Trust Region Policy Optimization

JOHN SCHULMAN, SERGEY LEVINE, PHILIPP MORITZ, MICHAEL I. JORDAN, PIETER ABBEEL

@ICML 2015

Presenter: 陈宏俊

Introduction

policy optimization

- **policy iteration method** — estimate value function & update policy
- **policy gradient** — $\theta = \theta_0 + \sum_t \alpha \nabla_{\theta_t} \log \pi_{\theta_t}(s_t, a_t) V_t$
- **derivative-free optimization** — cross-entropy method & covariance matrix adaptation

Policy Gradient

For i=1,2,…

Collect N trajectories for policy π_θ

Estimate advantage function A

Compute policy gradient g

Update policy parameter $\theta = \theta_{old} + \alpha g$

Problems of Policy Gradient

For $i=1,2,\ldots$

Collect N trajectories for policy π_θ

Estimate advantage function A

Compute policy gradient g

$\boxed{\text{Update policy parameter } \theta = \theta_{old} + \alpha g}$

Non stationary input data due to changing policy and reward distributions change

Problems of Policy Gradient

For i=1,2,...

Collect N trajectories for policy π_θ

$\boxed{\text{Estimate advantage function } A}$

Compute policy gradient g

Update policy parameter $\theta = \theta_{old} + \alpha g$

Advantage is very random initially

Problems of Policy Gradient

```
For i=1,2,...
        Collect N trajectories for policy π_θ
        Estimate advantage function A
        Compute policy gradient g
        Update policy parameter θ = θ_old + αg
```

$$\boxed{\text{Update policy parameter } \theta = \theta_{old} + \alpha g}$$

We need more carefully crafted policy update

We want improvement and not degradation

Main Idea

- We want to update old policy π_{old} to a new policy π_{new} such that they are “trusted” distance apart. Such conservative policy update allows quick and monotonical improvement instead of degradation.

Preliminaries

- Consider an infinite-horizon discounted Markov decision process (MDP), defined by the tuple $(S, A, P, r, \rho_0, \gamma)$
  - S: finite set of states
  - A: finite set of actions
  - P: S×A×S → R is the transition probability distribution
  - r: S → R is the reward function
  - ρ_0: S → R is the distribution of the initial state s_0
  - γ: $\gamma \in (0,1)$ is the discount factor

Expected Discounted Reward

- Let π denote a stochastic policy $\pi : S \times A \rightarrow [0,1]$, and let $\eta(\pi)$ denote its expected discounted reward:

$$\eta(\pi) = \mathbb{E}_{s_0, a_0, \ldots} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) \right], \text{ where}$$

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

Standard Definitions

$$Q_\pi(s_t, a_t) = \mathbb{E}_{s_{t+1}, a_{t+1}, \ldots} \left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l}) \right],$$

$$V_\pi(s_t) = \mathbb{E}_{a_t, s_{t+1}, \ldots} \left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l}) \right],$$

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s), \text{ where}$$

$$a_t \sim \pi(a_t | s_t), s_{t+1} \sim P(s_{t+1} | s_t, a_t) \text{ for } t \geq 0.$$

Expected Return for Another Policy

- The following useful identity expresses the expected return of another policy $\tilde{\pi}$ in terms of the advantage over π, accumu-lated over timesteps

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{s_0, a_0, \cdots \sim \tilde{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t) \right] \quad (1)$$

Proof

Proof. First note that $A_\pi(s,a) = \mathbb{E}_{s' \sim P(s'|s,a)}\left[r(s) + \gamma V_\pi(s') - V_\pi(s)\right]$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t A_\pi(s_t, a_t)\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t (r(s_t) + \gamma V_\pi(s_{t+1}) - V_\pi(s_t))\right] \\
&= \mathbb{E}_{\tau|\tilde{\pi}}\left[-V_\pi(s_0) + \sum_{t=0}^{\infty} \gamma^t r(s_t)\right] \\
&= -\mathbb{E}_{s_0}\left[V_\pi(s_0)\right] + \mathbb{E}_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right] \\
&= -\eta(\pi) + \eta(\tilde{\pi})
\end{aligned}$$

Discounted Visitation Frequencies

- We define ρ_π as the (unnormalized) discounted visitation frequencies:

$$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P(s_2 = s) + \ldots,$$

- where $s_0 \sim \rho_0$ and the actions are chosen according to π.

