CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 11

Sharan Vaswani

November 17, 2023

Recap

- **Tabular softmax policy parameterization**: There are SA parameters such that $\pi_\theta(\cdot|s) = h(\theta(s,\cdot))$. In this case, $[\nabla J(\theta)]_{s,a} = \frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta(s,a)} = \frac{d^{\pi_\theta}(s)}{1-\gamma}\, \pi_\theta(a|s)\, \mathfrak{a}^{\pi_\theta}(s,a)$, where $\mathfrak{a}^\pi(s,a) = q^\pi(s,a) - v^\pi(s)$ is the advantage function.
- **Softmax PG**: For the bandit setting with deterministic rewards, softmax PG with the tabular parameterization has the following update: $\theta_{t+1} = \theta_t + \eta\, \pi_{\theta_t}(a)\, [r(a) - \langle \pi_{\theta_t}, r \rangle]$.
- With exact gradients, softmax PG with the tabular parameterization converges to the optimal policy at an $O(1/T)$ rate for both bandits and general MDPs.
- **Natural policy gradient (NPG)**: It preconditions the policy gradient by the inverse Fisher information matrix $(F_\theta^\dagger)$ and results in faster convergence.
- For the tabular softmax parameterization, the preconditioned gradient direction is: $[F_\theta^\dagger\, \nabla J(\theta)]_{s,a} = \frac{\mathfrak{a}^{\pi_\theta}(s,a)}{1-\gamma}$, and the corresponding NPG update for (s,a) is given as: $\theta_{t+1}(s,a) = \theta_t(s,a) + \eta\, \frac{\mathfrak{a}^{\pi_t}(s,a)}{1-\gamma}$.

Natural Policy Gradient for Softmax Parametrization

Defining $\pi_t := \pi_{\theta_t}$, the NPG update corresponding to the tabular softmax parameterization, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ is given by: $\theta_{t+1}(s,a) = \theta_t(s,a) + \eta \frac{\mathfrak{a}^{\pi_t}(s,a)}{1-\gamma}$. Exponentiating both sides,

$\mathfrak{a}^{\pi_t}(s,a)$ 是 Advantage function

$$\exp(\theta_{t+1}(s,a)) = \exp(\theta_t(s,a)) \exp\left(\frac{\eta\, \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma}\right)$$

代入

$$\pi_{t+1}(a|s) \overset{定义}{=} \frac{\exp(\theta_{t+1}(s,a))}{\sum_{a'} \exp(\theta_{t+1}(s,a'))} = \frac{\exp(\theta_t(s,a)) \exp\left(\frac{\eta\, \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma}\right)}{\sum_{a'} \exp(\theta_t(s,a')) \exp\left(\frac{\eta\, \mathfrak{a}^{\pi_t}(s,a')}{1-\gamma}\right)}$$

硬提出来

distribution

$$= \boxed{\frac{\exp(\theta_t(s,a))}{\sum_{\tilde{a}} \exp(\theta_t(s,\tilde{a}))}} \exp\left(\frac{\eta\, \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma}\right) \frac{1}{\sum_{a'} \frac{\exp(\theta_t(s,a'))}{\sum_{\tilde{a}} \exp(\theta_t(s,\tilde{a}))} \exp\left(\frac{\eta\, \mathfrak{a}^{\pi_t}(s,a')}{1-\gamma}\right)}$$

distribution

用q(s,a)换advantage

$$\implies \pi_{t+1}(a|s) = \frac{\pi_t(a|s) \exp\left(\frac{\eta\, \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma}\right)}{\sum_{a'} \pi_t(a'|s) \exp\left(\frac{\eta\, \mathfrak{a}^{\pi_t}(s,a')}{1-\gamma}\right)} = \frac{\pi_t(a|s) \exp\left(\frac{\eta\, q^{\pi_t}(s,a)}{1-\gamma}\right)}{\sum_{a'} \pi_t(a'|s) \exp\left(\frac{\eta\, q^{\pi_t}(s,a')}{1-\gamma}\right)}$$

This is exactly the multiplicative weights from Lecture 9. Hence, for the softmax tabular policy parameterization, NPG is equivalent to mirror ascent with a negative entropy mirror map.

Convergence of Natural Policy Gradient for Softmax Parametrization

Similar to the proof for softmax PG, we will prove a non-uniform Lojasiewicz condition ? for NPG. We will do the proof for the bandits setting, where $J(\theta) = \langle \pi_\theta, r \rangle$ and the corresponding NPG update can be written as: for action a, $\pi_{t+1}(a) = \frac{\pi_t(a)\, \exp(\eta\, r(a))}{\sum_{a'} \pi_t(a')\, \exp(\eta\, r(a'))}$.

Claim: Define π' s.t. $\pi'(a) := \frac{\pi(a)\, \exp(\eta\, r(a))}{\sum_{a'} \pi(a')\, \exp(\eta\, r(a'))}$. Assuming that the arms are numbered in order of their rewards i.e. $r(1) > r(2) > \ldots$, $\Delta(a) := r(1) - r(a)$ and $\Delta := \min_{a \neq 1} \Delta(a) = r(1) - r(2)$, then, $\langle \pi' - \pi, r \rangle \geq \left[1 - \frac{1}{\pi(a^*)\,(\exp(\eta\Delta) - 1) + 1}\right] \langle \pi^* - \pi, r \rangle$.

- The LHS is the improvement in one step and is similar to the gradient for softmax PG.
- As the algorithm approaches a stationary point (such that $\pi' \approx \pi$), the LHS tends to zero. The RHS also tends to zero, meaning that π converges to the optimal policy.
- A similar Lojasiewicz property holds for general MDPs, and can be used to prove linear convergence to the optimal policy [MDX+21, Theorem 12].
- Importantly, for general MDPs, NPG can be proven to achieve a linear rate of convergence matching policy iteration and without a dependence on the distribution mismatch ratio [JPBR23, Theorem 1].

