Separated Trust Regions Policy Optimization Method

Luobao Zou*
Zhiwei Zhuang*
leiling@sjtu.edu.cn
zzw1993@sjtu.edu.cn
Shanghai Jiao Tong University
Shanghai, China

Yin Cheng
Shanghai Jiao Tong University
Shanghai, China
ycheng_sjtu@foxmail.com

Xuechun Wang
Shanghai Jiao Tong University
Shanghai, China
xuechun_wang@sjtu.edu.cn

Weidong Zhang†
Shanghai Jiao Tong University
Shanghai, China
wdzhang@sjtu.edu.cn

ABSTRACT

In this work, we propose a moderate policy update method for reinforcement learning, which encourages the agent to explore more boldly in early episodes but updates the policy more cautious. Based on the maximum entropy framework, we propose a softer objective with more conservative constraints and build the separated trust regions for optimization. To reduce the variance of expected entropy return, a calculated state policy entropy of Gaussian distribution is preferred instead of collecting log probability by sampling. This new method, which we call separated trust region for policy mean and variance (STRMV), can be view as an extension to proximal policy optimization (PPO) but it is gentler for policy update and more lively for exploration. We test our approach on a wide variety of continuous control benchmark tasks in the MuJoCo environment. The experiments demonstrate that STRMV outperforms the previous state of art on-policy methods, not only achieving higher rewards but also improving the sample efficiency.

KEYWORDS

Reinforcement learning, Trust region, Entropy maximization

ACM Reference Format:
Luobao Zou, Zhiwei Zhuang, Yin Cheng, Xuechun Wang, and Weidong Zhang. 2019. Separated Trust Regions Policy Optimization Method. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330892

1 INTRODUCTION

The increased computation power promotes the development of reinforcement learning (RL) for more difficult tasks. Especially, a variety of approaches have been proposed and successfully applied in some certain settings like games playing [11, 22] and robotic control [6]. The leading contenders in the domain of RL can be roughly divided into two categories said by Sutton [24]: (1) value-based methods, whose policies would not be available without an estimation of the action-value. (2) parameterized policy-based methods, whose agent selects actions without consulting a value function directly.

The typical representations of value-based approaches are the family of deep Q-learning methods [16, 26, 27] with experience replay buffer. Those Q-learning methods often obtain quick performance improvement in the early exploration episodes since the technique of experience replay can dramatically increase the sampling efficiency. By contrast, policy gradient methods [8–10, 18, 19, 23], the core of parameterized policy-based approaches, can master various continuous tasks rather than be restricted to discrete domains as Q-learning methods. However, most of those gradient methods suffer from the sample inefficiency and parameter sensitivity [9]. The brittleness of those policy gradient algorithms lies in the update size, which is controlled by a hyperparameter called learning rate. Each time when the set update size is not appropriate, the agent may be guided to execute a worse policy, which could circularly lead to the divergence of value estimation network. John Schulman et al. finds a surrogate estimation function that ensures policy improvement within a trust region [18]. This method, trust region policy optimization (TRPO), gains popularity for its robust and empirical performance on difficult continuous control tasks. But it is relatively complicated and impractical for scalability. To address this issue, Yuhuai Wu et al. suggest using Kronecker-factored approximate curvature with trust region to optimize both the actor and critic (ACKTR) [28]. Besides, proximal policy optimization (PPO) has been proposed using the clipped probability ratio in [20].

Moreover, another inspiring research pointed out that excessively aggressive rules of action selection adopted by the most mainstream algorithm are preventing further performance promotion [5, 13]. For the sake of exploration sufficiency, Ziebart et al. advocates a maximum entropy framework, which adds an entropy term to the standard maximum reward objective [30]. Instead of regarding the policy entropy as a regularizer [4, 10, 20], [13, 14, 25] begin considering the state policy entropy as rewards and maximizes both expected reward and entropy. Recently, Haarnoja et

*Both authors contributed equally to this research.
†Weidong Zhang is the corresponding author of this paper.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330892

al. successfully applied DDPG style policy gradients to the maximum entropy framework in [6]. This new approach called soft actor-critic (SAC) obtains impressive performance on a certain high-dimensional task. Nachum et al. tempt to extend trust region methods to entropy maximum framework and adopt the optimization objective in TRPO [18] intuitively, ignoring the changes in the policy entropy [14].

This paper seeks to propose a more reliable policy update method that guarantees the training stability and robustness while maintains the simplicity and data efficiency for the maximum entropy formulation. We introduce a new construction framework of policy entropy to reduce the variance of expected entropy return. From a unique perspective, we have deduced the lower bound of performance estimation function for the maximum entropy framework, yielding a softer policy objective with a stricter trust region. Based on this maximum problem, we build the isolated trust regions for policy mean and variance respectively and propose a more practical algorithm, which we call separated trust regions of mean-variance (STRMV). Our experiments have demonstrated that the proposed algorithm gains better performance against the prior state-of-the-art methods on most MuJoCo environments.

2 PRELIMINARIES

2.1 Standard Reinforcement Learning and Maximum Entropy

Consider a discounted markov decision process (MDP). At each discrete time step t, the agent observes a state $s_t \in S$ and takes a action $a_t \in A$ in accordance with its policy $\pi(a|s_t)$, receiving a reward $r_t \in \mathbb{R}$ from the environment and transferring to a new state s_{t+1} with the probability $P(s_{t+1}|s_t, a_t)$. Thus the standard value function V_π, the state-action value function Q_π^V, and the reward advantage function A_π^V can be defined as:

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k} \,|s_t = s \right\} \\ Q_\pi^V(s_t, a_t) &= r(s_t, a_t) + \gamma V(s_{t+1}) \\ A_\pi^V(s_t, a_t) &= Q_\pi^V(s_t, a_t) - V_\pi(s_t) \end{aligned} \tag{1}$$

The goal of standard reinforcement learning is to maximize the expectation of the discounted return

$$J_V(\pi) = \mathbb{E}_{\tau\sim\pi}[R_t] = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] \tag{2}$$

where $\mathbb{E}_{\tau\sim\pi}[\cdot]$ means that the sampling trajectory τ is obtained by taking actions from policy $\pi(\cdot|s)$

More generally, Ziebart [29, 30], R. Fox [3], Nachum [14] and Haarnoja [5, 6] consider maximizing both expected reward and policy entropy by adding the expected policy entropy over ρ_π into the policy performance:

$$J_T(\pi) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t [r(s_t, a_t) + \alpha \mathbb{H}(\pi(\cdot|s_t))] \tag{3}$$

Here, the hyperparameter α balances the magnitude of policy entropy and reward. As the time step increases, α keeps decreasing until zero is met. With this entropy maximization framework [6], the soft state value function T_π can be computed iteratively according to the modified Bellman equation:

$$T_\pi(s_t) = \mathbb{E}_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k (r_{t+k} - \log \pi(a_t|s_t)) \,|s_t = s \right\} \tag{4}$$

According to the softmax temporal consistency in [13], the discounted policy entropy $\mathbb{H}_\pi(s_t)$ at each state s_t can be perfectly isolated and defined recursively as:

$$\mathbb{H}_\pi(s_t) = \mathbb{E}_{a_t\sim\pi}[-\log \pi(a_t|s_t) + \gamma \mathbb{H}_\pi(s_{t+1})] \tag{5}$$

In this case, the soft state value can be expressed as a sum of the discounted reward and entropy term, which holds for all states.