Rewrite with over-state-sum

- We can rewrite Equation (1) with a sum over states instead of timesteps:

$$
\begin{aligned}
\eta(\tilde{\pi}) &= \eta(\pi) + \sum_{t=0}^{\infty} \sum_{s} P(s_t = s|\tilde{\pi}) \sum_{a} \tilde{\pi}(a|s) \gamma^t A_\pi(s,a) \\
&= \eta(\pi) + \sum_{s} \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\tilde{\pi}) \sum_{a} \tilde{\pi}(a|s) A_\pi(s,a) \\
&= \eta(\pi) + \sum_{s} \rho_{\tilde{\pi}}(s) \sum_{a} \tilde{\pi}(a|s) A_\pi(s,a). \qquad (2)
\end{aligned}
$$

Rewrite with over-state-sum

$$\boxed{\eta(\tilde{\pi})} = \boxed{\eta(\pi_{old})} + \sum_s \boxed{\rho_{\tilde{\pi}}(s)} \sum_a \tilde{\pi}(a|s) A_\pi(s,a)$$

Expected return of old policy

Expected return of new policy

Discounted visitation frequency

$\rho_\pi(s) = P(s_0 = s) + \gamma P(s_1 = s) + \gamma^2 P + \cdots$

Guarantee Increasing Policy Update

$$\eta(\tilde{\pi}) = \eta(\pi_{old}) + \sum_s \rho_{\tilde{\pi}}(s) \boxed{\sum_a \tilde{\pi}(a|s) A_\pi(s,a)} \geq \mathbf{0}$$

New Expected Return $>$ Old Expected Return

Guaranteed Improvement from $\pi_{old} \rightarrow \tilde{\pi}$

Difficulty of Rewrite

State visitation based on new policy

$$\eta(\tilde{\pi}) = \eta(\pi_{old}) + \sum_{s} \boxed{\rho_{\tilde{\pi}}(s)} \sum_{a} \boxed{\tilde{\pi}(a|s)} A_{\pi}(s,a)$$

New policy

The complex dependency of $\rho_{\tilde{\pi}}(s)$ on $\tilde{\pi}$ makes Equation (2) difficult to optimize directly

Local Approximation to η

$$\eta(\tilde{\pi}) = \eta(\pi_{old}) + \sum_{s} \boxed{\rho_{\tilde{\pi}}(s)} \sum_{a} \tilde{\pi}(a|s) A_{\pi}(s,a) \quad (2)$$

$$L(\tilde{\pi}) = \eta(\pi_{old}) + \sum_{s} \boxed{\rho_{\pi}(s)} \sum_{a} \tilde{\pi}(a|s) A_{\pi}(s,a) \quad (3)$$

Local approximation of $\eta(\tilde{\pi})$

Local Approximation to η

- $\pi_\theta(a|s)$ is a differentiable function of the parameter vector θ, then L_π matches η to first order. That is, for any parameter value θ_{old}:

$$L_{\pi_{\theta_{old}}}(\pi_{\theta_{old}}) = \eta(\pi_{\theta_{old}})$$

$$\nabla_\theta L_{\pi_{\theta_{old}}}(\pi_\theta)\,|_{\theta=\theta old} = \nabla_\theta \eta(\pi_\theta)|_{\theta=\theta_{old}} \quad (4)$$

Proof

- For the first equation:

$$L_{\pi_{\theta_{old}}}(\pi_{\theta_{old}}) = \eta(\pi_{\theta_{old}}) + \sum_s \rho_{\pi_{\theta_{old}}}(s) \sum_a \pi_{\theta_{old}}(a \mid s) A_{\pi_{\theta_{old}}}(s, a) = \eta(\pi_{\theta_{old}})$$

