{\pi(a\,|\,s)}$ are an important indicator in TRPO and PPO training as they are closely related to the total variation divergence [10]. Figure 2 shows that the ratios in both TRPO and TREFree are better bounded than in PPO, where the ratios grow without bound. However, TREFree constrains ratios in a dramatically different way from TRPO: TRPO bounds ratios between $[-2, 1]$ (log-scale) in a symmetrical way, while TREFree bounds ratios between $[0, 1]$, which implies that the policy is most often updated to increase the probability at empirical samples. This contrast in ratio ranges suggests that TREFree is fundamentally different from trust region methods.
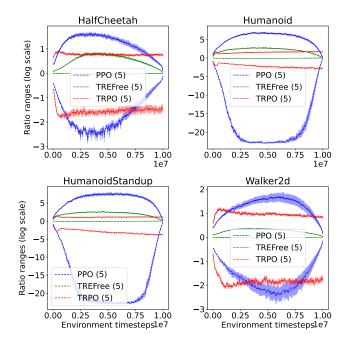

Figure 2: Contrasting ratio ranges.

References

[1] Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. *arXiv preprint arXiv:1806.06920*, 2018.

[2] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

[3] Riad Akrour, Joni Pajarinen, Jan Peters, and Gerhard Neumann. Projections for approximate policy iteration algorithms. In *International Conference on Machine Learning*, pages 181–190. PMLR, 2019.

[4] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[5] Matteo Hessel, Ivo Danihelka, Fabio Viola, Arthur Guez, Simon Schmitt, Laurent Sifre, Theophane Weber, David Silver, and Hado van Hasselt. Muesli: Combining improvements in policy optimization. *arXiv preprint arXiv:2104.06159*, 2021.

[6] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *In Proc. 19th International Conference on Machine Learning*. Citeseer, 2002.

[7] Sham M Kakade. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

[8] Fabian Otto, Philipp Becker, Ngo Anh Vien, Hanna Carolin Ziesche, and Gerhard Neumann. Differentiable trust region layers for deep reinforcement learning. *arXiv preprint arXiv:2101.09207*, 2021.

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

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

[11] Mingfei Sun, Vitaly Kurin, Guoqing Liu, Sam Devlin, Tao Qin, Katja Hofmann, and Shimon Whiteson. You may not need ratio clipping in ppo. *arXiv preprint arXiv:2202.00079*, 2022.

[12] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[13] Yuhuai Wu, Elman Mansimov, Roger B Grosse, Shun Liao, and Jimmy Ba. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. *Advances in neural information processing systems*, 30:5279–5288, 2017.