$$T_\pi(s) = V_\pi(s) + \alpha \mathbb{H}_\pi(s) \tag{6}$$

2.2 Trust Region and Proximal Policy Optimization

In [7], the difference of policy performance under a new policy $\tilde{\pi}$ relative to the current policy π can be expressed in terms of the reward advantage function A_π^V over policy π:

$$J_V(\tilde{\pi}) - J_V(\pi) = \mathbb{E}_{\tau\sim\tilde{\pi}}[\sum_{t=0}^{\infty} \gamma^t A_\pi^V(s_t, a_t)] \tag{7}$$

Note that the trajectory τ is generated under the new policy $\tilde{\pi}$. John Schulman et al. propose a new objective function subjected to a step-size constraint of the policy update (see TRPO [18] for specific derivation).

$$\begin{aligned} &\underset{\theta}{maximize}\, \mathbb{E}_{(s,a)\sim\rho_{\theta_{old}}}[\frac{\pi_\theta(a|s)}{\pi_{\theta_{old}}(a|s)} A_{\theta_{old}}^V(s,a)] \\ &\text{subject to } \mathbb{E}_{s\sim\rho_{\theta_{old}}}\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot|s) \,\|\, \pi_\theta(\cdot|s)\right)\right] \le \delta \end{aligned} \tag{8}$$

where θ_{old} denotes the parameters vector of the old policy π, i.e. π, $\pi_{\theta_{old}}$ and θ_{old} all denote the policy. In the following section, we will use these representations alternately depending on whether network parameters are involved for succinctness.

To simplify the calculation process of TRPO, Schulman et al. modify the surrogate objective with the clipped probability ratios and propose a simpler and more reliable approach [20]. Specifically,

$$\underset{\theta}{\text{maximize}}\, \mathbb{E}_{(s,a)\sim\rho_{\theta_{old}}}\left[\min\left(f_t(\theta)\hat{A}_t, \text{clip}\left(f_t(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_t\right)\right] \tag{9}$$

where $f_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ and $\hat{A}_t = A_{\theta_{old}}^V(s_t, a_t)$.

Here, $f_t(\theta)$ denotes the probability ratio and ϵ determines the clip margin.

3 SOFT POLICY WITH TIGHT TRUST REGION

3.1 Variance-reduced Entropy Function

In the entropy framework of prior works, when taking an action a_t at the state s_t obeying the policy π, the agent observes a partial policy entropy $e(s_t, a_t) = -\log \pi(a_t|s_t)$ rather than the complete state policy entropy. But actually, when the action distribution $a_t \sim \pi(a|s_t)$ is available for the agent, the state policy entropy $E(s_t, \pi)$ can be easily calculated over actions, $E(s_t, \pi) = \mathbb{E}_{a_t\sim\pi}[-\log \pi(a_t|s_t)] = -\sum_a \pi(a|s_t) \log \pi(a|s_t)$.

More importantly, gathering cumulative entropy by collecting log probability of samples could lead to huge variance, hindering the performance improvement of the entropy approximator. As with previous studies in the continuous action domains, we assume the action policy in each state as a Gaussian with mean and variance given by neural networks, i.e. $\pi(a_t|s_t) \sim \mathcal{N}(\mu(s_t, \pi), \sigma(s_t, \pi)^2)$.

Thus, the state policy entropy can be obtained by (See Appendix A for details)

$$\begin{aligned} E(s_t,\pi) &= -\int_{-\infty}^{\infty} \pi(a\,|s_t)\log\pi(a\,|s_t)\,da \\ &= \tfrac{1}{2}\log\left(2\pi e\sigma(s_t,\pi)^2\right) = \log\sigma(s_t,\pi) + b \end{aligned} \tag{10}$$

Here, the constant $b = \frac{1}{2}\log(2\pi e)$ can be ignored and the agent receives a state policy entropy $E(s_t,\pi) = \log\sigma(s_t,\pi)$ for being a state s_t instead.

From the equation above, the state policy entropy determined only by variance is irrelevant to the mean of the distribution. This discovery gives birth to the following individual optimization of policy mean and policy variance. With the variance-reduced entropy construction, we define the discounted entropy performance

$$J_H(\pi) = \mathbb{E}_{\tau\sim\pi}[E(s_t,\pi)] = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty}\gamma^t E(s_t,a_t)] \tag{11}$$

Moreover, the state and state-action entropy function be written recursively as:

$$\begin{aligned} \mathbb{H}_\pi(s_t) &= E(s_t,\pi) + \gamma\mathbb{E}_{a_t\sim\pi}\left[\mathbb{H}_\pi(s_{t+1})\right] \\ Q_\pi^H(s_t,a_t) &= E(s_t,\pi) + \gamma\mathbb{H}_\pi(s_{t+1}) \end{aligned} \tag{12}$$

Note that $\mathbb{H}_\pi(s_t)$ and $Q_\pi^H(s_t,a_t)$ are similar in form, but $\mathbb{H}_\pi(s_t)$ is the entropy expectation over all the actions at a given state s_t while $Q_\pi^H(s_t,a_t)$ only on action a_t.

3.2 Soft Policy With Tight Trust Region

To guarantee that the performance cannot get worse with policy updated each time, an estimation function that indicates how much improvement has been made with a new policy compared to an old one is required. Here comes the first significant lemma for the maximum entropy framework, which describes that the difference of soft policy performance $J_T(\tilde\pi) - J_T(\pi)$ can be decomposed as several independent parts.

Lemma 3.1. *Consider the definition of soft policy performance in eq.3 and Bellman equations eq.1 & eq.12. Given two policies $\tilde\pi$ and π,*

$$\begin{aligned} J_T(\tilde\pi) - J_T(\pi) = \mathbb{E}_{\tau\sim\tilde\pi}\Bigg[\sum_{t=0}^{\infty}\gamma^t\Big(A_\pi^V(s_t,a_t) \\ +\alpha\left(A_\pi^H(s_t,a_t) + E(s_t,\tilde\pi) - E(s_t,\tilde\pi)\right)\Big)\Bigg] \end{aligned} \tag{13}$$

where $A_\pi^V(s_t,a_t) = Q_\pi^V(s_t,a_t) - V_\pi(s_t)$, $A_\pi^H(s_t,a_t) = Q_\pi^H(s_t,a_t) - \mathbb{H}_\pi(s_t)$ are the advantage function of discounted reward and entropy respectively. (See Appendix B for proof)

From the perspective of the softmax temporal consistency [13] and equation(6), it's easy to have $J_T(\pi) = J_V(\pi) + \alpha J_H(\pi)$. Let Equation(13) subtract Equation(7), we get

$$J_H(\tilde\pi) - J_H(\pi) = \mathbb{E}_{\tau\sim\tilde\pi}[\sum_{t=0}^{\infty}\gamma^t(A_\pi^H(s_t,a_t) + E(s_t,\tilde\pi) - E(s_t,\pi))] \tag{14}$$

Equation(14) indicates that the difference on the discounted entropy performance is the sum of per-time-step entropy advantages and relative state policy entropy, while the entropy regularized expected reward objective proposed in [14] has totally ignored the relative entropy term, which is considered significant for wide exploration in this paper.In the following sections, we will directly use A_π^V and A_π^H to represent the reward and entropy advantage functions.

The state marginals of the trajectory distribution ρ_π records the discounted visitation probability of state s under policy π and it can be expressed by the normalized visitation frequency [18]. Here we ignored the normalization for the time being, and get

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \cdots \tag{15}$$

To eliminate the dependence of formulas on time series, equation(13) can be reorganized as a sum over states (see Appendix B the detailed process)

$$\begin{aligned} J_T(\tilde\pi) - J_T(\pi) = \alpha\log\frac{\sigma(s_0,\tilde\pi)}{\sigma(s_0,\tilde\pi)} \\ +\sum_s\rho_{\tilde\pi}(s)\sum_a\tilde\pi(a\,|s)\left[A_\pi^V + \alpha\left(A_\pi^H + \log\frac{\sigma(s',\tilde\pi)}{\sigma(s',\pi)}\right)\right] \end{aligned} \tag{16}$$

where s_0 denotes the initial state and s' the next state.

