Trust Region Policy Optimization

Hong Yige
116010071@link.cuhk.edu.cn

Junior, SSE, CUHKSZ

March 4, 2019

Introduction

- **Generalized Policy Iteration**
  Policy Evaluation: Estimate V_π
  Any Policy Evaluation Algorithm
  Policy Improvement: Generate $\pi' \geq \pi$
  Any policy Improvement Algorithm
- Policy iteration methods(value based) and policy gradient methods(policy based) can be viewed under this framework.
- Due to inaccurate estimate of V_π, new policy doesn't necessarily improve. Due to this, many algorithms based on generalized policy iteration have unsatisfactory result on some problems. For example, previously on Teris or locomotion, policy iteration or policy gradient cannot beat gradient-free methods like cross-entropy method(CME) and covariance matrix adaptation.
- TRPO makes several approximations to a procedure with guaranteed monotonic improvement.
- Effective for optimizing large nonlinear policies.

Overview

1. Problem Setup and Notations
2. Concepts and Theorems
3. Prototype Algorithm with Guaranteed Monotonic Improvement
4. Trust Region Policy Optimization
5. Proximal Policy Optimization

Problem Setup and Notations

- MDP is defined as $(S, A, P, r, \rho_0, \gamma)$. S state space, A action space,$P : S \times A \times S \to \mathbb{R}$ transition probability, $r : S \times A \to \mathbb{R}$ reward function, $\rho_0 : S \to \mathbb{R}$ distribution of initial state, $\gamma \in (0, 1)$ discount factor.
- Stochastic policy $\pi : S \times A \to \mathbb{R}$
- Value function, state-action value and advantage:

$$Q_\pi(s, a) = \mathbb{E}_{s_0, a_0, s_1, a_1 \ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a] \tag{1}$$

$$V_\pi(s) = \mathbb{E}_{s_0, a_0, s_1, a_1 \ldots}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s] \tag{2}$$

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s) \tag{3}$$

where $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$, $a_t \sim \pi(a_t|s_t)$ where $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$, $a_t \sim \pi(a_t|s_t)$

Problem Setup and Notations

- Discounted visitation frequency

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s)... \tag{4}$$

where $s_0, a_0, s_1, a_1...$ is a sequence sampled according to policy π

- A measure of policy performance: long run expected reward

$$\eta(\pi) = \mathbb{E}_{s_0, a_0, s_1, a_1...}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)] = \mathbb{E}_{s_0 \sim \rho_0}[V_\pi(s_0)] \tag{5}$$

where $s_0 \sim \rho_0$, $a_t \sim \pi(a_t|s_t)$, $s_{t+1} \sim P(s_{t+1}|s_t, a_t)$

Measuring Degree of Policy Improvement

Lemma1

$$\eta(\widetilde{\pi}) - \eta(\pi) = \mathbb{E}_{s_0, a_0, \ldots \sim \widetilde{\pi}}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)] \tag{6}$$

We can rewrite it in terms of state-distribution:

$$\eta(\widetilde{\pi}) - \eta(\pi) = \mathbb{E}_{s_0, a_0, \ldots \sim \widetilde{\pi}}[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)] \tag{7}$$

$$= \sum_{t=0}^{\infty} \sum_{s \in S} P(s_t = s|\widetilde{\pi}) \sum_{a \in A} \widetilde{\pi}(a|s) \gamma^t A_\pi(s, a) \tag{8}$$

$$= \sum_{s \in S} \sum_{t=0}^{\infty} P(s_t = s|\widetilde{\pi}) \sum_{a \in A} \widetilde{\pi}(a|s) \gamma^t A_\pi(s, a) \tag{9}$$

$$= \sum_{s \in S} \rho_{\widetilde{\pi}}(s) \sum_{a \in A} \widetilde{\pi}(a|s) A_\pi(s, a) \tag{10}$$

Proof of Lemma1

Proof.

$A_\pi(s,a) = \mathbb{E}_{s'\sim P(s'|s,a)}[r(s,a) + \gamma V_\pi(s') - V_\pi(s)]$, therefore

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

$$=\mathbb{E}_{\tau|\widetilde{\pi}}[\sum_{t=0}^{\infty}\gamma^t (r(s_t,a_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t))] \tag{12}$$

$$=\mathbb{E}_{\tau|\widetilde{\pi}}[-V_\pi(s_0) + \sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)] \tag{13}$$

$$= -\mathbb{E}_{s_0\sim\rho_0}[V_\pi(s_0)] + \mathbb{E}_{\tau|\widetilde{\pi}}[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)] \tag{14}$$

$$= -\eta(\pi) + \eta(\widetilde{\pi}) \tag{15}$$

Approximating Degree of Policy Improvement

Since the dependency of $\rho_{\widetilde{\pi}}(a)$ on $\widetilde{\pi}$ is complicated, and require access to system model, we define a "local approximation" to $\eta(\widetilde{\pi})$, called "policy advantage". It serves as a surrogate function.