Convergence of Natural Policy Gradient for Softmax Parametrization

Proof: $(\pi' - \pi)^\top r = \sum_{i=1}^K [\pi'(i)\, r(i) - \pi(i)\, r(i)] = \sum_{i=1}^K \left[\frac{\pi(i)\, e^{\eta\, r(i)}\, r(i)}{\sum_{j=1}^K \pi(j)\, e^{\eta\, r(j)}} - \pi(i)\, r(i) \right]$

$$= \frac{1}{\sum_{j=1}^K \pi(j)\, e^{\eta\, r(j)}} \underbrace{\left(\sum_{i=1}^K \pi(i)\, e^{\eta\, r(i)}\, r(i) - \sum_{i=1}^K \pi(i)\, r(i) \sum_{j=1}^K \pi(j)\, e^{\eta\, r(j)} \right)}_{\text{(i)}}$$

把 i=j 拆出来,

$$\text{(i)} = \sum_{i=1}^K \pi(i)\, e^{\eta\, r(i)}\, r(i) - \sum_{i=1}^K [\pi(i)]^2\, r(i)\, e^{\eta\, r(i)} - \sum_{i=1}^K \pi(i)\, r(i) \boxed{\sum_{j=1, j\neq i}^K \pi(j)\, e^{\eta\, r(j)}}$$

$$= \sum_{i=1}^K \underbrace{\pi(i)}_{a_i} \underbrace{e^{\eta r(i)} r(i)}_{b_i} \boxed{\sum_{j=1, j\neq i}^K \underbrace{\pi(j)}_{a_j}} - \sum_{i=1}^K \pi(i)\, r(i) \sum_{j=1, j\neq i}^K \pi(j)\, e^{\eta\, r(j)} \qquad \boxed{(1 - \pi(i) = \textstyle\sum_{j\neq i} \pi(j))}$$

$$= \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^K \pi(j)\, [e^{\eta r(i)} r(i) + e^{\eta r(j)} r(j)] - \sum_{i=1}^K \pi(i)\, r(i) \sum_{j=1, j\neq i}^K \pi(j)\, e^{\eta\, r(j)}$$

(For any a_i, b_i, $\sum_{i=1}^K \underline{\underline{a_i}}\, \underline{\underline{b_i}} \sum_{j=1, j\neq i}^K \underline{\underline{a_j}} = \sum_{i=1}^{K-1} a_i \sum_{j=i+1}^K a_j\, [b_i + b_j]$)

Convergence of Natural Policy Gradient for Softmax Parametrization

Recall that $(i) = \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j) \, [e^{\eta r(i)} r(i) + e^{\eta r(j)} r(j)] - \sum_{i=1}^{K} \pi(i) \, r(i) \sum_{j=1, j\neq i}^{K} \pi(j) \, e^{\eta \, r(j)}$

$$\sum_{i=1}^{K} \pi(i) \, r(i) \sum_{j=1, j\neq i}^{K} \pi(j) \, e^{\eta \, r(j)} = \sum_{i=1}^{K} \underbrace{\pi(i) e^{\eta \, r(i)}}_{a_i} \underbrace{\frac{r(i)}{e^{\eta \, r(i)}}}_{b_i} \sum_{j=1, j\neq i}^{K} \underbrace{\pi(j) \, e^{\eta \, r(j)}}_{a_j}$$

$$= \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j) \, [e^{\eta \, r(j)} \, r(i) + e^{\eta \, r(i)} \, r(j)]$$

$$\left(\sum_{i=1}^{K} a_i b_i \sum_{j=1, j\neq i}^{K} a_j = \sum_{i=1}^{K-1} a_i \sum_{j=i+1}^{K} a_j \, [b_i + b_j]\right)$$

$$\implies (i) = \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j) \, [e^{\eta r(i)} r(i) + e^{\eta r(j)} r(j)] - \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j) \, [e^{\eta r(j)} r(i) + e^{\eta r(i)} r(j)]$$

$$= \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j) \, [e^{\eta r(i)} - e^{\eta r(j)}] \, [r(i) - r(j)]$$

Convergence of Natural Policy Gradient for Softmax Parametrization

Recall that $(\pi' - \pi)^\top r = \frac{(i)}{\sum_{j=\mathbf{1}}^{K} \pi(j)\, e^{\eta\, r(j)}}$, (i) $= \sum_{i=1}^{K-1} \pi(i) \sum_{j=i+1}^{K} \pi(j) \, [e^{\eta r(i)} - e^{\eta r(j)}] \, [r(i) - r(j)]$.

$$
\begin{aligned}
(i) &\geq \pi(1) \sum_{j=2}^{K} \pi(j) \left[e^{\eta\, r(1)} - e^{\eta\, r(j)} \right] [r(1) - r(j)] && \text{(Only using the first term)} \\
&\geq \pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1 \right) \sum_{j=2}^{K} \pi(j)\, [r(1) - r(j)] && (r(j) \leq r(2),\ \Delta = r(1) - r(2)) \\
&= \pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1 \right) \sum_{a \neq a^*} \pi(a)\, \Delta(a) && \text{(Arm 1 is the optimal arm)} \\
&= \pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1 \right) \sum_{a} \pi(a)\, \Delta(a) && (\Delta(a^*) = 0) \\
&= \pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1 \right) (\pi^* - \pi)^\top r && (\text{Since } \pi^*(a^*) = 1)
\end{aligned}
$$

$$
\implies (\pi' - \pi)^\top r \geq \frac{\pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1 \right)}{\sum_{j=1}^{K} \pi(j)\, e^{\eta\, r(j)}} \, (\pi^* - \pi)^\top r
$$