- For the second equation:

$$\nabla_\theta L_{\pi_{\theta_{old}}}(\pi_\theta)\,|_{\theta=\theta_{old}} = \sum_s \rho_{\pi_{\theta_{old}}}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s) A_{\pi_{\theta_{old}}}(s, a)|_{\theta=\theta_{old}}$$

$$\nabla_\theta \eta(\pi_\theta)\,|_{\theta=\theta_{old}} = \sum_s \rho_{\pi_\theta}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s) A_{\pi_{\theta_{old}}}(s, a)|_{\theta=\theta_{old}}$$

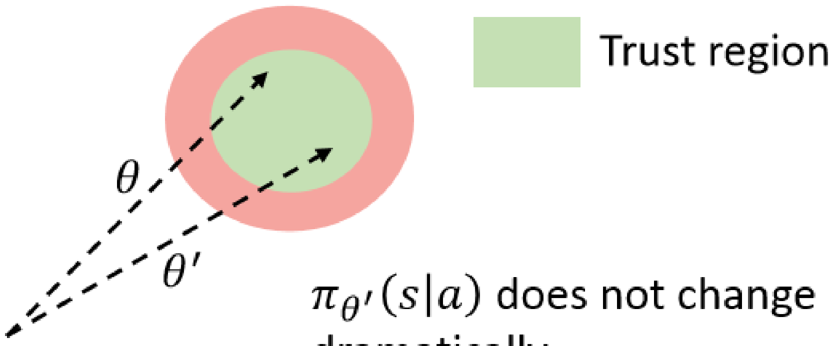
- In practice, $\sum_s \rho_{\pi_\theta}(s)$ is obtained from samples. When $\theta = \theta_{old}$, $\sum_s \rho_{\pi_\theta}(s) = \sum_s \rho_{\pi_{\theta_{old}}}(s)$, then both sides match.

Update with Small Step Size

$$L(\tilde{\pi}) = \eta(\pi_{old}) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_{\pi_{old}}(s,a)$$

The approximation is accurate within step size δ (trust region)

Monotonic improvement guaranteed



$\pi_{\theta'}(s|a)$ does not change dramatically.

Equation (4) implies that a sufficiently small step $\pi_{\theta_{old}} \to \tilde{\pi}$ that improves $L_{\pi_{\theta_{old}}}$ will also improve η, but does not give us any guidance on how big of a step to take.

Lower Bounds on the improvement of η

- To address this issue, Kakade & Langford(2002) proposed a policy updating scheme called **conservative policy iteration**, for which they could provide explicit lower bounds on the improvement of η.
- We define $\pi' = \arg\max_{\pi'} L_{\pi_{\text{old}}}(\pi')$
- The new policy π_{new} was defined to be the following mixture:

$$\pi_{\text{new}}(a|s) = (1 - \alpha)\pi_{\text{old}}(a|s) + \alpha\pi'(a|s). \qquad (5)$$

Lower Bounds on the improvement of η

- Kakade and Langford derived the following lower bound:

$$\eta(\pi_{\text{new}}) \geq L_{\pi_{\text{old}}}(\pi_{\text{new}}) - \frac{2\epsilon\gamma}{(1-\gamma)^2}\alpha^2$$
$$\text{where } \epsilon = \max_s \left| \mathbb{E}_{a\sim\pi'(a|s)}\left[A_\pi(s,a)\right] \right|. \qquad (6)$$

- However, that so far this bound only applies to mixture policies generated by Equation (5). This policy class is unwieldy and restrictive in practice, and it is desirable for a practical policy update scheme to be applicable to all general stochastic policy classes.

Monotonic Improvement Guarantee for General Stochastic Policies

- Our principal theoretical result is that the policy improvement bound in Equation (6) can be extended to general stochastic policies, by:
  - replacing α with a distance measure between π and $\tilde{\pi}$
  - changing the constant ϵ appropriately.