$$L_\pi(\widetilde{\pi}) = \eta(\pi) + \sum_{s \in S} \rho_\pi(s) \sum_{a \in A} \widetilde{\pi}(a|s) A_\pi(s,a) \tag{16}$$

$$= \eta_(\pi) + \mathbb{E}_{s_0, a_0, \ldots \sim \pi}[\sum_{t=0}^{\infty} \gamma^t \frac{\widetilde{\pi}(a_t|s_t)}{\pi(a_t|s_t)} A_\pi(s_t, a_t)] \tag{17}$$

$L_\pi(\widetilde{\pi})$ is a first order approximation to $\eta(\widetilde{\pi})$:

$$L_{\pi_{\theta_0}}(\pi_{\theta_0}) = \eta(\pi_0) \tag{18}$$

$$\nabla_\theta L_{\pi_{\theta_0}}(\pi_\theta)\big|_{\theta=\theta_0} = \nabla_\theta \eta(\theta)\big|_{\theta=\theta_0} \tag{19}$$

Lower Bounding Degree of Policy Improvement

- When we use approximated $\hat{A}_\pi$ instead of exact A_π, optimizing $L_\pi(\widetilde{\pi})$ doesn't necessarily give improved $\eta(\widetilde{\pi})$.
- A lower bound of $\eta(\widetilde{\pi})$ in terms of "distance" to the current policy

Lower Bounding Degree of Policy Improvement

Theorem

$$\eta(\widetilde{\pi}) \geq L_{\pi}(\widetilde{\pi}) - \frac{4\epsilon\gamma}{(1-\gamma)^2} D_{KL}^{max}(\pi \Big|\Big| \widetilde{\pi}) \tag{20}$$

$\epsilon = \max_{s,a} \left|A_{\pi}(s,a)\right|$, $D_{KL}^{max}(\pi||\widetilde{\pi}) = \max_s D_{KL}(\pi(\cdot,s)||\widetilde{\pi}(\cdot,s))$

Theorem's Proof

Definition

Two distributions p and q are α-coupled if there is a joint distribution (p, q) with marginal p and q. For $(X, Y) \sim (p, q)$, $P(X \neq Y) \leq \alpha$. Two policies π $\widetilde{\pi}$ are α-coupled if $\forall s$ $\pi(\cdot, s)$ and $\widetilde{\pi}(\cdot, s)$ are α-coupled.

Lemma2

When $D_{KL}(p||q) \leq \alpha^2$, p, q are α-coupled.

Theorem's Proof

Lemma3

$$\left|\bar{A}(s)\right| \leq 2\alpha \max_{s,a} \left|A_\pi(s,a)\right| \tag{21}$$

where $\bar{A}(s) = \sum_{a\in A} \tilde{\pi}(a|s) A_\pi(s,a)$

Proof of Lemma3:
Since $\mathbb{E}_{a\sim\pi}[A_\pi(s,a)] = \mathbb{E}_{a\sim\pi}[Q_\pi(s,a) - V_\pi(s)] = 0$

$$\begin{aligned}
\bar{A}(s) =& \mathbb{E}_{\tilde{a}\sim\tilde{\pi}}[A_\pi(s,\tilde{a})] && (22)\\
=& \mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})}[A_\pi(s,\tilde{a}) - A_\pi(s,a)] && (23)\\
=& P(a\neq\tilde{a})\mathbb{E}_{(a,\tilde{a})\sim(\pi,\tilde{\pi})|a\neq\tilde{a}}[A_\pi(s,\tilde{a}) - A_\pi(s,a)] && (24)\\
\leq& 2\alpha \max_{s,a} \left|A_\pi(s,a)\right| && (25)
\end{aligned}$$

Theorem's Proof

Lemma4

Let π and $\widetilde{\pi}$ be α-coupled. Then

$$\left|\mathbb{E}_{s_t\sim\widetilde{\pi}}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)]\right| \leq 4\alpha(1-(1-\alpha)^t)\max\left|A_\pi(s,a)\right| \quad (26)$$

Proof of Lemma4

Consider the trajectory generated by $\widetilde{\pi}$:$\{s'_0, a'_0, s'_1, a'_1, ...\}$ and π:$\{s_0, a_0, s_1, a_1, ...\}$. Let n_t be the number of times $a'_i \neq a_i$ for $i < t$.

$$\mathbb{E}_{s_t\sim\widetilde{\pi}}[\bar{A}(s_t)] = P(n_t=0)\mathbb{E}_{s_t\sim\widetilde{\pi}|n_t=0}[\bar{A}(s_t)] + P(n_t>0)\mathbb{E}_{s_t\sim\widetilde{\pi}|n_t>0}[\bar{A}(s_t)] \quad (27)$$

$$\mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)] = P(n_t=0)\mathbb{E}_{s_t\sim\pi|n_t=0}[\bar{A}(s_t)] + P(n_t>0)\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)] \quad (28)$$

Theorem's Proof

$$\mathbb{E}_{s_t\sim\widetilde{\pi}|n_t=0}[\bar{A}(s_t)] = \mathbb{E}_{s_t\sim\pi|n_t=0}[\bar{A}(s_t)] \tag{29}$$