This lemma indicates that when the policy entropy is included, the agent also needs to consider the relative state policy entropy and the potential entropy of those trajectories. As in equation(16), $\log\frac{\sigma(s',\tilde\pi)}{\sigma(s',\pi)}$ is the relative entropy of the new policy over the old one, and A_π^H denotes the relative measure of value of the expected entropy when selecting the action a_t at a given state s_t. Thus, the agent will be guided towards not only high-reward regions but also high-entropy regions and be encouraged to explore more widely while abandoning apparently hopeless avenues in the meantime.

To simplify this equation, we replace the visitation frequency $\rho_{\tilde\pi}(s)$ with $\rho_\pi(s)$ as in [7, 18] and get a new function:

$$\begin{aligned} F_\pi(\tilde\pi) = J_T(\pi) + \alpha\log\frac{\sigma(s_0,\tilde\pi)}{\sigma(s_0,\pi)} \\ +\sum_s\rho_\pi(s)\sum_a\tilde\pi(a\,|s)\left[A_\pi^V + \alpha\left(A_\pi^H + \log\frac{\sigma(s',\tilde\pi)}{\sigma(s',\pi)}\right)\right] \end{aligned} \tag{17}$$

Note that $F_\pi(\tilde\pi)$ has ignored the variety of state visitation frequency caused by the new policy $\tilde\pi$. For the purpose of finding out the mathematical relationship between $F_\pi(\tilde\pi)$ and $J_T(\tilde\pi)$, [7] has proposed a conservative update rule that makes the update amplitude controllable with parameter β :

$$\pi_\theta(\cdot\,|s) = (1-\beta)\,\pi_{\theta_{old}}(\cdot\,|s) + \beta\pi_{\theta'}(\cdot\,|s) \tag{18}$$

Here, β evaluates the gap between the new policy π_θ and the current one $\pi_{\theta_{old}}$, which is used to identify the estimation offset with the variety of state visitation frequency ignored. Remember that $\tilde\pi$ and π_θ both imply the new policy as the mentioned before and $\pi_{\theta'}$ denotes an expected update policy. With the conservative update rule, we have the following theorem (see Appendix C for details).

Theorem 3.2. *Let $\delta = \max_s|A_\pi^V(s_t,a_t) + \alpha A_\pi^H(s_t,a_t)|$. If the expected update policy $\pi_{\theta'}(\cdot|s)$ satisfies $\max_s|\log\frac{\sigma(s,\pi_{\theta'})}{\sigma(s,\pi_{\theta_{old}})}| < \omega$, and correspondingly let $k = \arg\max_k\{|\log(1-\beta)^2 + \beta^2 k)|\}, k \in \{e^\omega, e^{-\omega}\}$. The following conditional bound holds:*

$$J_T(\tilde\pi) \geq F_\pi(\tilde\pi) - \frac{2\beta\gamma\left(\beta\delta + \alpha\left|\log\left((1-\beta)^2 + \beta^2 k\right)\right|\right)}{(1-\gamma)^2} \tag{19}$$

It is worth mentioning that KL-divergence between the new and old policy (using Gaussian distribution) satisfies

$$D_{KL}(\pi(\cdot\,|s))\,\|(\tilde\pi(\cdot\,|s)) \geq \frac{1}{2}\log\left((1-\beta)^2 + \beta^2\chi\right) + m\beta^2 - \frac{1}{2} \tag{20}$$

where $\chi = \frac{\sigma_2^2}{\sigma_1^2}$ and $m = \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}$ can be seen as two constant. It's interesting to note that inequality of KL-divergence is similar to our difference term as shown in equation(19). We believe this general theorem can partially explain why TRPO has chosen KL-divergence to replace β^2 [18].

However, KL-divergence cannot be directly used here because the difference term has introduced ABS function and there is no guarantee that maximal KL-divergence in all states is always larger than the difference term. Besides, this theorem is over complicated and the constraint $\max_s |\log \frac{\sigma(s,\pi_{\theta'})}{\sigma(s,\pi_{\theta_{old}})}| < \omega$ can not be met directly since we operates on the new policy π_θ instead of $\pi_{\theta'}$.

To simplify theorem 3.2, we directly make a constraint on the new policy π_θ and get a more practical result:

if $\max_s |\log \frac{\sigma(s,\pi_\theta)}{\sigma(s,\pi_{\theta_{old}})}| \le \frac{\zeta}{\alpha}$ holds with the new policy $\tilde{\pi}$, then

$$J_T(\tilde{\pi}) \ge F_\pi(\tilde{\pi}) - \frac{2\beta\gamma(\beta\delta+\zeta)}{(1-\gamma)^2} \tag{21}$$

Since the temperature parameter α will gradually decrease to zero, We add a small positive number ℓ to the denominator of the relative policy entropy boundary to avoid a division by zero, i.e. $\max_s |\log \frac{\sigma(s,\pi_\theta)}{\sigma(s,\pi_{\theta_{old}})}| \le \frac{\zeta}{\alpha+\ell}$.

At this point, we turn to optimize function $F_\pi(\tilde{\pi}) - \frac{2\beta\gamma(\beta\delta+\zeta)}{(1-\gamma)^2}$ at each iteration, which can still guarantee that the true objective $J_T(\tilde{\pi})$ will not decrease as discussed by [18]. Rather than a penalty, and a constraint can be more robust practically, i.e.

$$\begin{aligned} &\underset{\theta}{\text{maximize}}\ F_{\pi_{\theta_{old}}}(\pi_\theta) \\ &s.t. A\beta^2 + B\beta \le \kappa \\ &\max_s \left|\log \frac{\sigma(s,\pi_\theta)}{\sigma\left(s,\pi_{\theta_{old}}\right)}\right| \le \frac{\zeta}{\alpha+\ell} \end{aligned} \tag{22}$$

where $A = \frac{2\gamma\delta}{(1-\gamma)^2}, B = \frac{2\gamma\zeta}{(1-\gamma)^2}$.

Note that in the objective function $F_\pi(\tilde{\pi})$, $J_T(\pi)$ is a constant w.r.t θ and $\alpha \log \frac{\sigma(s_0,\tilde{\pi})}{\sigma(s_0,\pi)}$ worked only when initial states are visited with a discounted visitation probability $\frac{1}{1-\gamma} \ll 1$. Both of those two terms can be ignored for simpler optimization without degrading its performance. Besides, since $A\beta^2 + B\beta$ is an increasing function w.r.t β when $\beta > 0$, the constraint $A\beta^2 + B\beta \le \kappa$ is equivalent to $\beta^2 \le o$, where $o = \left(\sqrt{\kappa + \frac{B^2}{2A}} - \frac{B}{2A}\right)^2$.