Total Variation Divergence

- The particular distance measure we use is the total variation divergence, which is defined by $D_{TV}(p \parallel q) = \frac{1}{2}\sum_i |p_i - q_i|$ for discrete probability distributions p, q.
- Define $D_{TV}^{max}(\pi, \tilde{\pi})$ as $$D_{\mathrm{TV}}^{\max}(\pi, \tilde{\pi}) = \max_s D_{TV}(\pi(\cdot|s) \parallel \tilde{\pi}(\cdot|s)). \quad (7)$$

Theorem 1

- Let $\alpha = D_{TV}^{max}(\pi, \tilde{\pi})$. Then the following bound holds:

$$\eta(\pi_{\text{new}}) \geq L_{\pi_{\text{old}}}(\pi_{\text{new}}) - \frac{4\epsilon\gamma}{(1-\gamma)^2}\alpha^2$$

$$\textit{where } \epsilon = \max_{s,a} |A_\pi(s,a)| \qquad (8)$$

Introducing KL Divergence

- Next, we note the following relationship between the **total variation divergence** and the **KL divergence :** $D_{TV}(p \parallel q)^2 \leq D_{\mathrm{KL}}(p \parallel q)$
- Let $D_{\mathrm{KL}}^{\max}(\pi, \tilde{\pi}) = \max_s D_{\mathrm{KL}}(\pi(\cdot|s) \parallel \tilde{\pi}(\cdot|s))$
- The following bound then follows directly from Theorem 1:

$$\eta(\tilde{\pi}) \geq L_\pi(\tilde{\pi}) - C D_{\mathrm{KL}}^{\max}(\pi, \tilde{\pi}),$$

$$\text{where } C = \frac{4\epsilon\gamma}{(1-\gamma)^2}. \qquad (9)$$

Guarantee improving policies

- Define i as the turn of iteration.
- Let $M_i(\pi) = L_{\pi_i}(\pi) - CD_{KL}^{max}(\pi_i, \pi)$. Then

$$
\begin{aligned}
&\eta(\pi_{i+1}) \geq M_i(\pi_{i+1}) \text{ by Equation (9)} \\
&\eta(\pi_i) = M_i(\pi_i), \text{ therefore,} \\
&\eta(\pi_{i+1}) - \eta(\pi_i) \geq M_i(\pi_{i+1}) - M(\pi_i). \qquad (10)
\end{aligned}
$$

- By maximizing M_i at each iteration, we guarantee that the true objective η is non-decreasing