Subtracting (28)(29), we have

$$\begin{aligned}
&\mathbb{E}_{s_t\sim\widetilde{\pi}}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi}[\bar{A}(s_t)] && (30)\\
=&(\mathbb{E}_{s_t\sim\widetilde{\pi}|n_t>0}[\bar{A}(s_t)] - \mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)])P(n_t>0) && (31)\\
\leq&\left(\left|\mathbb{E}_{s_t\sim\widetilde{\pi}|n_t>0}[\bar{A}(s_t)]\right| + \left|\mathbb{E}_{s_t\sim\pi|n_t>0}[\bar{A}(s_t)]\right|\right)P(n_t>0) && (32)\\
\leq&4\alpha\max\left|A_\pi(s,a)\right|(1-(1-\alpha)^t) && (33)
\end{aligned}$$

The last inequality is because Lemma3 and that π and $\widetilde{\pi}$ are α-coupled.

Theorem's Proof

Proof of the theorem:

Let $D_{KL}^{max}(\pi||\widetilde{\pi}) = \alpha^2$, by lemma2, π and $\widetilde{\pi}$ are α-coupled.

$$\eta(\widetilde{\pi}) = \eta(\pi) + \sum_{s\in S} \rho_{\widetilde{\pi}}(s) \sum_{a\in A} \widetilde{\pi}(a|s) A_\pi(s,a) \tag{34}$$

$$= \eta(\pi) + \sum_{s\in S} \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\widetilde{\pi}) \sum_{a\in A} \widetilde{\pi}(a|s) A_\pi(s,a) \tag{35}$$

$$= \mathbb{E}_{\tau\sim\widetilde{\pi}}[\sum_{t=0}^{\infty} \gamma^t \bar{A}(s_t)] \tag{36}$$

$$L_\pi(\widetilde{\pi}) = \eta(\pi) + \sum_{s\in S} \rho_\pi(s) \sum_{a\in A} \widetilde{\pi}(a|s) A_\pi(s,a) \tag{37}$$

$$= \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t \bar{A}(s_t)] \tag{38}$$

Theorem's Proof

$$\left|\eta(\widetilde{\pi}) - L_{\pi}(\widetilde{\pi})\right| \leq \sum_{t=0}^{\infty} \gamma^t \left|\mathbb{E}_{\tau\sim\widetilde{\pi}}[\bar{A}(s_t)] - \mathbb{E}_{\tau\sim\pi}[\bar{A}(s_t)]\right| \tag{39}$$

$$\leq \sum_{t=0}^{\infty} \gamma^t 4\alpha\epsilon(1-(1-\alpha)^t) \tag{40}$$

$$= 4\alpha\epsilon\left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)}\right) \tag{41}$$

$$= \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)(1-\gamma(1-\alpha))} \tag{42}$$

$$\leq \frac{4\alpha^2\gamma\epsilon}{(1-\gamma)^2} \tag{43}$$

where $\epsilon = \max\left|A_{\pi}(s,a)\right|$.

Monotonic Policy Improvement

Optimize over the bound will give improved policies:
let $\pi^* = argmax_{\widetilde{\pi}}[L_\pi(\widetilde{\pi}) - CD_{KL}^{max}(\pi||\widetilde{\pi})]$. Then

$$\eta(\pi^*) \geq \max_{\widetilde{\pi}} L_\pi(\widetilde{\pi}) - CD_{KL}^{max}(\pi||\widetilde{\pi}) \geq L_\pi(\pi) - CD_{KL}^{max}(\pi||\pi) = \eta(\pi) \quad (44)$$

This is actually a kind of *minorization-maximization algorithm*:
Objective $\eta(\pi)$, find lower bound $M_\pi(\widetilde{\pi}) \leq \eta(\widetilde{\pi})$ and $M_\pi(\pi) = \eta(\pi)$.
Repeat the following steps:

1. $\pi' \leftarrow argmax M_\pi(\widetilde{\pi})$
2. $\pi \leftarrow \pi'$

Then for generated $\pi_1, \pi_2, ...$, we have

$$\eta(\pi_t) \geq M_{\pi_{t-1}}(\pi_t) \geq M_{\pi_{t-1}}(\pi_{t-1}) \geq \eta(\pi_{t-1}) \quad (45)$$

which improves the objective monotonically.

Monotonic Policy Improvement

Algorithm 1 Policy iteration algorithm guaranteeing non-decreasing expected return η

Initialize π_0.

for $i = 0, 1, 2, \ldots$ until convergence **do**

Compute all advantage values $A_{\pi_i}(s, a)$.