Hence, when we replace β^2 with the average KL-divergence, the simplified optimization objective can be

$$\begin{aligned} &\underset{\theta}{\text{maximize}}\ \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_{\theta_{old}}}\left[\frac{\pi_\theta(\cdot|s)}{\pi_{\theta_{old}}(\cdot|s)}\left(A^V_{\theta_{old}} + \log\frac{\sigma(s',\pi_\theta)}{\sigma(s',\pi_{old})}\right)\right] \\ &s.t. \mathbb{E}_{(s,a)\sim\rho_{\theta_{old}}}\left[D_{KL}\left(\pi_{\theta_{old}}(\cdot|s)\,|\pi_\theta(\cdot|s)|\right)\right] \le \xi \\ &\max_s \left|\log\frac{\sigma(s',\pi_\theta)}{\sigma(s',\pi_{old})}\right| \le \frac{\varsigma}{\alpha+\ell} \end{aligned} \tag{23}$$

Here, $\sum_s \rho_\pi(s)[\cdot]$ in the objective can be replaced by the expectation $\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\rho_{\theta_{old}}}[\cdot]$. Note that each symbol appearing above, such as $\omega, \zeta, \kappa, o, \xi$ and ℓ, refers to a small positive super-parameter.

From our optimization problem in equation(23), we discover an interesting phenomenon that when the agent is guided to explore more widely, its trust region of policy update would correspondingly shake, that is, the policy update should be more cautious. Besides, as the temperature parameter drops slowly, the constraint on relative state policy entropy will become looser until the optimization problem recovers to a normal reward-guided problem.

3.3 Separated Trust Regions of Mean-Variance

One may consider that the bound on relative policy entropy repeats KL-divergence constraint, i.e. ζ has a strong coupling relationship with β. And a more practical problem is how to realize the update constraint on policy variance for each state. Actually, looking into the objective function, we will find that the agent is guided to maximize $F_\pi(\tilde{\pi})$ within a small neighborhood around the old policy $\pi_{\theta_{old}}$(namely trust region). To achieve similar effect, a natural idea is to clip the action probability ratio of new and old policy into a small range around 1 and turn to optimize the clipped surrogate objective function. This approach comes from PPO [20] and gains reputation for its simplicity and sample efficiency.

Similarly, we apply the clip skill to our optimization problem by clipping both the action probability ratio and the relative state policy entropy, which plays a role similar to two constraints in equation(22). And our clipped surrogate objective can be constructed as:

$$\begin{aligned} C(\theta) = \mathbb{E}_{(s,a)\sim\rho_{\theta_{old}}}[&\min(f_t(\theta)(\hat{A}^T_t + \alpha v_{t+1}(\theta)), \\ &clip(f_t(\theta), 1-\epsilon, 1+\epsilon)(\hat{A}^T_t + \alpha\cdot clip(v_{t+1}(\theta), \tfrac{-\eta}{\alpha+\ell}, \tfrac{\eta}{\alpha+\ell}))] \end{aligned} \tag{24}$$

Here, $f_t(\theta) = \frac{\pi_\theta(\cdot|s)}{\pi_{\theta_{old}}(\cdot|s)}$ is the action probability ratio and $v_{t+1}(\theta) = \log\frac{\sigma(s_{t+1},\pi_\theta)}{\sigma(s_{t+1},\pi_{\theta_{old}})}$ indicates the relative policy entropy of next state, so $f_t(\theta_{old}) = 1$ and $v_{t+1}(\theta_{old}) = 0$. ϵ and η are the clip margins of the probability ratio f and relative entropy v respectively and ℓ is a much smaller positive value, avoiding a division by zero error. Besides, $\hat{A}^T_t = A^V_{\theta_{old}}(s_t,a_t) + \alpha A^H_{\theta_{old}}(s_t,a_t)$ presents the soft advantage function. However, this objective always leads to network divergence and infinite policy variance.

Our analysis believes that the clipped probability ratio should be blamed since it may allow policy variance to vary over a wide range while keeping the probability ratio $f_t(\theta)$ within a certain range. Consider the logarithm of probability ratio,

$$\begin{aligned} \log f_t(\theta) &= \log\frac{\pi_\theta(a_t|s)}{\pi_{\theta_{old}}(a_t|s)} \\ &= \frac{\left(a_t-\mu_{\theta_{old}}\right)^2}{2\sigma_{\theta_{old}}{}^2} - \frac{(a_t-\mu_\theta)^2}{2\sigma_\theta{}^2} - \log\frac{\sigma_\theta}{\sigma_{\theta_{old}}} = \psi_t(\theta) - v_t(\theta) \end{aligned} \tag{25}$$

where $\psi_t(\theta) = \frac{(a_t-\mu_{\theta_{old}})^2}{2\sigma_{\theta_{old}}{}^2} - \frac{(a_t-\mu_\theta)^2}{2\sigma_\theta{}^2}$. Note that when $\sigma_{\theta_{old}}{}^2 \gg 0$ (in the early episodes), σ_θ within a small neighborhood around $\sigma_{\theta_{old}}$ can hardly influence $\psi_t(\theta)$ while μ_θ contributes much to it. Thus, the impact of updated variance on $\psi_t(\theta)$ is almost negligible, which means $\psi_t(\theta)$ can be roughly regarded as a function of policy mean μ. This split achieves a domain separation of policy mean and policy variance, which essentially decouples two constraints on relative policy entropy and KL-divergence in as shown in equation(23). Based on this domain separation, our clipped surrogate

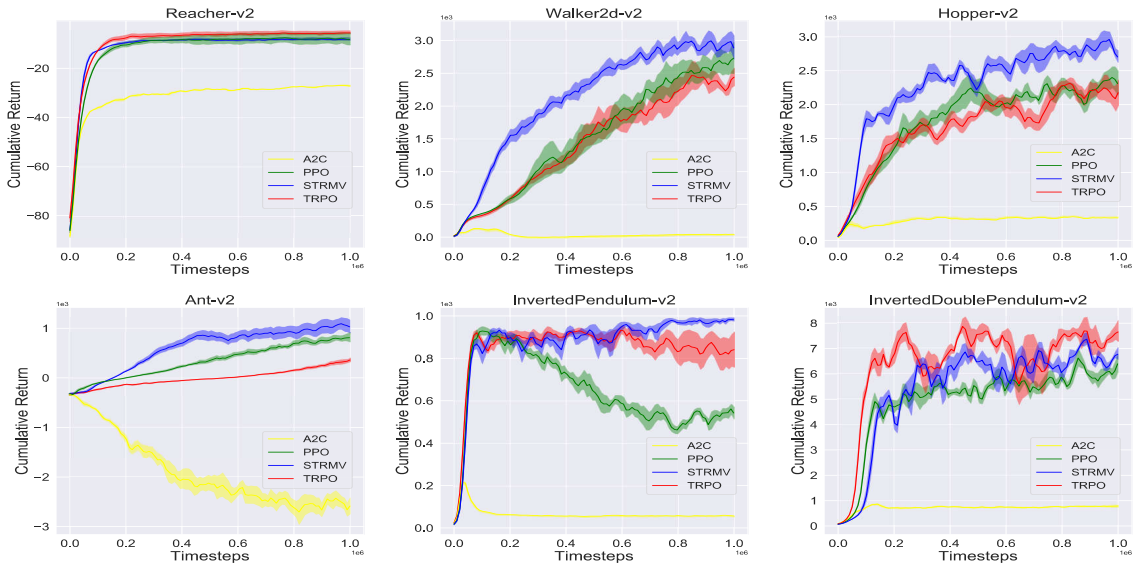

Figure 1: The comparison of A2C,PPO,TRPO,STRVM on several MuJoCo environments within one million time-steps. These curves are generated by averaged 10 random seeds and the shadow areas means one standard deviation. The benchmark algorithms are used with default hyper-parameters.

objective can be modified as

$$L(\theta) = \mathbb{E}_{(s,a)\sim\rho_{\theta_{old}}}\Big[\min\Big(f_t(\theta)\Big(\hat{A}_t^T + \alpha v_{t+1}(\theta)\Big)\Big), \\ e^{s_t^{clip}(\theta)}\Big(\hat{A}_t^T + \alpha\cdot clip\Big(v_{t+1}(\theta), \frac{-\eta}{\alpha+\ell'}, \frac{\eta}{\alpha+\ell}\Big)\Big)\Big] \tag{26}$$

where $s_t^{clip}(\theta) = clip(\psi_t(\theta), -\epsilon, +\epsilon) - clip(v_t(\theta), \frac{-\eta}{\alpha+\ell'}, \frac{\eta}{\alpha+\ell'})$ constructs the trust regions of mean and variance separately in the probability ratio f.Thus we call separated trust region for policy mean and variance (STRMV). Besides, a proper positive value ℓ' is set for the bound constraint that the probability ratio r imposes on the policy variance, i.e. $|v_t(\theta)| \leq \frac{-\eta}{\ell'} = \lim_{\alpha\to 0}\frac{-\eta}{\alpha+\ell'}$.