Convergence of Natural Policy Gradient for Softmax Parametrization

Recall that $(\pi' - \pi)^\top r \geq \frac{\pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1\right)}{\sum_{j=1}^K \pi(j)\, e^{\eta\, r(j)}} (\pi^* - \pi)^\top r$. Simplifying,

$$\frac{\pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1\right)}{\sum_{j=1}^K \pi(j)\, e^{\eta\, r(j)}} = \frac{\pi(1)\, e^{\eta\, r(2)} \left(e^{\eta\, \Delta} - 1\right)}{\pi(1)\, e^{\eta\, r(1)} + \sum_{j=2}^K \pi(j)\, e^{\eta\, r(j)}}$$

除以$e^{\eta r(2)}$

$$= \frac{\pi(1) \left(e^{\eta\, \Delta} - 1\right)}{\pi(1)\, e^{\eta\, \Delta} + \sum_{j=2}^K \pi(j)\, e^{\eta\, [r(j) - r(2)]}} \geq \frac{\pi(1) \left(e^{\eta\, \Delta} - 1\right)}{\pi(1)\, e^{\eta\, \Delta} + \sum_{j=2}^K \pi(j)}$$

(Since $r(j) \leq r(2)$ for $j \geq 2$)

$$= \frac{\pi(1) \left(e^{\eta\, \Delta} - 1\right)}{\pi(1)\, e^{\eta\, \Delta} + 1 - \pi(1)} = \frac{\pi(1) \left(e^{\eta\, \Delta} - 1\right)}{\pi(1) \left(e^{\eta\, \Delta} - 1\right) + 1} = 1 - \frac{1}{\pi(a^*) \left(e^{\eta\, \Delta} - 1\right) + 1}$$

$$\implies (\pi' - \pi)^\top r \geq \left[1 - \frac{1}{\pi(a^*) \left(e^{\eta\, \Delta} - 1\right) + 1}\right] (\pi^* - \pi)^\top r \quad \square$$

Goal: We will now use this non-uniform Lojasiewicz condition to prove global convergence to the optimal policy for NPG.

Convergence of Natural Policy Gradient for Softmax Parametrization

Claim: For a bandit problem with deterministic rewards and $\Delta := r(a^*) - \max_{a \neq a^*} r(a)$, NPG with the softmax tabular policy parameterization, any step-size η and T iterations results in the following convergence: if $\delta_t := \langle \pi^*, r \rangle - \langle \pi_{\theta_t}, r \rangle$, then, $\delta_T \leq \exp(-cT)\,\delta_0$ where $c := \log\left(\pi_{\theta_{\mathbf{0}}}(a^*)\left(e^{\eta\,\Delta} - 1\right)\right) + 1)$.

Proof: $\delta_{t+1} = \langle \pi^*, r \rangle - \langle \pi_{\theta_{t+\mathbf{1}}}, r \rangle = \delta_t - \langle \pi_{\theta_{t+\mathbf{1}}} - \pi_{\theta_t}, r \rangle$. Recall that the NPG update is $\pi_{t+1}(a) = \frac{\pi_t(a)\,\exp(\eta\, r(a))}{\sum_{a'} \pi_t(a')\,\exp(\eta\, r(a'))}$. Using the non-uniform Lojasiewicz condition,

$$\delta_{t+1} \leq \delta_t - \left[1 - \frac{1}{\pi_{\theta_t}(a^*)\,(e^{\eta\,\Delta} - 1) + 1}\right](\pi^* - \pi_{\theta_t})^\top r = \frac{\delta_t}{\pi_{\theta_t}(a^*)\,(e^{\eta\,\Delta} - 1) + 1}$$

$$\pi_{\theta_{t+\mathbf{1}}}(a^*) = \pi_{t+1}(a^*) = \frac{\pi_t(a^*)\,\exp(\eta\, r(a^*))}{\sum_{a'} \pi_t(a')\,\exp(\eta\, r(a'))} = \frac{\pi_t(a^*)}{\sum_{a'} \pi_t(a')\,\exp(\eta\,[r(a') - r(a^*)])} \geq \pi_t(a^*)$$

$$\implies \pi_t(a^*) \geq \pi_0(a^*) \implies \delta_{t+1} \leq \frac{\delta_t}{\pi_{\theta_{\mathbf{0}}}(a^*)\,(e^{\eta\,\Delta} - 1) + 1}$$

$$\implies \delta_T \leq \frac{\delta_0}{\left[\pi_{\theta_{\mathbf{0}}}(a^*)\,(e^{\eta\,\Delta} - 1) + 1\right]^T} = \exp(-cT)\,\delta_0 \quad \square$$

Handling Stochasticity

Stochastic Softmax Policy Gradient for Bandits

- Until now, we have assumed that we have access to the full gradient $\nabla J(\theta)$. For bandits, the full gradient involves computing $\pi_\theta(a)[r(a) - \langle \pi_\theta, r \rangle]$ for all a in each iteration.
- In order to make the resulting algorithms more practical, we now focus on *stochastic PG methods* for bandits with deterministic rewards. The algorithm pulls only one arm in each iteration to compute a gradient estimate.
- Importance-weighted reward estimator at iteration t: $\hat{r}_t(a) := \frac{\mathcal{I}\{a_t = a\}}{\pi_\theta(a)} \, r(a)$ where a_t is the arm pulled at iteration t. Hence, $\mathbb{E}_{a_t \sim \pi_\theta}[\hat{r}_t(a)] = r(a)$.
- Stochastic softmax PG update:

$$\theta_{t+1} = \theta_t + \eta_t \, \tilde{\nabla} J(\theta_t) \quad ; \quad [\tilde{\nabla} J(\theta)]_a := \frac{\partial \langle \pi_\theta, \hat{r}_t \rangle}{\partial \theta(a)} = \pi_\theta(a) \, [\hat{r}_t(a) - \langle \pi_\theta, \hat{r}_t \rangle] \, .$$

- We will first show that the gradient estimator $\tilde{\nabla} J(\theta_t)$ is unbiased and has bounded variance.

Stochastic Softmax Policy Gradient for Bandits

Claim: The estimator $\tilde{\nabla} J(\theta)$ is unbiased, i.e. $\mathbb{E}_{a_t \sim \pi_\theta}[\tilde{\nabla} J(\theta)] = \nabla J(\theta)$.