Solve the constrained optimization problem

$$\pi_{i+1} = \arg\max_{\pi} \left[L_{\pi_i}(\pi) - C D_{\mathrm{KL}}^{\max}(\pi_i, \pi) \right]$$

where $C = 4\epsilon\gamma/(1-\gamma)^2$

and $L_{\pi_i}(\pi) = \eta(\pi_i) + \sum_s \rho_{\pi_i}(s) \sum_a \pi(a|s) A_{\pi_i}(s, a)$

end for

Trust Region Policy Optimization

1 Change from penalty to hard-constraint Let current policy parameter be θ, and improved policy parameter is $\widetilde{\theta}$. In prototype algorithm, the policy improvement step solves an optimization problem

$$\max_{\widetilde{\theta}}[L_\theta(\widetilde{\theta}) - CD_{KL}^{max}(\theta||\widetilde{\theta})] \tag{46}$$

However,

- $C = \frac{4\epsilon\gamma}{(1-\gamma)^2}$ is too large, and results in too small step size.
- C as a hyper-parameter hard to adjust

Change to a hard constraint on KL-divergence:

$$\max_{\theta} L_\theta(\widetilde{\theta}) \tag{47}$$

$$\text{subject to } D_{KL}^{max}(\widetilde{\theta}||\widetilde{\theta}) \leq \delta \tag{48}$$

for some $\delta > 0$

Trust Region Policy Optimization

2 Change from max KL divergence to average KL divergence

Computing maximal KL divergence D_{KL}^{max} is also impractical. We use average KL divergence instead:

$$D_{KL}^{\rho_\pi}(\theta||\widetilde{\theta}) = \mathbb{E}_{s\sim\rho_\pi}[\pi_\theta(\cdot,s)||\pi_{\widetilde{\theta}}(\cdot,s)] \tag{49}$$

$$D_{KL}^{\rho_\pi}(\theta||\widetilde{\theta}) \approx D_{KL}^{max}(\theta||\widetilde{\theta}) \tag{50}$$

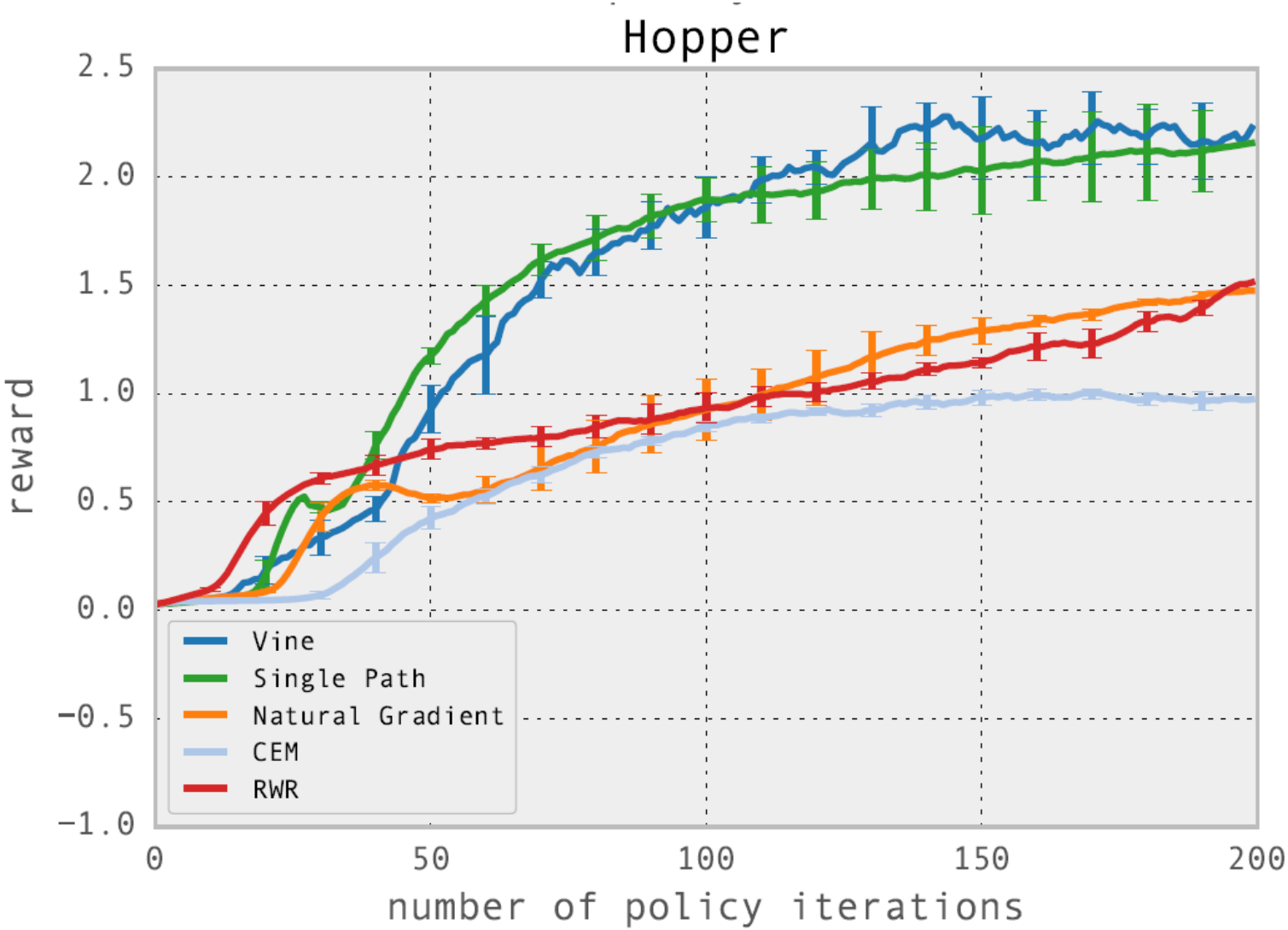
The Trust Region Policy Optimization is finally formulated as