Algorithm 1

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

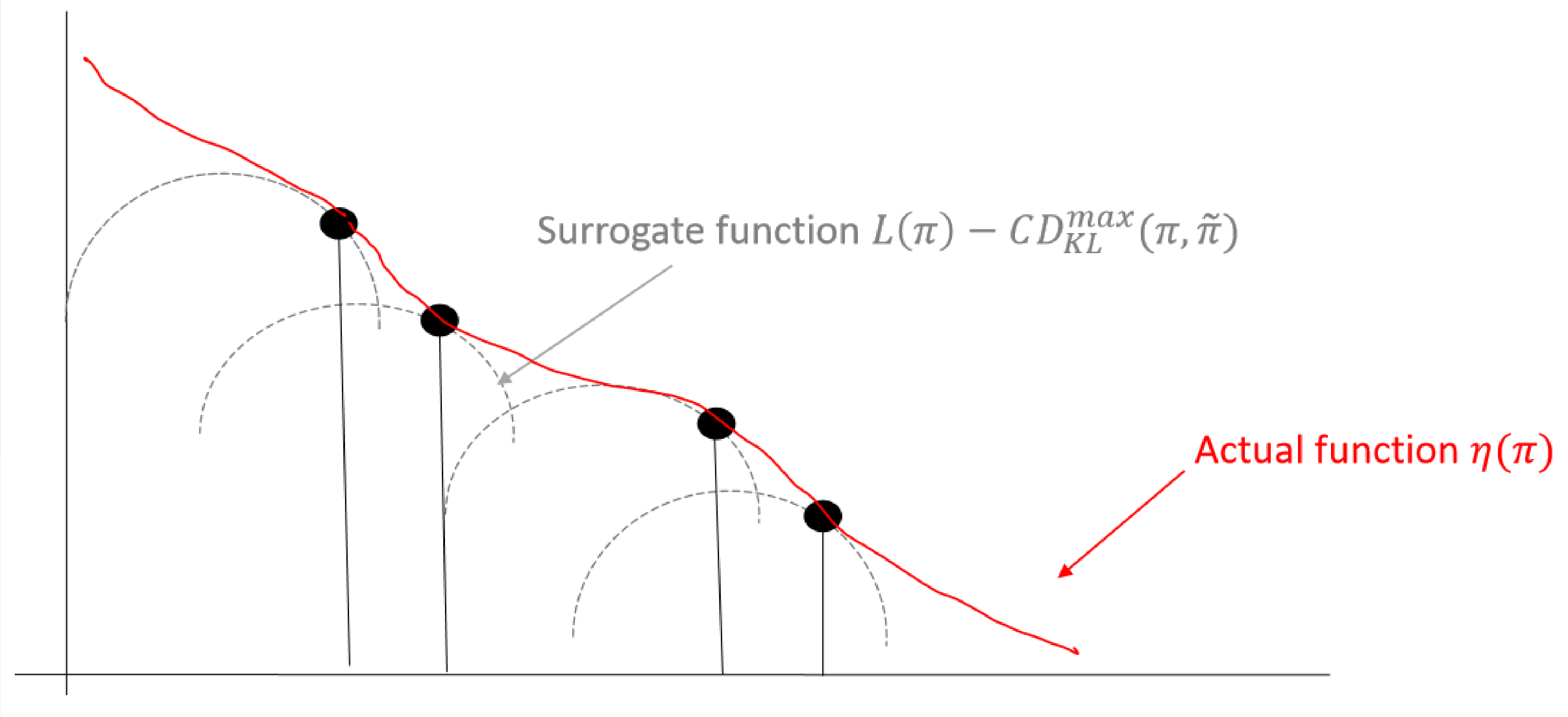
$$\pi_{i+1} = \arg\max_{\pi} \left[L_{\pi_i}(\pi) - C D_{\mathrm{KL}}^{\max}(\pi_i, \pi)\right]$$

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

Minorization-maximization (MM)



M_i is the surrogate function that minorizes η with equality at π_i.

TRPO

- Trust region policy optimization, which we propose in the following section, is an approximation to Algorithm 1, which uses a constraint on the KL divergence rather than a penalty to robustly allow large updates

Policy Update

- The preceding section showed that $\eta(\theta) = L_{\theta_{old}}(\theta) - CD_{KL}^{max}(\theta_{old}, \theta)$, with equality at $\theta = \theta_{old}$.
- Thus, by performing the following maximization, we are guaranteed to improve the true objective η :

$$\underset{\theta}{\text{maximize}} \left[L_{\theta_{\text{old}}}(\theta) - C D_{\text{KL}}^{\max}(\theta_{\text{old}}, \theta) \right]$$

To enlarge the step sizes

- One way to take larger steps in a robust way is to use a constraint on the KL divergence between the new policy and the old policy.
- **Trust region constraint:** $D_{\mathrm{KL}}^{\max}(\theta_{\mathrm{old}}, \theta) \leq \delta.$

$$\begin{aligned}
&\underset{\theta}{\operatorname{maximize}}\, L_{\theta_{\mathrm{old}}}(\theta) && (11)\\
&\text{subject to } D_{\mathrm{KL}}^{\max}(\theta_{\mathrm{old}}, \theta) \leq \delta.
\end{aligned}$$