Proof: Recall that $\frac{\partial \langle \pi_\theta, r \rangle}{\partial \theta(a)} = \pi_\theta(a)[r(a) - \langle \pi_\theta, r \rangle]$.

$$[\tilde{\nabla} J(\theta)]_a = \frac{\partial \langle \pi_\theta, \hat{r}_t \rangle}{\partial \theta(a)} = \pi_\theta(a)[\hat{r}_t(a) - \langle \pi_\theta, \hat{r}_t \rangle] = \pi_\theta(a) \left[\frac{\mathcal{I}\left\{a_t = a\right\} \, r(a)}{\pi_\theta(a)} - \sum_{a'} \pi_\theta(a') \, \hat{r}_t(a') \right]$$

$$= \mathcal{I}\left\{a_t = a\right\} \, r(a) - \pi_\theta(a) \sum_{a'} \pi_\theta(a') \, \frac{\mathcal{I}\left\{a_t = a'\right\} \, r(a')}{\pi_\theta(a')}$$

$$= \mathcal{I}\left\{a_t = a\right\} \, r(a) - \pi_\theta(a) \, r(a_t)$$

$$\implies \mathbb{E}_{a_t \sim \pi_\theta} \left[\frac{\partial \langle \pi_\theta, \hat{r}_t \rangle}{\partial \theta(a)} \right] = \sum_{a_t} \pi_\theta(a_t) \, [\tilde{\nabla} J(\theta)]_a = \sum_{a_t} \pi_\theta(a_t) \, [\mathcal{I}\left\{a_t = a\right\} \, r(a) - \pi_\theta(a) \, r(a_t)]$$

$$= \pi_\theta(a) \, r(a) - \pi_\theta(a) \sum_{a_t} \pi_\theta(a_t) \, r(a_t) = \pi_\theta(a) \, [r(a) - \langle \pi_\theta, r \rangle]$$

$$\implies \mathbb{E}_{a_t \sim \pi_\theta} \left[\frac{\partial \langle \pi_\theta, \hat{r}_t \rangle}{\partial \theta(a)} \right] = \frac{\partial \langle \pi_\theta, r \rangle}{\partial \theta(a)} \implies \mathbb{E}_{a_t \sim \pi_\theta} \left[\frac{\partial \langle \pi_\theta, \hat{r}_t \rangle}{\partial \theta} \right] = \frac{\partial \langle \pi_\theta, r \rangle}{\partial \theta} \quad \square$$

Stochastic Softmax Policy Gradient for Bandits

Claim: For rewards in $[0, 1]$, $\left\|\tilde{\nabla} J(\theta)\right\|^2 \leq 2$.

Proof: $\left\|\tilde{\nabla} J(\theta)\right\|^2 = \sum_a \left(\frac{\partial\langle\pi_\theta, \hat{r}_t\rangle}{\partial\theta(a)}\right)^2 = \sum_a [\pi_\theta(a)]^2 \overbrace{[\hat{r}_t(a) - \langle\pi_\theta, \hat{r}_t\rangle]^2}^{\text{(i)}}$.

$$\text{(i)} = \frac{\mathcal{I}\{a_t = a\}\,[r(a)]^2}{[\pi_\theta(a)]^2} - \frac{2\,\mathcal{I}\{a_t = a\}\,r(a)}{\pi_\theta(a)} \sum_{a'} \mathcal{I}\{a_t = a'\}\,r(a') + \left(\sum_{a'} \mathcal{I}\{a_t = a'\}\,r(a')\right)^2$$

$$= \frac{\mathcal{I}\{a_t = a\}\,[r(a)]^2}{[\pi_\theta(a)]^2} - \frac{2\,\mathcal{I}\{a_t = a\}\,r(a)\,r(a_t)}{\pi_\theta(a)} + [r(a_t)]^2$$

$$\implies \left\|\tilde{\nabla} J(\theta)\right\|^2 = \sum_a \left[\mathcal{I}\{a_t = a\}\,[r(a)]^2 - 2\,\mathcal{I}\{a_t = a\}\,r(a)\,r(a_t)\,\pi_\theta(a) + [\pi_\theta(a)]^2\,[r(a_t)]^2\right]$$

$$= [r(a_t)]^2 - 2\pi_\theta(a_t)\,[r(a_t)]^2 + \sum_a [\pi_\theta(a)]^2\,[r(a_t)]^2$$

$$= (1 - \pi_\theta(a_t))\,[r(a_t)]^2 - \pi_\theta(a_t)\,[r(a_t)]^2 + [\pi_\theta(a_t)]^2\,[r(a_t)]^2 + \sum_{a \neq a_t} [\pi_\theta(a)]^2\,[r(a_t)]^2$$

$$= (1 - \pi_\theta(a_t))^2\,[r(a_t)]^2 + \sum_{a \neq a_t} [\pi_\theta(a)]^2\,[r(a_t)]^2$$

Stochastic Softmax Policy Gradient for Bandits

Recall that $\left\|\nabla\tilde{J}(\theta)\right\|^2 \leq (1-\pi_\theta(a_t))^2\,[r(a_t)]^2 + \sum_{a\neq a_t}[\pi_\theta(a)]^2\,[r(a_t)]^2$. Taking expectation w.r.t π_θ,

$$\mathbb{E}_{a_t\sim\pi_\theta}\left\|\nabla\tilde{J}(\theta)\right\|^2 = \sum_{a_t}\pi_\theta(a_t)\,\left[(1-\pi_\theta(a_t))^2\,[r(a_t)]^2 + \sum_{a\neq a_t}[\pi_\theta(a)]^2\,[r(a_t)]^2\right]$$

$$\leq \sum_{a_t}\pi_\theta(a_t)\,(1-\pi_\theta(a_t))^2\,[r(a_t)]^2 + \sum_{a_t}\pi_\theta(a_t)\,[r(a_t)]^2\,\left[\sum_{a\neq a_t}\pi_\theta(a)\right]^2 \qquad (\textstyle\sum x_i^2 \leq (\sum x_i)^2)$$

$$= 2\sum_{a_t}\pi_\theta(a_t)\,(1-\pi_\theta(a_t))^2\,[r(a_t)]^2 \leq 2\sum_{a_t}\pi_\theta(a_t)\,(1-\pi_\theta(a_t))^2 \qquad (r(a)\in[0,1])$$

$$\implies \mathbb{E}_{a_t\sim\pi_\theta}\left\|\nabla\tilde{J}(\theta)\right\|^2 \leq 2\sum_{a_t}\pi_\theta(a_t) = 2 \quad \square$$