In particular, we store the recent experiences at each time-step in the form $e_t = (s_t, a_t, r_t, \sigma_t, s_{t+1})$, and both reward and entropy appproximators are updated applying truncated version of generalized advantage estimator (GAE) [19] within time-steps T. Also, fixed-length trajectory segments are used in this work as well as PPO [20].

4 EXPERIMENT

In this section, we desire to answer three questions: (a) how much improvement in sample efficiency and cumulative reward have been achieved by STRMV compared to other state of art on-policy baselines. (b) How to keep a balance between exploration and exploitation with a correct temperature parameter α. (c) Which component plays a core role in STRMV.

We conducted a series of experiments on a variety of challenging continuous control benchmark environments in OpenAI Gym benchmark suites [1], using the MuJoCo engine [21]. As shown in figure 1, we evaluate our algorithms on six continuous control tasks, namely Reacher, Walker2d, Ant, Hopper, InvertedPendulum and InvertedDoublePendulum. In those tasks, the agents are required to control the robots' joint torques by observing their position information and kinematic parameters (please refer to [1] for more details).

In all comparative experiments, the same input pre-processing and model architecture are adopted as [12]. The state value networks of reward V_π and entropy $\mathbb{H}_\pi$ have the same structure but share no parameters. We anneal the temperature parameter α over the course of training together with learning rate. To obtain a stable training process, we use the neural networks only to learn the policy mean rather than output both mean and variance while policy variance is simply set to be a learnable variable shared by all states, which has been commonly adopted by most actor-critic algorithms [15, 18, 20]. Besides, the same hyperparameters for all MuJoCo environments are used without further fine-tuning and provided in Appendix E.

4.1 Comparative Evaluation

We compare our method to proximal policy optimization (PPO) [20], trust region policy optimization (TRPO) [18] and advantage actor critic (A2C) [15]. These three are all the best-known on-policy algorithms in the continuous domains of reinforcement learning.

Figure 1 shows the cumulative reward during 1 million time-steps training over 10 different random seeds respectively for STRMV(blue), PPO(green), TRPO(red) and A2C(yellow). The solid curves report the average score and the shaded regions are formed by the minimum and maximum return bound at each step over the 10 trials.

The results show that, A2C can hardly make any progress almost in all harder tasks within 1 million time-steps, which has revealed its sample-inefficiency. On the contrary, STRMV performs quite

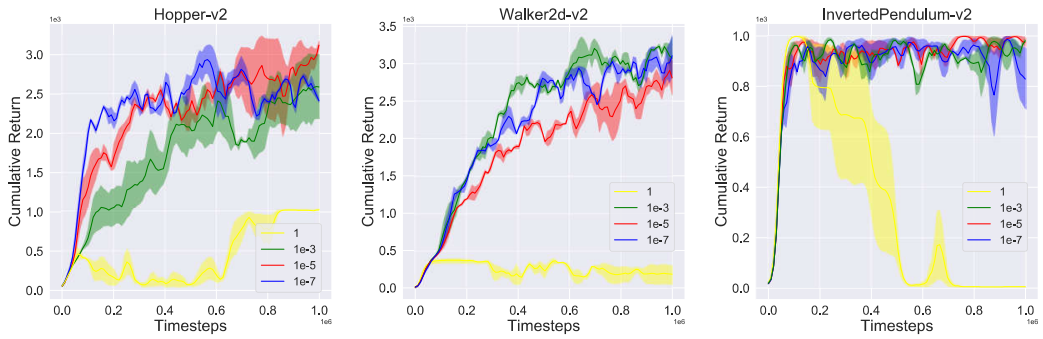

Figure 2: Sensitivity analysis of STRVM to temperature parameter α on some MuJoCo environments.

robust across various environments and learns faster than PPO and TRPO with higher or comparable reward scores on all tasks, except for InvertedDoublePendulum, where STRMV performs better than PPO but slightly worse than TRPO. For the easier task Reacher, all algorithms achieve the similar learning speed and reward score except A2C, but the standard deviation of STRMV is apparently smaller than TRPO and PPO, which means that STRMV gets more stable across different random seeds. More generally, our method gains nearly twice the learning speed of TRPO and PPO in the early episodes on Hopper, Ant, Walker2d and obtains highest final scores on 4 tasks out of 6. The A2C, TRPO and PPO MuJoCo baseline results are provided by the OpenAI team [2] [1]. And the source code of our STRMV implementation and pretrained model weights are available online[2].

4.2 Temperature Parameter Analysis

The temperature parameter α, which determines the relative importance of the entropy term against the reward, has a vital role in the balance between exploration and exploitation. However, when the entropy advantage A_t^H and the relative state policy entropy $v_{t+1}(\theta)$ share the same coefficient α, the actor network could be brittle and easy to get divergent in the early episodes. The direct cause lies in the fact that the values of A_t^H and $v_{t+1}(\theta)$ differ by orders of magnitude, even A_t^H is normalized in a random sampled mini-batch. When α (initial value) is too small, entropy may cast no effect, followed by insufficient explorations. Otherwise, the relative entropy term would be so aggressive that overbold exploration is encouraged, resulting in the network divergence.

Therefore, different weight coefficients are required for A_t^H and $v_{t+1}(\theta)$, expressed as α_H, α_v. Here, we simply set $\alpha_H = 1$, and the reward score during training for several different values of weight ratio $\frac{\alpha_v}{\alpha_H}$ are presented in Figure 2.

We design the hyper-parameter analysis experiments on Hopper, Walker2d and IntertedPendulum. The results show that the best weight ratio differs from task to task, which are mainly determined by the amplitude of entropy advantage function. Yet good performance is achieved in most tasks when weight ratio falls in [1e-5,1e-7]. Besides, apparently the agent can't learn nothing when a shared coefficient α is used, i.e. $\frac{\alpha_v}{\alpha_H} = 1$.

[1]https://github.com/openai/baselines-results
[2]https://github.com/AndyWolfZwei/STRMV

4.3 Component Analysis

As mentioned before, the entropy advantage A_t^H guides the agent to explore on the trajectory with high cumulative entropy, and the relative entropy $v_t(\theta)$ leads to explore more randomly. Besides, separated trust regions are also built for a more stable training process. To figure out which component matters most in STRMV, we set up a group of control trials. Each trial compares STRMV with and without one component. For the component analysis of the entropy advantage and the relative entropy, we can simply set their weight coefficient α_H and α_v to zero respectively. As for separated trust regions, we adopt the original clipped surrogate objective $C(\theta)$ in equation(24). The details of performance degradation on Ant, Hopper and Walker2d are shown in figure 3.