A Heuristic Approximation on KL Divergence

- This problem imposes a constraint that the KL divergence is bounded at every point in the state space. While it is motivated by the theory, this problem is impractical to solve due to the large number of constraints.
- Instead, we can use a heuristic approximation which considers the average KL divergence:

$$\overline{D}_{\mathrm{KL}}^{\rho}(\theta_1, \theta_2) := \mathbb{E}_{s \sim \rho}\left[D_{\mathrm{KL}}(\pi_{\theta_1}(\cdot|s) \,\|\, \pi_{\theta_2}(\cdot|s))\right].$$

Solving the optimization problem

$$\underset{\theta}{\text{maximize}}\, L_{\theta_{\text{old}}}(\theta) \tag{11}$$
$$\text{subject to } D_{\text{KL}}^{\max}(\theta_{\text{old}}, \theta) \le \delta.$$

$$\underset{\theta}{\text{maximize}}\, L_{\theta_{\text{old}}}(\theta) \tag{12}$$
$$\text{subject to } \overline{D}_{\text{KL}}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}}, \theta) \le \delta.$$

Sample-Based Estimation of the Objective and Constraint

- The previous section proposed a constrained optimization problem on the policy parameters (Equation (12)), which optimizes an estimate of the expected total reward η subject to a constraint on the change in the policy at each update.
- This section describes how the objective and constraint functions can be approximated using Monte Carlo simulation.

Solving the Optimization Problem

$$\underset{\theta}{\text{maximize}}\, L_{\theta_{\text{old}}}(\theta) \tag{12}$$

$$\text{subject to } \overline{D}_{\text{KL}}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}}, \theta) \leq \delta.$$

$$L_\pi(\tilde{\pi}) = \eta(\pi) + \sum_s \rho_\pi(s) \sum_a \tilde{\pi}(a|s) A_\pi(s,a). \tag{3}$$

$$\underset{\theta}{\text{maximize}} \sum_s \rho_{\theta_{\text{old}}}(s) \sum_a \pi_\theta(a|s) A_{\theta_{\text{old}}}(s,a)$$

$$\text{subject to } \overline{D}_{\text{KL}}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}}, \theta) \leq \delta. \tag{13}$$

Rewrite the Optimization Problem

We first replace $\sum_s \rho_{\theta_{\text{old}}}(s)\,[\ldots]$ in the objective by the expectation $\frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}}[\ldots]$. Next, we replace the advantage values $A_{\theta_{\text{old}}}$ by the Q-values $Q_{\theta_{\text{old}}}$ in Equation (13), which only changes the objective by a constant. Last, we replace the sum over the actions by an importance sampling estimator. Using q to denote the sampling distribution, the contribution of a single s_n to the loss function is

Rewrite the Optimization Problem

$$\underset{\theta}{\text{maximize}} \boxed{\sum_s \rho_{\theta_{\text{old}}}(s)} \boxed{\sum_a \pi_\theta(a|s) \boxed{A_{\theta_{\text{old}}}(s,a)}}$$

$$\text{subject to } \overline{D}_{\mathrm{KL}}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}}, \theta) \le \delta. \qquad (13)$$

$$\underset{\theta}{\text{maximize}} \boxed{\mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}}}, \boxed{a\sim q \left[\frac{\pi_\theta(a|s)}{q(a|s)} \boxed{Q_{\theta_{\text{old}}}(s,a)}\right]} \qquad (14)$$