Hence, we have a bound on the variance of the stochastic gradient estimator.

$$\sigma^2 := \mathbb{E}\left\|\tilde{\nabla}J(\theta) - \mathbb{E}\,[\tilde{\nabla}J(\theta)]\right\|^2 \leq \mathbb{E}\left\|\tilde{\nabla}J(\theta)\right\|^2 \leq 2\,.$$

Similarly, we can construct an unbiased and σ^2-bounded variance stochastic gradient estimator for MDPs [MDX$^+$21, Lemma 11]. We will use these properties to prove convergence to a stationary point.

Stationary point Convergence of Stochastic Softmax Policy Gradient

Claim: Assuming $J(\theta)$ is L-smooth, stochastic softmax PG with an unbiased and σ^2-bounded variance stochastic gradient estimator and step-size $\eta = \min \left\{ {}^{1}\!/_{2L}, {}^{1}\!/_{\sigma\sqrt{T}} \right\}$ converges as:

$$\min_{t \in \{0, \dots T-1\}} \mathbb{E}[\|\nabla J(\theta_t)\|^2] \leq \frac{4L}{(1-\gamma)\,T} + \frac{\sigma\,\left[{}^{2}\!/_{1-\gamma} + L\right]}{\sqrt{T}}\,.$$

Proof: Using smoothness of $J(\theta)$ and the update $\theta_{t+1} = \theta_t + \eta \tilde{\nabla} J(\theta_t)$.

$$J(\theta_{t+1}) \geq J(\theta_t) + \eta\,\langle \nabla J(\theta_t), \tilde{\nabla} J(\theta_t) \rangle - \frac{L\,\eta^2}{2}\,\left\|\tilde{\nabla} J(\theta_t)\right\|^2$$

Taking expectation w.r.t the randomness in iteration t. Since $\mathbb{E}[\tilde{\nabla} J(\theta_t)] = \nabla J(\theta_t)$,

$$\begin{aligned}
\mathbb{E}[J(\theta_{t+1})] &\geq J(\theta_t) + \eta\,\|\nabla J(\theta_t)\|^2 - \frac{L\,\eta^2}{2}\,\mathbb{E}\left[\left\|\tilde{\nabla} J(\theta_t)\right\|^2\right] \\
&= J(\theta_t) + \eta\,\|\nabla J(\theta_t)\|^2 - \frac{L\,\eta^2}{2}\,\mathbb{E}\left[\left\|\tilde{\nabla} J(\theta_t) - \nabla J(\theta_t) + \nabla J(\theta_t)\right\|^2\right] \\
&= J(\theta_t) + \eta\,\|\nabla J(\theta_t)\|^2 - \frac{L\,\eta^2}{2}\,\left[\mathbb{E}[\|\nabla J(\theta_t)\|^2] + \mathbb{E}\left\|\nabla \tilde{J}(\theta_t) - \mathbb{E}[\nabla \tilde{J}(\theta_t)]\right\|^2\right]
\end{aligned}$$

Stationary point Convergence of Stochastic Softmax Policy Gradient

Recall that $\mathbb{E}[J(\theta_{t+1})] \geq J(\theta_t) + \eta \left\|\nabla J(\theta_t)\right\|^2 - \frac{L\eta^2}{2} \left[\mathbb{E}[\left\|\nabla J(\theta_t)\right\|^2] + \mathbb{E}\left\|\nabla \tilde{J}(\theta_t) - \mathbb{E}[\nabla \tilde{J}(\theta_t)]\right\|^2\right]$

$$\mathbb{E}[J(\theta_{t+1})] \geq J(\theta_t) + \eta \left\|\nabla J(\theta_t)\right\|^2 - \frac{L\,\eta^2}{2} \left[\mathbb{E}[\left\|\nabla J(\theta_t)\right\|^2] + \sigma^2\right] \qquad \text{(Def. of } \sigma^2\text{)}$$

Taking expectation w.r.t to the randomness in iterations $t = 0$ to $T - 1$ and summing,

$$\implies \sum_{t=0}^{T-1} \left(\eta - \frac{L\eta^2}{2}\right) \mathbb{E}[\left\|\nabla J(\theta_t)\right\|^2] \leq \sum_{t=0}^{T-1} \mathbb{E}[J(\theta_{t+1}) - J(\theta_t)] + \frac{L\,\eta^2\,\sigma^2\,T}{2}$$

$$\implies \left(\eta - \frac{L\eta^2}{2}\right) \frac{\sum_{t=0}^{T-1} \mathbb{E}[\left\|\nabla J(\theta_t)\right\|^2]}{T} \leq \frac{J(\theta_T) - J(\theta_0)}{T} + \frac{L\eta^2\,\sigma^2}{2} \leq \frac{1}{(1-\gamma)\,T} + \frac{L\eta^2\,\sigma^2}{2}$$

Since $\eta = \min\left\{\frac{1}{2L}, \frac{1}{\sigma\sqrt{T}}\right\}$, $\eta < \frac{1}{L} \implies \left(\eta - \frac{L\eta^2}{2}\right) \geq \frac{\eta}{2}$. Since min is smaller than the average,

$$\min_{t \in \{0, \dots T-1\}} \mathbb{E}[\left\|\nabla J(\theta_t)\right\|^2] \leq \frac{2}{\eta\,(1-\gamma)\,T} + L\eta\,\sigma^2 \leq \frac{\left(4L + 2\sigma\sqrt{T}\right)}{(1-\gamma)\,T} + \frac{L\sigma}{\sqrt{T}} \quad \square$$