$$\max_{\widetilde{\theta}} L_\theta(\widetilde{\theta}) \tag{51}$$

$$\text{subject to } D_{KL}^{\rho_\pi}(\theta||\widetilde{\theta}) \leq \delta \tag{52}$$

for some fiexd $\delta > 0$

Simulated Robotic Locomotion

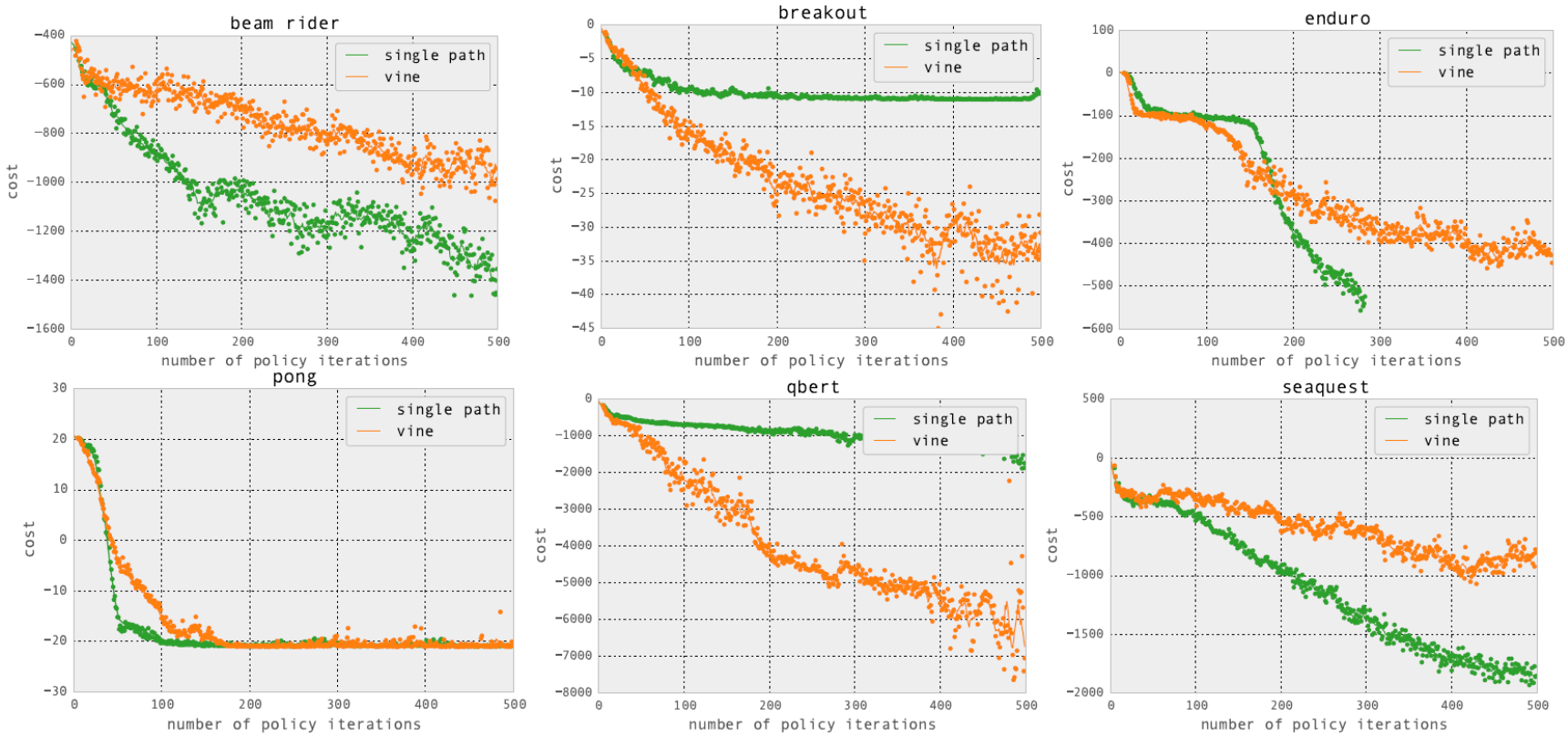


Atari Games

	B. Rider	*Breakout*	*Enduro*	*Pong*	*Q*bert*	*Seaquest*	*S. Invaders*
Random	354	1.2	0	−20.4	157	110	179
Human (Mnih et al., 2013)	7456	31.0	368	−3.0	18900	28010	3690
Deep Q Learning (Mnih et al., 2013)	4092	168.0	470	20.0	1952	1705	581
UCC-I (Guo et al., 2014)	5702	380	741	21	20025	2995	692
TRPO - single path	1425.2	10.8	534.6	20.9	1973.5	1908.6	568.4
TRPO - vine	859.5	34.2	430.8	20.9	7732.5	788.4	450.2

- using same set of parameters; discrete tasks.