Apparently, the performance of STRMV without separated trust regions is greatly degraded, while STRMV without entropy advantage and relative entropy perform much better. In fact, this work is the first to consider separating the constraints on policy mean and variance. Correspondingly, the results have also demonstrated that separated trust regions shed more light on our algorithm. Furthermore, compared to the relative entropy, the entropy advantage function is more important since it considers the potential entropy over those trajectories starting from a given state. This result can also confirm that a multistep entropy regularizer is more promising than an one-step one.

5 CONCLUSION

We present a simple and computationally inexpensive on-policy maximum entropy algorithm that not only possesses the robust and reliable property of trust region methods but also retains the stable and sample-efficient merit of entropy maximization framework. Our theoretical work derives a more general but conditional bound for policy performance improvement, and obtains a softer objective with stricter trust region that incorporates second-order information. By revealing the essence of state policy entropy, we build separated trust regions for policy mean and variance for further optimization. A series of experiments above has empirically demonstrated that our algorithm is useful and robust across several continuous control tasks and apparently outperforms the most popular on-policy methods on most MuJoCo environments in both sample efficiency and final reward score. This suggests that how to achieve the balance between exploration and exploitation is still

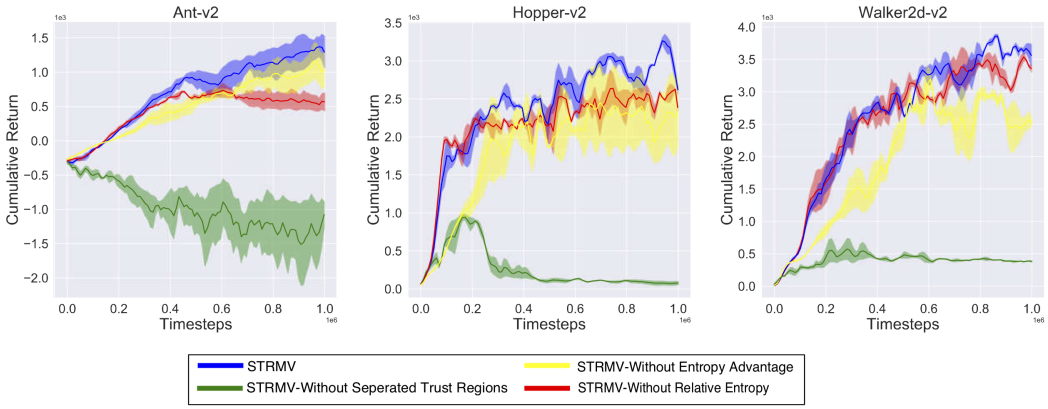

Figure 3: Comparison of STRVM algorithms across different tricks on several MuJoCo environments.

the crucial problem in the continuous domains of reinforcement learning, which has been actually overlooked in most standard policy gradient methods. Instead of simply using one-step entropy as a regularizer, multistep entropy return is a promising research avenue for future work.

6 ACKNOWLEDGMENTS

This paper is partly supported by the National Natural Science Foundation of China (61627810, U1509211) and the National Key R&D Program of China (SQ2017YFGH001005).

7 APPENDIX

A. proof for the differential entropy of Gaussian distribution.

Assume that the action at a given state s_t over policy π subjects to a Gaussian distribution, $\pi(\cdot|s_t) \sim \mathcal{N}(\mu, \sigma^2)$. Hence, we have $\pi(a_t|s_t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(a_t-\mu)^2}{2\sigma^2}}$. Then the state policy entropy can be defined as

$$\begin{aligned}
&E(s_t,\pi) = -\int_{-\infty}^{\infty} \pi(a|s_t)\log\pi(a|s_t)\,da \\
&= -\int_{-\infty}^{\infty} \pi(a|s_t)\log(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(a-\mu)^2}{2\sigma^2}})da \\
&= -\int_{-\infty}^{\infty} \pi(a|s_t)\log\frac{1}{\sqrt{2\pi\sigma^2}}da - \log(e)\int_{-\infty}^{\infty}\pi(a|s_t)\left(-\frac{(a-\mu)^2}{2\sigma^2}\right)da \\
&= \frac{1}{2}\log\left(2\pi\sigma^2\right) + \log(e)\,\frac{\sigma^2}{2\sigma^2} \\
&= \frac{1}{2}\log\left(2\pi e\sigma^2\right)
\end{aligned} \tag{27}$$

B. proof of performance improvement.

This proof refers to the some importance techniques from the proof of Theorem 4.1 in [29]. We extend this lower bound theorem to the maximum entropy framework and a more general result has been obtained. Besides, in this proof, we will demonstrate why TRPO has chosen the KL divergence to replace β (a distance measure between the old policy π and a new one $\tilde{\pi}$) from another perspective. This proof mainly depends on the notion of coupling, as demonstrated in TRPO[12].

$F_\pi(\tilde{\pi})$ can be seen as a result when the action a_t at time step t selected by policy $\tilde{\pi}$ disagrees with π for the first time over the whole episode. Therefore, the error between $F_\pi(\tilde{\pi})$ and $J_T(\tilde{\pi})$ is induced by those other circumstances where the first time that $\tilde{\pi}$ disagrees with π happens before time step t.

For a complete explanation, we begin the proof with a lemma, which describes the difference of soft policy performance $J_T(\tilde{\pi}) - J_T(\pi)$ in terms of the advantage functions and relative state policy entropy.

Lemma 1: *Given two policies $\tilde{\pi}$ and π,*

$$\begin{aligned}
J_T(\tilde{\pi}) - J_T(\pi) = E_{\tau\sim\tilde{\pi}}\Bigg[\sum_{t=0}^{\infty}\gamma^t\Big(A_\pi^V(s_t,a_t) \\
+\alpha\left(A_\pi^H(s_t,a_t) + E(s_t,\tilde{\pi}) - E(s_t,\tilde{\pi})\right)\Big)\Bigg]
\end{aligned} \tag{28}$$

Proof. Here $A_\pi^V(s_t,a_t) = r(s_t,a_t) + \gamma V(s_{t+1}) - V(s_t)$, and $A_\pi^H(s_t,a_t) = E(s_t,\pi) + \gamma\mathbb{H}_\pi(s_{t+1}) - \mathbb{H}_\pi(s_t)$, thus we have

$$\begin{aligned}
&\mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(A_\pi^V(s_t,a_t) + \alpha\left(A_\pi^H(s_t,a_t) + E(s_t,\tilde{\pi}) - E(s_t,\pi)\right)\right)\right] \\
&= \mathbb{E}_{\tau\sim\tilde{\pi}}\left[-V_\pi(s_0) - H_\pi(s_t) + \sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t) + \alpha E(s_t,\tilde{\pi})\right)\right] \\
&= \mathbb{E}_{\tau\sim\tilde{\pi}}\left[-T_\pi(s_0)\right] + J_T(\tilde{\pi}) \\
&= J_T(\tilde{\pi}) - J_T(\pi)
\end{aligned} \tag{29}$$

Let $\rho_\pi(s)$ denotes the discounted visitation frequencies, then the equation can be reorganized as:

$$\begin{aligned}
&J_T(\tilde{\pi}) - J_T(\pi) \\
&= \sum_{t=0}^{\infty}\sum_{s}P(s_t = s|\tilde{\pi})\left[\alpha\gamma^t\log\frac{\sigma(s,\tilde{\pi})}{\sigma(s,\pi)} + \sum_a\tilde{\pi}(a|s)\gamma^t\left(A_\pi^V + \alpha A_\pi^H\right)\right] \\
&= \sum_s\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\tilde{\pi})\left[\alpha\log\frac{\sigma(s,\tilde{\pi})}{\sigma(s,\pi)} + \sum_a\tilde{\pi}(a|s)\left(A_\pi^V + \alpha A_\pi^H\right)\right] \\
&= \sum_s\rho_{\tilde{\pi}}(s)\left[\alpha\log\frac{\sigma(s,\tilde{\pi})}{\sigma(s,\pi)} + \sum_a\tilde{\pi}(a|s)\left(A_\pi^V + \alpha A_\pi^H\right)\right]
\end{aligned} \tag{30}$$