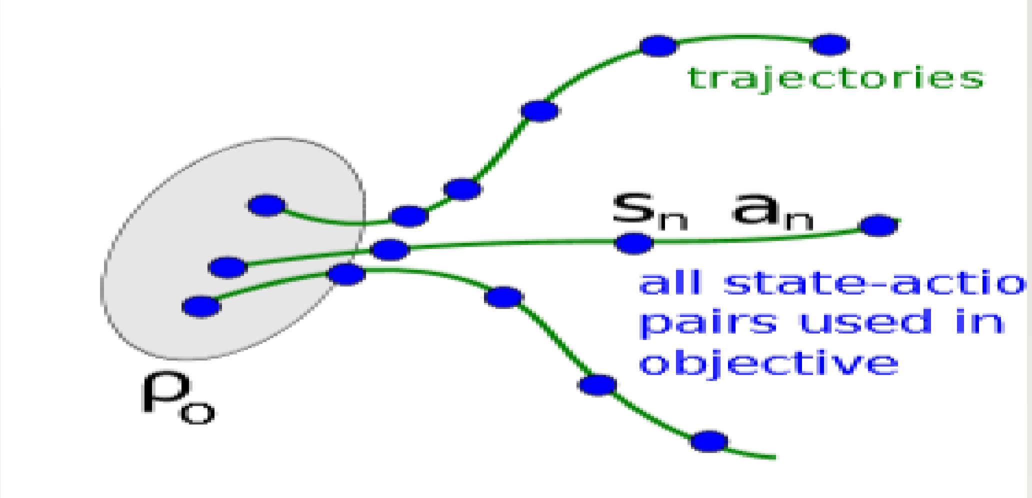
$$\text{subject to } \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}} \left[D_{\mathrm{KL}}(\pi_{\theta_{\text{old}}}(\cdot|s) \,\|\, \pi_\theta(\cdot|s))\right] \le \delta.$$

Sample-Based Estimation

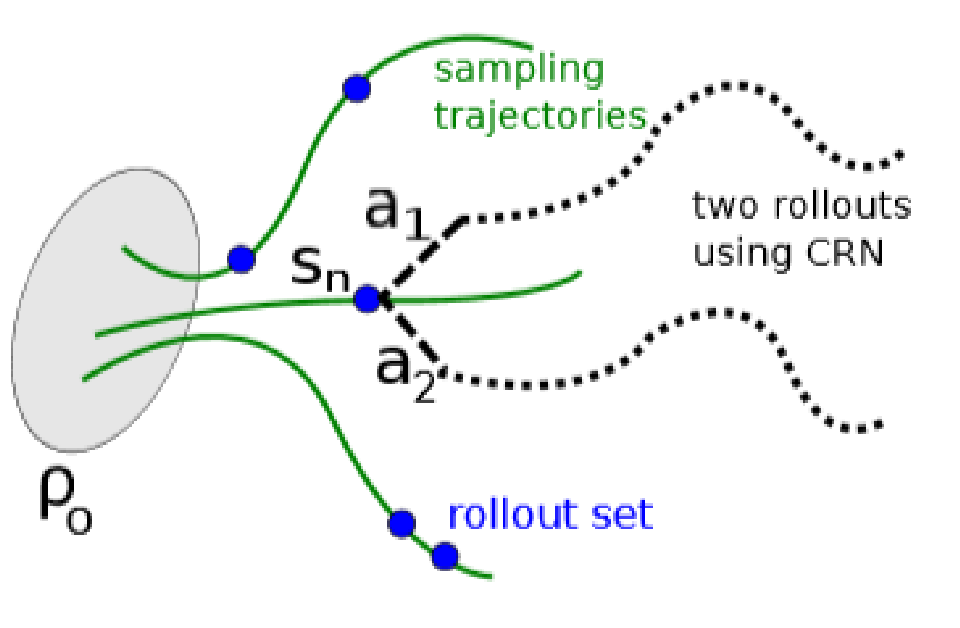
- All that remains is to replace the expectations by sample averages and replace the Q value by an empirical estimate. The following sections describe two different schemes for performing this estimation.
  - Single Path
  - Vine

Single Path

- In this estimation procedure, we collect a sequence of states by sampling $s_0 \sim \rho_0$ and then simulating the policy $\pi_{\theta_{old}}$ for some number of timesteps to generate a trajectory $s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T$. Hence $q(a|s) = \pi_{\theta_{\text{old}}}(a|s)$. $Q_{\theta_{\text{old}}}(s, a)$ is computed at each state-action pair(st,at) by taking the discounted sum of future rewards along the trajectory.