Atari Games



- Mostly monotonic
- Sometimes vine is better, sometimes single path is better.

Policy Evaluation Step

- To compute the surrogate objective $L_\theta(\widetilde{\theta})$, we need to evaluate $A_{\pi_\theta}(s,a)$ or $Q_{\pi_\theta}(s,a)$ under current policy π_θ.

$$L_\theta(\widetilde{\theta}) - \eta(\theta) = \sum_{s\in S} \rho_{\pi_\theta}(s) \sum_{a\in A} \pi_{\widetilde{\theta}}(a|s) A_{\pi_\theta}(s,a) \tag{53}$$

$$\text{or } L_\theta(\widetilde{\theta}) = \sum_{s\in S} \rho_{\pi_\theta}(s) \sum_{a\in A} \pi_{\widetilde{\theta}}(a|s) Q_{\pi_\theta}(s,a) \tag{54}$$

Policy Evaluation Step

1 Single Path

1. Sample T of $s_0 \sim \rho_0$. For each s_0, generate a trajectory $\tau : s_0, a_0, r_0, s_1, a_1, r_1 ... s_N, a_N, r_N$ using π_θ.
2. For each trajectory, objective function

$$L_\theta(\widetilde{\theta}) = \mathbb{E}_{s_0, a_0, \ldots \sim \pi}[\sum_{t=0}^{\infty} \gamma^t \frac{\widetilde{\pi}(a_t|s_t)}{\pi(a_t|s_t)} Q_\pi(s_t, a_t)] \tag{55}$$

$$\approx \frac{1}{T} \sum_{\tau} \sum_{t=0}^{N} \gamma^t \frac{\pi_{\widetilde{\theta}}(a_t|s_t)}{\pi_\theta(a_t|s_t)} \hat{Q}(s_t, a_t) \tag{56}$$

where for (s, a) appearing on the trajectories, suppose it appears for m times

$$\hat{Q}(s, a) = \frac{1}{m} \sum_{\tau} \sum_{t: s_t = s, a_t = a} \sum_{i=t}^{N} \gamma^{i-t} r(s_i, a_i) \tag{57}$$

Policy Evaluation Step

2 Vine

Generate T trajectories of length $N+1$ according to ρ_0 and π_θ, collect the states to form a "rollout set" $D = \cup_{j=1}^{T}\{s_{j0}, s_{j1}..s_{jN}\}$. Use rollouts to estimate $\hat{Q}(s,a)$ for some of the $a \in A$.

When action space is small, estimate $\hat{Q}(s,a)$ for all a:

$$L_\theta(\widetilde{\theta}) = \sum_{s\in S}\rho_\pi(s)\sum_{a\in A}\widetilde{\pi}(a|s)Q_\pi(s,a) \approx \frac{1}{T}\sum_{j=1}^{T}\sum_{n=0}^{N}\gamma^n\sum_{a\in A}\pi_{\widetilde{\theta}}(s_{jn},a)\hat{Q}(s_{jn},a) \tag{58}$$

Otherwise sample $\{a_0, a_1, ...a_K\}$with some distribution q, q can be π_θ or uniform distribution.

$$L_\theta(\widetilde{\theta}) \approx \frac{1}{T}\sum_{j=1}^{T}\sum_{n=0}^{N}\gamma^n\frac{\sum_{k=1}^{K}\frac{\pi_{\widetilde{\theta}}(s_{jn},a_k)}{q(s_{jn},a_k)}\hat{Q}(s_{jn},a_k)}{\sum_{k=1}^{K}\frac{\pi_{\widetilde{\theta}}(s_{jn},a_k)}{q(s_{jn},a_k)}} \tag{59}$$