Actually, the selected actions have no effect on current state policy entropy but determine the expectation of the next state policy entropy. Besides, the policy entropy of the next state is already

available to the agent after taking an action, so the equation above is also equivalent to the following form:

$$\begin{aligned} &J_T(\tilde{\pi}) - J_T(\pi) \\ &= \alpha \log \frac{\sigma(s_0,\tilde{\pi})}{\sigma(s_0,\pi)} + \sum_s \rho_{\tilde{\pi}}(s) \sum_a \tilde{\pi}(a|s) \left[A_\pi^V + \alpha \left(A_\pi^H + \log \frac{\sigma(s',\tilde{\pi})}{\sigma(s',\pi)} \right) \right] \end{aligned} \tag{31}$$

where s' denotes the next state and s_0 is the initial state.

C. proof of performance bound.

Theory 1: *Let* $\delta = \max_s |A_\pi^V(s_t,a_t) + \alpha A_\pi^H(s_t,a_t)|$. *For the conservative update rule, the following conditional bound holds: if*

$$\max_s |\log \frac{\sigma(s,\pi_\theta)}{\sigma(s,\pi_{\theta_{old}})}| \le \zeta \tag{32}$$

holds with the new policy $\tilde{\pi}$, then

$$J_T(\tilde{\pi}) \ge F_\pi(\tilde{\pi}) - \frac{2\beta\gamma(\beta\delta+\zeta)}{(1-\gamma)^2} \tag{33}$$

proof : To make the proof Brief, we define

$$\bar{T}(s) = \sum_a \tilde{\pi}(a|s) \left[A_\pi^V + \alpha \left(A_\pi^H + \log \frac{\sigma(s',\tilde{\pi})}{\sigma(s',\pi)} \right) \right] \tag{34}$$

then the soft performance can be

$$J_T(\tilde{\pi}) - J_T(\pi) = \alpha \log \frac{\sigma(s_0,\tilde{\pi})}{\sigma(s_0,\pi)} + \mathbb{E}_{\tau\sim\tilde{\pi}}[\sum_{t=0}^{\infty} \gamma^t \bar{T}(s_t)] \tag{35}$$

and correspondingly

$$F_\pi(\tilde{\pi}) - J_T(\pi) = \alpha \log \frac{\sigma(s_0,\tilde{\pi})}{\sigma(s_0,\pi)} + \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t \bar{T}(s_t)] \tag{36}$$

Note that $F_\pi(\tilde{\pi})$ collects $\bar{T}(s)$ over the old policy π. To bound the gap between $F_\pi(\tilde{\pi})$ and $J_T(\tilde{\pi})$, we divide all sample trajectories into two cases: $\tilde{\pi}$ totally agrees with π and otherwise. For each time step, the probability that the actions sampled using π and $\tilde{\pi}$ are the same is β. we use c_t to count the number of times π and $\tilde{\pi}$ disagree with each other before time t. Then, $P(c_t = 0) = (1-\beta)^t$ while $P(c_t > 0) = 1 - (1-\beta)^t$. Let's first consider the difference contributed by the trajectories with time-step t,

$$\begin{aligned} &\left| \mathbb{E}_{\tau\sim\tilde{\pi}} \left[\bar{T}(s_t) \right] - \mathbb{E}_{\tau\sim\pi} \left[\bar{T}(s_t) \right] \right| \\ &= P(c_t > 0) \left| \mathbb{E}_{\tau\sim\tilde{\pi}|c_t>0} \left[\bar{T}(s_t) \right] - \mathbb{E}_{\tau\sim\pi|c_t>0} \left[\bar{T}(s_t) \right] \right| \\ &\le \left(1-(1-\beta)^t\right) \left(\left| \mathbb{E}_{\tau\sim\tilde{\pi}|c_t>0} \left[\bar{T}(s_t) \right] \right| + \left| \mathbb{E}_{\tau\sim\pi|c_t>0} [\bar{T}(s_t)] \right| \right) \\ &\le 2\left(1-(1-\beta)^t\right) \max \left| \bar{T}(s_t) \right| \end{aligned} \tag{37}$$

Here, we naturally know that $\mathbb{E}_{\tau\sim\tilde{\pi}|c_t=0}[\bar{T}(s_t)] = \mathbb{E}_{\tau\sim\pi|c_t=0}[\bar{T}(s_t)]$ when π and $\tilde{\pi}$ totally agree with each other before time step t, thus the difference comes from the cases where a time $\tilde{a}_i \neq a_i$ happens for $i < t$ with probability $P(c_t > 0) = 1-(1-\beta)^t$. Next, we evaluate the difference between $F_\pi(\tilde{\pi})$ and $J_T(\tilde{\pi})$ over trajectories of all time steps and we get

$$\begin{aligned} &|J_T(\tilde{\pi}) - F_\pi(\tilde{\pi})| \\ &= \left| \mathbb{E}_{\tau\sim\tilde{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \bar{T}(s_t) \right] - \mathbb{E}_{\tau\sim\pi} \left[\sum_{t=0}^{\infty} \gamma^t \bar{T}(s_t) \right] \right| \\ &= \sum_{t=0}^{\infty} \gamma^t \left| \mathbb{E}_{\tau(t)\sim\tilde{\pi}} \left[\bar{T}(s_t) \right] - \mathbb{E}_{\tau(t)\sim\pi} \left[\bar{T}(s_t) \right] \right| \\ &\le 2 \sum_{t=0}^{\infty} \gamma^t \left(1-(1-\beta)^t\right) \max \left| \bar{T}(s_t) \right| \end{aligned} \tag{38}$$

The rest of the work is how to bound $\bar{T}(s)$ at different time-step t. Here, $\bar{T}(s)$ can be divided into two parts $\varphi_{\tilde{\pi}}(s)$ and $\alpha e_{\tilde{\pi}}(s)$, where

$$\begin{aligned} \varphi_{\tilde{\pi}}(s) &= \sum_a \tilde{\pi}(a|s) \left(A_\pi^V + \alpha A_\pi^H \right) \\ e_{\tilde{\pi}}(s) &= \sum_a \tilde{\pi}(a|s) \log \frac{\sigma(s',\tilde{\pi})}{\sigma(s',\pi)} \end{aligned} \tag{39}$$

and we separately find out their upper bound. Let's focus on $\varphi(\tilde{\pi})$ first, and it's easy to get

$$\begin{aligned} &|\varphi_{\tilde{\pi}}(s)| \\ &= \left| \sum_a \pi_\theta(a|s) \left(A_{\theta_{old}}^V + \alpha A_{\theta_{old}}^H \right) \right| \\ &= \sum_a \left((1-\beta)\pi_{\theta_{old}}(a|s) + \beta\pi_{\theta'}(a|s) \right) \left| A_{\theta_{old}}^V + \alpha A_{\theta_{old}}^H \right| \\ &= \beta \sum_a \pi_{\theta'}(a|s) \left| A_{\theta_{old}}^V + \alpha A_{\theta_{old}}^H \right| \\ &\le \beta \max \left| A_{\theta_{old}}^V + \alpha A_{\theta_{old}}^H \right| \cdot \sum_a \pi_{\theta'}(a|s) \\ &\le \beta\delta \end{aligned} \tag{40}$$

