# Trust Region Policy Optimization

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

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

$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$

Policy Evaluation

$V_{i+1}^{\pi_k}(s) = \sum_{s'} P(s'|s, \pi_k(s))[r + \gamma V_i^{\pi_k}(s')]$

Policy Improvement

$\pi_{k+1}(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s, a)[r + \gamma V^{\pi_k}(s')]$

n – step return:

$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^n \hat{v}(s_{t+n})$

$\nabla J(\theta) = \mathbb{E}_\pi \left[ G_t \frac{\nabla \pi(a|s, \theta)}{\pi(a|s, \theta)} \right]$

$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a \left( q_\pi(s, a) - b(s) \right) \nabla \pi(a|s, \theta)$

$\nabla J(\theta) = \mathbb{E}_\pi[\nabla_\theta \log \pi_\theta(a_t|s_t) \hat{A}_t]$

Expected discounted reward for a stochastic policy $\pi_\theta$:

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

$a_t \sim \pi(a_t|s_t, \theta)$, $s_{t+1} \sim P(s_{t+1}|s_t, a_t)\ \forall\ t \geq 0$

State 's' distribution: $\rho(s)$

$\eta(\beta) = \eta(\theta) + \mathbb{E}_{s_0, a_0, \ldots \sim \pi_\beta}[\sum_{t=0}^{\infty} \gamma^t A_{\pi_\theta}(s_t, a_t)]$

$\eta(\beta) = \eta(\theta) + \sum_s \rho_\beta(s) \sum_a \pi_\beta(a|s) A_{\pi_\theta}(s, a)$

Complex dependency of $\rho_\beta(s)$on $\beta$ is difficult to optimize

Local approximation to $\eta$:

$J_{\pi_\theta}(\beta) = \eta(\theta) + \sum_s \rho_{\pi_\theta}(s) \sum_a \pi_\beta(a|s) A_{\pi_\theta}(s, a)$

Doesn't specify a step size

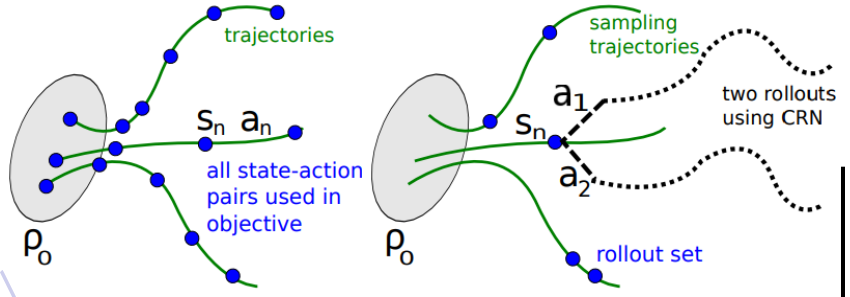
$\theta$: Current policy parameters

$\beta$: Behaviour parameters

$\theta'$: New policy parameters

- Performance function (at $timestep = t$) optimization problem in policy gradient methods:
  - $\nabla J(\theta) = \mathbb{E}_\pi[\nabla_\theta \ln \pi_\theta(a_t|s_t) \hat{A}_t]$ ← Gradient equation
  - $J(\theta) = \mathbb{E}_\pi[\ln \pi_\theta(a_t|s_t) \hat{A}_t]$ ← Maximize this cost function
  - $\theta_{t+1} = \theta_t + \alpha \hat{A}_t \nabla \ln \pi(a_t|s_t, \theta_t) \Rightarrow \theta_{new} \gg \theta_{old}$
  - How to set step size for update?
- Conservative policy (Kakade & Langford, 2002):
  - $\pi_{new}(a|s) = (1 - \alpha)\pi_{old}(a|s) + \alpha \pi'(a|s)$
  - Where, $\pi' = \arg\max_{\pi'} J_{\pi_{old}}(\pi')$
- Kakade & Langford derived the following lower bound:
  - $\eta(\pi_{new}) \geq J_{\pi_{old}}(\pi_{new}) - \frac{2\epsilon\gamma}{(1-\gamma)^2} \alpha^2$
  - $\epsilon = \max_s \left| \mathbb{E}_{a \sim \pi'(a|s)}[A_\pi(s, a)] \right|$
- To generalize for stochastic policies:
  - $\alpha = D_{TV}^{max}(\pi_{old}, \pi_{new}) = \max_s D_{TV}\left(\pi_{old}(\cdot|s) || \pi_{new}(\cdot|s)\right)$
  - $\epsilon = \max_{s, a} |A_\pi(s, a)|$
- Theorem 1: For $\alpha = D_{TV}^{max}(\pi_{old}, \pi_{new})$. Following bound holds:
  - $\eta(\pi_{new}) \geq J_{\pi_{old}}(\pi_{new}) - \frac{4\epsilon\gamma}{(1-\gamma)^2} \alpha^2$
- Theorem 2: $D_{TV}(p \,||\, q)^2 \leq D_{KL}(p \,||\, q)$
  - $\eta(\pi_{new}) \geq J_{\pi_{old}}(\pi_{new}) - \frac{4\epsilon\gamma}{(1-\gamma)^2} D_{KL}^{max}(\pi_{old}, \pi_{new})$

$C$




Single path Vine

- To improve true objective $\eta$:
  - $maximize_{\pi_{new}}[J_{\pi_{old}}(\pi_{new}) - CD_{KL}^{max}(\pi_{old}, \pi_{new})]$
  - Or, $maximize_{\theta'}[J_\theta(\theta') - CD_{KL}^{max}(\theta, \theta')]$
- Practically, $C$ reduces the step sizes. Therefore, to take larger steps, a limiting constraint $\delta$ is preferred
  - $maximize_{\theta'}\ J_\theta(\theta')$
  - $s.t. D_{KL}^{max}(\theta, \theta') \leq \delta$
- Sample-based estimation of objective & constraint:
  - $maximize_{\theta'}\ \mathbb{E}_{s \sim \rho_\theta,\, a \sim q}\left[\frac{\pi_{\theta'}(a|s)}{q(a|s)} Q_\theta(s, a)\right]$
  - $s.t.\ \mathbb{E}_{s \sim \rho_\theta}\left[D_{KL}\left(\pi_\theta(\cdot|s) \parallel \pi_{\theta'}(\cdot|s)\right)\right] \leq \delta$

**Algorithm:** TRPO Algorithm

1. Use single path or vine to collect a set of $s - a$ pairs along with Monte Carlo estimates of their Q-values.
2. $maximize_{\theta'}\ \mathbb{E}_{s \sim \rho_\theta,\, a \sim q}\left[\frac{\pi_{\theta'}(a|s)}{q(a|s)} Q_\theta(s, a)\right]$
   $s.t.\ \mathbb{E}_{s \sim \rho_\theta}\left[D_{KL}\left(\pi_\theta(\cdot|s) \parallel \pi_{\theta'}(\cdot|s)\right)\right] \leq \delta$
3. Approximately solve this constrained optimization problem to update the policy's parameter vector $\theta'$. A conjugate gradient algorithm followed by line search is used.

https://www.profrkumar.co.in/

Schulman, John, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. "Trust region policy optimization." In International conference on machine learning, pp. 1889-1897. PMLR, 2015.