Policy Evaluation Step

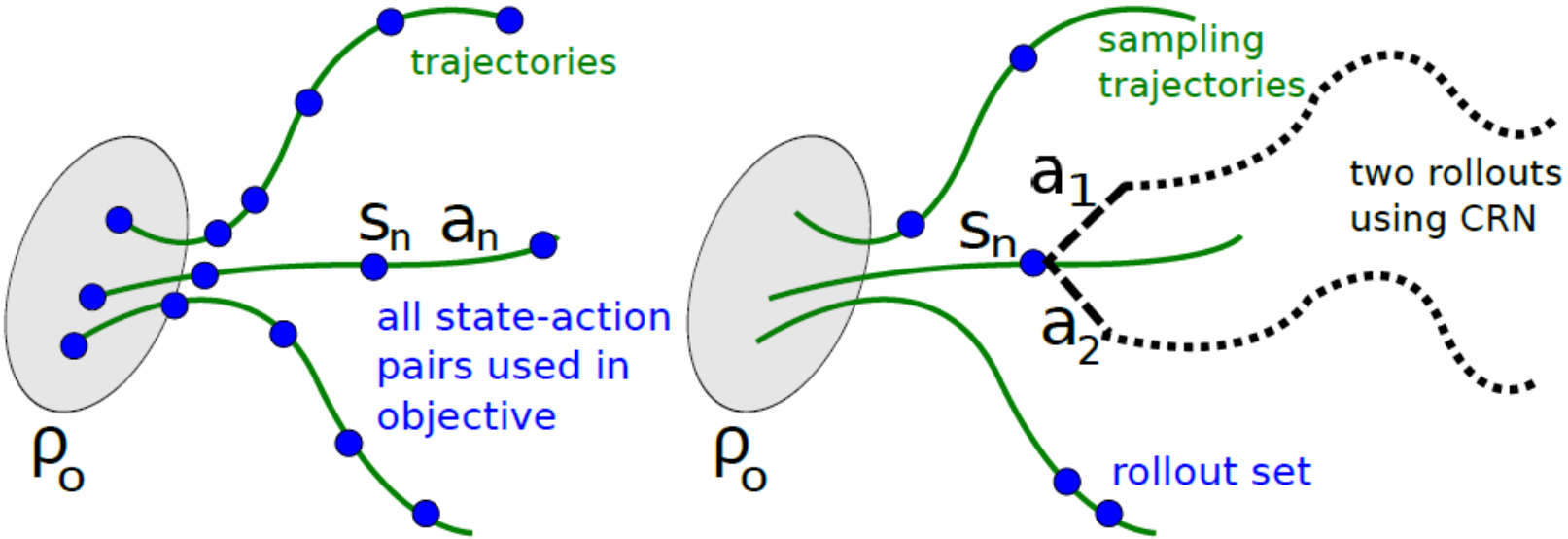


Figure: left: single path; right: vine

Policy Improvement Step

1 Find the update direction and maximal step size from the approximated problem

Perform Taylor expansion. First-order derivative of KL divergence is 0.

$$L_\theta(\widetilde{\theta}) \approx \eta(\theta) + (\widetilde{\theta} - \theta)^T \nabla_{\widetilde{\theta}} L_\theta(\widetilde{\theta})\big|_{\widetilde{\theta}=\theta} \tag{60}$$

$$D_{KL}^{\rho_\pi}(\theta \big|\big| \widetilde{\theta}) \approx (\widetilde{\theta} - \theta)^T A(\theta)(\widetilde{\theta} - \theta) \tag{61}$$

where $A(\theta)_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} D_{KL}^{\rho_\pi}(\theta \big|\big| \widetilde{\theta})\big|_{\widetilde{\theta}=\theta}$

Note that

- $\nabla_{\widetilde{\theta}} L_\theta(\widetilde{\theta})|_{\widetilde{\theta}=\theta} = \nabla_{\widetilde{\theta}} \eta(\widetilde{\theta})\big|_{\widetilde{\theta}=\theta}$, it's just the policy gradient.
- $A(\theta)$ can be shown to be the Fisher Information Matrix.

Policy Improvement Step

- Find direction and maximal step size by solving

$$\max_{\widetilde{\theta}}[(\widetilde{\theta}-\theta)^T \nabla_{\widetilde{\theta}} L_\theta(\widetilde{\theta})] \tag{62}$$

$$\text{subject to } \frac{1}{2}(\widetilde{\theta}-\theta)^T A(\theta)(\widetilde{\theta}-\theta) \leq \delta \tag{63}$$

- It solves $\widetilde{\theta} = \theta + \beta s$, where direction $s = A(\theta)^{-1}\nabla_{\widetilde{\theta}} L_\theta(\widetilde{\theta})$, step size $\beta = \sqrt{2\delta / s^T A(\theta) s}$.
- s is solved using a "Conjugate Gradient Algorithm" without forming and invert the whole Fisher Information Matrix $A(\theta)$.

Policy Improvement Step

2 Determine step size

- Perform a line search in direction s, starting from max step size β, until the objective (25) $L_\theta(\widetilde{\theta})$ improves.

Pseudo Code

For each iteration:

1. Solve $s = A(\theta)^{-1}\nabla_{\widetilde{\theta}} L_\theta(\widetilde{\theta})$
2. $\beta \leftarrow \sqrt{2\delta / s^T A(\theta) s}$
3. Do
   $\widetilde{\theta} \leftarrow \theta + \beta s$
   $\beta \leftarrow \beta / C$
   while $L_\theta(\widetilde{\theta}) < \eta(\theta)$
4. Return $\widetilde{\theta}$

Proximal Policy Optimization

Perform conservative updates like TRPO, while being simpler to implement

Version1: Clipped Objective

- $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}$, $\hat{\mathbb{E}}$ denotes the empirical expectation with sampling distribution being the old distribution, ϵ is a hyper parameter

$$clip(r, a, b) = \begin{cases} a \text{ if } r < a \\ b \text{ if } r > b \\ r \text{ otherwise} \end{cases} \tag{64}$$

- Optimize the new objective for several epochs:

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{65}$$

Clipped Surrogate Loss

When $\left|r_t(\theta) - 1\right| > \epsilon$, the gradient vanishes.
New policy π_θ doesn't deviate too much from current policy π_{old}, which in effect is similar to TRPO.