Vine



Small action space:

$$L_n(\theta) = \sum_{k=1}^{K} \pi_\theta(a_k|s_n)\hat{Q}(s_n, a_k), \tag{15}$$

Large or continuous action space:

$$L_n(\theta) = \frac{\sum_{k=1}^{K} \frac{\pi_\theta(a_{n,k}|s_n)}{\pi_{\theta_{\text{old}}}(a_{n,k}|s_n)} \hat{Q}(s_n, a_{n,k})}{\sum_{k=1}^{K} \frac{\pi_\theta(a_{n,k}|s_n)}{\pi_{\theta_{\text{old}}}(a_{n,k}|s_n)}}, \tag{16}$$

TRPO: KL-Constrained

- Unconstrained problem: $\underset{\theta}{\text{maximize}}\, L(\theta) - C.\overline{D_{KL}}(\theta_{old}, \theta)$
- Constrained problem: $\underset{\theta}{\text{maximize}}\, L(\theta)$ subject to $C.\overline{D_{KL}}(\theta_{old}, \theta) \leq \delta$
- δ is a hyper-parameter, remains fixed over whole learning process
- Solve constrained quadratic problem: compute $F^{-1}g$ and then rescale step to get correct KL
  - $\underset{\theta}{\text{maximize}}\, g\, .\, (\theta - \theta_{old})$ subject to $\frac{1}{2}(\theta - \theta_{old})^T F(\theta - \theta_{old}) \leq \delta$
  - Lagrangian: $\mathcal{L}(\theta, \lambda) = g\, .\, (\theta - \theta_{old}) - \frac{\lambda}{2}[(\theta - \theta_{old})^T F(\theta - \theta_{old}) - \delta]$
  - Differentiate wrt θ and get $\theta - \theta_{old} = \frac{1}{\lambda}F^{-1}g$
  - We want $\frac{1}{2}s^T F s = \delta$

$$g = \nabla l(\theta_{old})^T = \frac{\partial}{\partial\theta} l(\theta)|_{\theta=\theta_{old}}$$

$$F = H(kl)(\theta_{old}) = \frac{\partial^2}{\partial^2\theta} kl(\theta)\ |_{\theta=\theta_{old}}$$

TRPO: KL-Constrained

- Unconstrained problem: $\underset{\theta}{\text{maximize}}\, L(\theta) - C.\overline{D_{KL}}(\theta_{old}, \theta)$
- 近似为二次型：

$$\max_{\theta} g(\theta - \theta_{old}) - \frac{C}{2}(\theta - \theta_{old})^T F(\theta - \theta_{old})$$

$$g = \nabla l(\theta_{old})^T = \frac{\partial}{\partial \theta} l(\theta)|_{\theta=\theta_{old}}$$

$$F = H(kl)(\theta_{old}) = \frac{\partial^2}{\partial^2 \theta} kl(\theta)\ |_{\theta=\theta_{old}}$$

TRPO: KL-Constrained

solution:

$$\theta - \theta_{old} = \frac{1}{C} F^{-1} g$$

- 用共轭梯度法（Hessian Free）去计算 $F^{-1} g$

$$g = \nabla l(\theta_{old})^T = \frac{\partial}{\partial \theta} l(\theta)|_{\theta=\theta_{old}}$$

$$F = H(kl)(\theta_{old}) = \frac{\partial^2}{\partial^2 \theta} kl(\theta)\ |_{\theta=\theta_{old}}$$

TRPO Algorithm

For i=1,2,…

Collect N trajectories for policy π_θ
Estimate advantage function A
Compute policy gradient g
Use CG to compute $H^{-1}g$
Compute rescaled step $s = \alpha H^{-1}g$ with rescaling and line search
Apply update: $\theta = \theta_{old} + \alpha H^{-1}g$

$\underset{\theta}{\text{maximize}}\ L(\theta)$ subject to $C.\overline{D_{KL}}(\theta_{old}, \theta) \leq \delta$