(Since $1/\min\{a,b\} = \max\{1/a, 1/b\}$ and $\max\{a, b\} \leq a + b$ for $a, b \geq 0$)

Convergence of Stochastic Softmax Policy Gradient

- We have shown that stochastic softmax PG converges to a stationary point (in expectation) at an $O(1/T + \sigma/\sqrt{T})$ rate.
- We can use the Lojasiewicz condition and prove convergence to the optimal policy at an $O(1/T^{1/4})$ rate. For the bandits case, global convergence to the optimal policy requires that $\min_{t \geq 0} \pi_{\theta_t}(a^*) > 0$. For softmax PG, this property can also be proven in the stochastic case [MZD+23, Theorem 5.1].
- By exploiting non-uniform smoothness, the convergence rate to the optimal policy can be improved to $O(1/\sqrt{T})$ [MDX+21, Theorem 2]. By further exploiting a growth condition on the stochastic gradients, the rate can be improved to $O(1/T)$ [MZD+23, Theorem 5.5].
- The stochastic softmax PG algorithm and the corresponding analysis can be extended to the general multi-armed bandit setting where the rewards are stochastic and sampled from some underlying distribution. The resulting algorithm thus handles exploration in an "automatic" manner and results in an $O(\sqrt{T})$ regret similar to UCB [MZD+23].
- For general MDPs, current results can prove convergence to the optimal policy at an $O(1/\sqrt{T})$ rate [MDX+21, Theorem 13].

Stochastic Natural Policy Gradient

- In the deterministic case, we have shown that NPG converges to the optimal policy at a faster $O(\exp(-T))$ rate. For achieving fast convergence in the stochastic setting, the immediate idea is to use NPG with an importance-weighted reward estimate. For bandits with deterministic rewards, the resulting update is: $\pi_{t+1}(a) = \frac{\pi_t(a)\ \exp(\eta\ \hat{r}_t(a))}{\sum_{a'} \pi_t(a')\ \exp(\eta\ \hat{r}_t(a'))}$.
- For stochastic NPG, $\mathbb{E} \left\|\tilde{\nabla} J(\theta)\right\|^2 = \sum_a \frac{[r(a)]^2}{\pi_\theta(a)}$. Hence, as $\pi_\theta(a) \to 0$ for any action a, the variance becomes unbounded and our previous analysis does not apply.
- In fact, with some non-zero probability, the resulting update does not converge to the optimal policy [MDX+21, Theorem 3] i.e. $\lim_{t\to\infty} \sum_{a\neq a^*} \pi_{\theta_t}(a) \to 1$. Intuitively, the stochastic NPG update is too aggressive and commits to a sub-optimal action early.
- There is a geometry-convergence trade-off in stochastic policy optimization – a "good" algorithm (such as softmax PG, NPG) can only exhibit at most one of the following two behaviours: (i) convergence to the optimal policy with probability 1 at a rate no better than $O(1/T)$ (e.g. a *stable* algorithm like stochastic softmax PG), or (ii) convergence at a rate faster than $O(1/T)$ but failure to converge to the optimal policy with some non-zero probability (e.g. an *aggressive* algorithm like stochastic NPG).

TRPO & PPO

Trust Region Policy Optimization

- Both softmax PG and NPG need to compute the policy gradient for each update to the policy. In scenarios where computing the (approximate) PG is computationally expensive (e.g. involves interaction with a real-world environment or an expensive simulator), these methods can be inefficient.
- PG methods used in practice use the policy gradient to iteratively construct *surrogate functions*, and update the policy parameters in order to maximize these surrogates.
- While forming the surrogate function requires computing the policy gradient, maximizing it and updating the policy parameters does not. Hence, these PG methods can do multiple parameter updates and better re-use the data acquired from the environment.
- Trust Region Policy Optimization (TRPO) is one of the most common PG methods that iteratively constructs such a surrogate function.

Trust Region Policy Optimization

Given a set of feasible policies Π_θ (e.g. those that can be expressed using a model parameterized by θ), TRPO maximizes the following surrogate function (β, δ are parameters) at iteration t:

有theory proof

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_\theta} h_t(\pi) := \left[v^{\pi_t}(\rho) + \frac{\mathbb{E}_{s\sim d^{\pi_t}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma} - \beta \max_s \mathrm{KL}(\pi_t(\cdot|s)||\pi(\cdot|s)) \right] \quad \textbf{(v1)}$$

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_\theta} h_t(\pi) := \left[v^{\pi_t}(\rho) + \frac{\mathbb{E}_{s\sim d^{\pi_t}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\pi_t}(s,a)}{1-\gamma} \right] \quad \text{s.t. } \mathbb{E}_{s\sim d^{\pi_t}} \mathrm{KL}(\pi_t(\cdot|s)||\pi(\cdot|s)) \le \delta \quad \textbf{(v2)}$$

多用这个

- The set Π_θ depends on the policy parameterization, and solving $\max_{\pi\in\Pi_\theta} h_t(\pi)$ by an iterative method such as gradient ascent results in multiple policy updates.
- Theoretical guarantees are proved for **(v1)**, whereas **(v2)** is used in practice (using **(v1)** in practice results in overly conservative updates.)
- β in **(v1)** will be determined theoretically, whereas δ in **(v2)** needs to be tuned empirically.
- Using a linear approximation of $\mathbb{E}_{s\sim d^{\pi_t}} \mathbb{E}_{a\sim\pi(\cdot|s)} \mathfrak{a}^{\pi_t}(s,a)$ and a quadratic approximation of $\mathrm{KL}(\pi_t||\pi)$ leads to a closed-form solution and the resulting update is the same as NPG for the tabular parameterization. (Prove in Assignment 4!).