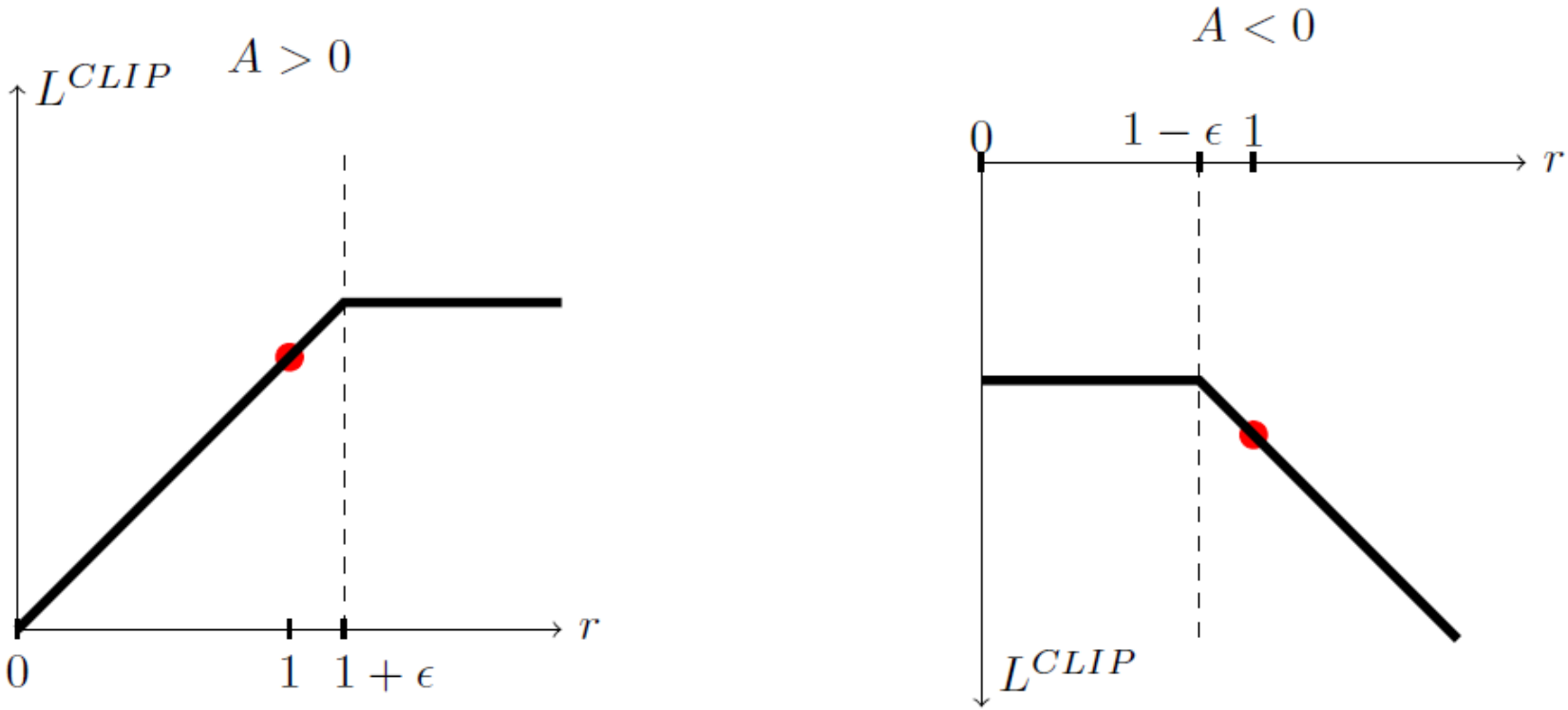


Figure: clipped surrogate objective vs r

Adaptive KL-penalty

Version2: Adaptive KL-penalty

- Optimize KL-penalized objective for several epochs

$$L^{KLPEN}(\theta) = \hat{\mathbb{E}}_t[\frac{\pi_\theta(a_t|s_t)}{\pi_{old}(a_t|s_t)}\hat{A}_t - \beta D_{KL}^{\rho_{\pi_{old}}}[\pi_{old}(\cdot|s_t), \pi_{theta}(\cdot|s_t)]] \quad (66)$$

- Adapt β after each policy update. Compute
  $d = KL[\pi_{old}(\cdot|s_t), \pi_{theta}(\cdot|s_t)]$.
  If $d \leq d_{targ}/1.5$, $\beta \leftarrow \beta/2$
  If $d \geq d_{targ} * 1.5$, $\beta \leftarrow \beta * 2$

Proximal Policy Optimization

Pseudo Code

Algorithm 1 PPO, Actor-Critic Style

for iteration=$1, 2, \ldots$ **do**
  for actor=$1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

References

- Kakade, S., & Langford, J. (2002, July). Approximately optimal approximate reinforcement learning. In ICML (Vol. 2, pp. 267-274).
- Schulman, J., Levine, S., Abbeel, P., Jordan, M. I., & Moritz, P. (2015, July). Trust Region Policy Optimization. In Icml (Vol. 37, pp. 1889-1897).
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
- Kakade, S. M. (2002). A natural policy gradient. In Advances in neural information processing systems (pp. 1531-1538).