The above result can be easily obtained because

$$\sum_a \pi_{\theta_{old}}(a|s) \left(A_{\theta_{old}}^V + \alpha A_{\theta_{old}}^H \right) = 0 \tag{41}$$

As for $e_{\tilde{\pi}}(s)$, we assume that the old policy subjects to $\pi_{\theta_{old}}(\cdot|s_t) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and the expected update policy $\pi_{\theta'}(\cdot|s_t) \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then the action at state $_t$ selected by the new policy $\pi_\theta(\cdot|s)$ has a distribution of $\mathcal{N}((1-\beta)\mu_1 + \beta\mu_2, (1-\beta)^2\sigma_1^2 + \beta^2\sigma_2^2)$. The relative entropy will be

$$\begin{aligned} \log \frac{\sigma(s,\pi_\theta)}{\sigma\left(s,\pi_{\theta_{old}}\right)} &= \log \frac{(1-\beta)^2\sigma_1^2 + \beta^2\sigma_2^2}{\sigma_1^2} \\ &= \log \left((1-\beta)^2 + \beta^2 \frac{\sigma_2^2}{\sigma_1^2} \right) \end{aligned} \tag{42}$$

Here, we make a constraint on the output variance of policy $\pi_{\theta'}(\cdot|s)$ for all states, i.e. $\max_s |\log \frac{\sigma_2}{\sigma_1}| < \omega$, and correspondingly let $k = \arg\max_k \{|\log(1-\beta)^2 + \beta^2 k)|\}, k \in \{e^\omega, e^{-\omega}\}$. since the new policy $\pi_\theta(\cdot|s)$ at state s will not have an impact on the relative policy entropy of next state s', we can simply get $|e_{\tilde{\pi}}(s)| \le \max |\log \frac{\sigma(s',\pi_\theta)}{\sigma(s',\pi_{\theta_{old}})}| \cdot \sum_a \pi_\theta(a|s) \le |\log(1-\beta)^2 + \beta^2 k)|$. Thus, $|\bar{T}(s_t)| \le |\varphi_{\tilde{\pi}}(s)| + \alpha|e_{\tilde{\pi}}(s)| \le \beta\delta + \alpha|\log(1-\beta)^2 + \beta^2 k)|$ holds for each state. Finally, we combine those inequalities together and get

$$\begin{aligned} &|J_T(\tilde{\pi}) - F_\pi(\tilde{\pi})| \\ &\le 2 \sum_{t=0}^{\infty} \gamma^t \left(1-(1-\beta)^t\right) \max \left| \bar{T}(s_t) \right| \\ &\le 2 \left(\beta\delta + \alpha \left| \log(1-\beta)^2 + \beta^2 k \right| \right) \sum_{t=0}^{\infty} \gamma^t \left(1-(1-\beta)^t\right) \\ &= 2 \left(\beta\delta + \alpha \left| \log(1-\beta)^2 + \beta^2 k \right| \right) \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\beta)} \right) \\ &= \frac{2\beta\gamma \left(\beta\delta + \alpha \left| \log(1-\beta)^2 + \beta^2 k \right| \right)}{(1-\gamma)(1-\gamma(1-\beta))} \\ &\le \frac{2\beta\gamma \left(\beta\delta + \alpha \left| \log(1-\beta)^2 + \beta^2 k \right| \right)}{(1-\gamma)^2} \end{aligned} \tag{43}$$

Thus, we have

$$J_T(\tilde{\pi}) \ge F_\pi(\tilde{\pi}) - \frac{2\beta\gamma \left(\beta\delta + \alpha \left| \log\left((1-\beta)^2 + \beta^2 k \right) \right| \right)}{(1-\gamma)^2} \tag{44}$$

D. KL-divergence between old and new policy.

It's easy to know that KL-divergence between two Gaussian distributions is

$$D_{KL}(p||q) = \log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2} \tag{45}$$

where $p \sim \mathcal{N}(\mu_p, \sigma_p^2)$ and $q \sim \mathcal{N}(\mu_q, \sigma_q^2)$. Thus, substitute the distribution of old and new policy into the above equation and we can get

$$D_{KL}(\pi(\cdot|s)||\tilde{\pi}(\cdot|s)) = \frac{1}{2}\log\left((1-\beta)^2 + \beta^2\chi\right) + \frac{1+\beta^2 z}{2((1-\beta)^2+\beta^2\chi)} - \frac{1}{2}$$

where $\chi = \frac{\sigma_2^2}{\sigma_1^2}$ and $z = \left(\frac{\mu_1-\mu_2}{\sigma_1}\right)^2$.

Here, χ and z can be seen as two constant. Since $0 \le \beta \le 1$ and $\chi > 0$, we have $2((1-\beta)^2 + \beta^2\chi) \le 2(1+\chi)$. Then

$$D_{KL}(\pi(\cdot|s)||\tilde{\pi}(\cdot|s)) \ge \frac{1}{2}\log\left((1-\beta)^2 + \beta^2\chi\right) + m\beta^2 - \frac{1}{2}$$

where $m = \frac{(\mu_1-\mu_2)^2}{2(1+\chi)\sigma_1^2} = \frac{(\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}$.

E. Hyperparameters

The following table lists the common parameters over all the on-policy algorithms and unique parameters of STRMV used in the comparative evaluation in Figure 1.

Common Hyper-parameters	
Parameter	Value
learning rate	3e-4(Linear decay)
Discount(γ)	0.99
GAE(λ)	0.95
hidden layer number	2
hidden units per layer	64
Mini-batch size	64
optimizer	Adam
Value loss coefficient	0.5
STRMV Hyperparameters	
Entropy loss coefficient	1
α for clipped margin	6e-3(Linear decay)
α_H	1(Linear decay)
α_v	1e-6(Linear decay)
ϵ	0.2
η	3e-4
ℓ	1e-8
ℓ'	1e-3

REFERENCES

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

[2] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. https://github.com/openai/baselines, 2017.

[3] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *arXiv preprint arXiv:1512.08562*, 2015.

[4] Audrunas Gruslys, Mohammad Gheshlaghi Azar, Marc G Bellemare, and Remi Munos. The reactor: A sample-efficient actor-critic architecture. *arXiv preprint arXiv:1704.04651*, 2017.

[5] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *arXiv preprint arXiv:1702.08165*, 2017.

[6] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

[7] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *ICML*, volume 2, pages 267–274, 2002.

[8] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[9] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[10] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[11] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[13] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2775–2785, 2017.

[14] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Trust-pcl: An off-policy trust region method for continuous control. *arXiv preprint arXiv:1707.01891*, 2017.

[15] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

[16] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[17] Nicol N Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural computation*, 14(7):1723–1738, 2002.

[18] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[19] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

[21] D Silver. D. silver, a. huang, cj maddison, a. guez, l. sifre, g. van den driessche, j. schrittwieser, i. antonoglou, v. panneershelvam, m. lanctot, s. dieleman, d. grewe, j. nham, n. kalchbrenner, i. sutskever, t. lillicrap, m. leach, k. kavukcuoglu, t. graepel, and d. hassabis, nature (london) 529, 484 (2016). *Nature (London)*, 529:484, 2016.

[22] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[23] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *ICML*, 2014.

[24] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[25] Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pages 1049–1056. ACM, 2009.

[26] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, volume 2, page 5. Phoenix, AZ, 2016.

[27] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015.

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

[29] Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010.

[30] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.