Trust Region Policy Optimization

- Given exact estimates of the advantage, **(v1)** has monotonic policy improvement guarantees, i.e. $v^{\pi_{t+1}}(\rho) \geq v^{\pi_t}(\rho)$ for all t. Since the function is upper-bounded from above by $\frac{1}{1-\gamma}$, **(v1)** results in convergence to a local maximum.
- Proving monotonic policy improvement relies on the fact that $h_t(\pi)$ is a minorization of $v^{\pi}(\rho)$, i.e. (i) for all π, $v^{\pi}(\rho) \geq h_t(\pi)$ and the inequality is tight at π_t, i.e. (ii) $h_t(\pi_t) = v^{\pi_t}(\rho)$. Given this result, since π_{t+1} is the maximizer of $h_t(\pi)$, (iii) $h_t(\pi_{t+1}) \geq h_t(\pi_t)$. Putting these results together,

$$v^{\pi_{t+1}}(\rho) \overset{(i)}{\geq} h_t(\pi_{t+1}) \overset{(iii)}{\geq} h_t(\pi_t) \overset{(ii)}{=} v^{\pi_t}(\rho)$$

  In order to show monotonic policy improvement for **(v1)**, we now show that $v^{\pi}(\rho) \geq h(\pi)$.

Trust Region Policy Optimization

Claim: For any policies π and $\tilde{\pi}$, $\beta = \frac{4\gamma}{(1-\gamma)^3}$,

$$v^{\pi}(\rho) \geq h(\pi) := \left[v^{\tilde{\pi}}(\rho) + \frac{\mathbb{E}_{s \sim d^{\tilde{\pi}}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \beta \max_s \mathrm{KL}(\tilde{\pi}(\cdot|s) || \pi(\cdot|s)) \right].$$

For iteration t of TRPO, $\tilde{\pi} = \pi_t$ and hence $h(\pi) = h_t(\pi)$.

Proof: The proof relies on the following lemma that bounds the difference in the values of arbitrary policies $\pi, \tilde{\pi}$: $v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)}[\mathfrak{a}^{\tilde{\pi}}(s,a)]$ (Prove in Assignment 4!).

$$\begin{aligned}
v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) \\
&= \frac{1}{1-\gamma} \left[\mathbb{E}_{s \sim d^{\tilde{\pi}}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) + \mathbb{E}_{s \sim d^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) - \mathbb{E}_{s \sim d^{\tilde{\pi}}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) \right] \\
&\qquad \text{(Add/Subtract } \mathbb{E}_{s \sim d^{\tilde{\pi}}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a)) \\
&\geq \frac{1}{1-\gamma} \left[\mathbb{E}_{s \sim d^{\tilde{\pi}}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) - \left| \mathbb{E}_{s \sim d^{\pi}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) - \mathbb{E}_{s \sim d^{\tilde{\pi}}} \mathbb{E}_{a \sim \pi(\cdot|s)} \mathfrak{a}^{\tilde{\pi}}(s,a) \right| \right] \\
&\qquad \text{(Since } x \geq -|x|)
\end{aligned}$$

Trust Region Policy Optimization

$$v^{\pi}(\rho) - v^{\tilde\pi}(\rho) \geq \tfrac{1}{1-\gamma}\ \left[\mathbb{E}_{s\sim d^{\tilde\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a) - |\mathbb{E}_{s\sim d^{\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a) - \mathbb{E}_{s\sim d^{\tilde\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a)|\right].$$

$$v^{\pi}(\rho) - v^{\tilde\pi}(\rho) \geq \frac{\mathbb{E}_{s\sim d^{\tilde\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a)}{1-\gamma} - \frac{|\max_s \left\{\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a)\right\}|\ \left\|d^{\pi} - d^{\tilde\pi}\right\|_1}{1-\gamma}$$

(By Holder's inequality, $|\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]| \leq |\max_x f(x)|\ \|P-Q\|_1$)

$$= \frac{\mathbb{E}_{s\sim d^{\tilde\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a)}{1-\gamma} - \frac{|\max_s \left\{\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a) - \mathbb{E}_{a\sim\tilde\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a)\right\}|\ \left\|d^{\pi} - d^{\tilde\pi}\right\|_1}{1-\gamma}$$

(Since $\mathbb{E}_{a\sim\tilde\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a) = 0$)

$$\geq \frac{\mathbb{E}_{s\sim d^{\tilde\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a)}{1-\gamma} - \frac{\left\|d^{\pi} - d^{\tilde\pi}\right\|_1\ \max_{s,a}|\mathfrak{a}^{\tilde\pi}(s,a)|\ \max_s \|\pi(\cdot|s) - \tilde\pi(\cdot|s)\|_1}{(1-\gamma)}$$

(By Holder's inequality, $|\mathbb{E}_{x\sim P}[f(x)] - \mathbb{E}_{x\sim Q}[f(x)]| \leq |\max_x f(x)|\ \|P-Q\|_1$)

$$\geq \frac{\mathbb{E}_{s\sim d^{\tilde\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde\pi}(s,a)}{1-\gamma} - \frac{2\left\|d^{\pi} - d^{\tilde\pi}\right\|_1\ \max_s \|\pi(\cdot|s) - \tilde\pi(\cdot|s)\|_1}{(1-\gamma)^2}\ (*) \qquad (\mathfrak{a}^{\pi}(s,a) \leq \tfrac{2}{1-\gamma})$$

Next, we will express $\left\|d^{\pi} - d^{\tilde\pi}\right\|_1$ in terms of $\|\pi(\cdot|s) - \tilde\pi(\cdot|s)\|_1$, and combine it with (*).

Trust Region Policy Optimization

$$\mathrm{Pr}^{\pi}(S_\tau = s') - \mathrm{Pr}^{\pi'}(S_\tau = s') = \sum_s \mathrm{Pr}^{\pi}(S_{\tau-1} = s)\,\mathbf{P}_\pi(s, s') - \sum_s \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1} = s)\,\mathbf{P}_{\tilde{\pi}}(s, s')$$

$$= \sum_{s,a} \left[\mathcal{P}(s'|s,a)\,\left[\mathrm{Pr}^{\pi}(S_{\tau-1} = s)\,\pi(a|s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1} = s)\,\tilde{\pi}(a|s)\right]\right]$$

$$= \sum_{s,a} \left[\mathcal{P}(s'|s,a)\,\left[\mathrm{Pr}^{\pi}(S_{\tau-1} = s)\,\left(\pi(a|s) - \tilde{\pi}(a|s)\right) + \tilde{\pi}(a|s)\,\left(\mathrm{Pr}^{\pi}(S_{\tau-1} = s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1} = s)\right)\right]\right]$$

Taking absolute values, using the triangle inequality and summing over s',

$$\begin{aligned}
\implies \sum_{s'} &|\mathrm{Pr}^{\pi}(S_\tau = s') - \mathrm{Pr}^{\pi'}(S_\tau = s')| \\
&\leq \underbrace{\sum_{s'} \sum_{s,a} \mathcal{P}(s'|s,a) |\mathrm{Pr}^{\pi}(S_{\tau-1} = s)\,\left(\pi(a|s) - \tilde{\pi}(a|s)\right)|}_{(i)} \\
&+ \underbrace{\sum_{s'} \sum_{s,a} \mathcal{P}(s'|s,a)\tilde{\pi}(a|s)\,|\mathrm{Pr}^{\pi}(S_{\tau-1} = s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1} = s)|}_{(ii)}
\end{aligned}$$

Trust Region Policy Optimization

$$
\begin{aligned}
(i) &= \sum_{s,a} |\mathrm{Pr}^{\pi}(S_{\tau-1} = s)\ (\pi(a|s) - \tilde{\pi}(a|s))\,| \sum_{s'} \mathcal{P}(s'|s,a) = \sum_{s,a} |\mathrm{Pr}^{\pi}(S_{\tau-1} = s)\ (\pi(a|s) - \tilde{\pi}(a|s))\,| \\
&= \sum_{s} \mathrm{Pr}^{\pi}(S_{\tau-1} = s) \sum_{a} |\,(\pi(a|s) - \tilde{\pi}(a|s))\,| = \sum_{s} \mathrm{Pr}^{\pi}(S_{\tau-1} = s)\ \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1 \\
&\leq \max_{s} \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1 \\
(ii) &= \sum_{s} |\mathrm{Pr}^{\pi}(S_{\tau-1} = s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1} = s)| \sum_{a} \tilde{\pi}(a|s) \sum_{s'} \mathcal{P}(s'|s,a) \\
&= \sum_{s} |\mathrm{Pr}^{\pi}(S_{\tau-1} = s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1} = s)|
\end{aligned}
$$

Hence, $\sum_{s'} |\mathrm{Pr}^{\pi}(S_{\tau} = s') - \mathrm{Pr}^{\pi'}(S_{\tau} = s')| \leq \max_s \|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1$
$+ \sum_{s} |\mathrm{Pr}^{\pi}(S_{\tau-1} = s) - \mathrm{Pr}^{\tilde{\pi}}(S_{\tau-1} = s)|$. By recursing over τ, we get that,

$$
\sum_{s'} |\mathrm{Pr}^{\pi}(S_{\tau} = s') - \mathrm{Pr}^{\pi'}(S_{\tau} = s')| \leq \tau \max_{s} \{\|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1\}
$$

Trust Region Policy Optimization

Recall that $\sum_{s'} \left|\Pr^{\pi}(S_\tau = s') - \Pr^{\pi'}(S_\tau = s')\right| \leq \tau \max_s \left\{\|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1\right\}$.

$$[d^{\pi} - d^{\tilde{\pi}}](s')$$

$$= (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^{\tau} \Pr^{\pi}[S_\tau = s' | S_0 = s_0] - (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^{t} \Pr^{\tilde{\pi}}[S_\tau = s' | S_0 = s_0]$$

$$\left\| d^{\pi} - d^{\tilde{\pi}} \right\|_1 = \sum_{s'} \left| (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^{\tau} \left[\Pr^{\pi}[S_\tau = s' | S_0 = s_0] - \Pr^{\tilde{\pi}}[S_\tau = s' | S_0 = s_0] \right] \right|$$

$$\leq (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{\tau=0}^{\infty} \gamma^{\tau} \sum_{s'} \left|\Pr^{\pi}(S_\tau = s') - \Pr^{\pi'}(S_\tau = s')\right| \qquad \text{(Triangle inequality)}$$

$$\leq (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^{\tau} \tau \max_s \left\{\|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1\right\} = (1-\gamma) \max_s \left\{\|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1\right\} \sum_{\tau=0}^{\infty} \gamma^{\tau} \tau$$

$$\leq \frac{\gamma \max_s \left\{\|\pi(\cdot|s) - \tilde{\pi}(\cdot|s)\|_1\right\}}{1-\gamma} \qquad \left(\text{Since } \sum_{\tau=0}^{\infty} \gamma^{\tau} \tau \leq \tfrac{\gamma}{(1-\gamma)^2}\right)$$

Trust Region Policy Optimization

Recalling inequality (*), $v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) \geq \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{2\left\|d^{\pi}-d^{\tilde{\pi}}\right\|_1 \max_s\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1}{(1-\gamma)^2}$.

We also know that $\left\|d^{\pi} - d^{\tilde{\pi}}\right\|_1 \leq \frac{\gamma\ \max_s\left\{\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1\right\}}{1-\gamma}$. Hence,

$$v^{\pi}(\rho) - v^{\tilde{\pi}}(\rho) \geq \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{2\gamma\ [\max_s\|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1]^2]}{(1-\gamma)^3}$$

$$\implies v^{\pi}(\rho) \geq \left[v^{\tilde{\pi}}(\rho) + \frac{\mathbb{E}_{s\sim d^{\tilde{\pi}}}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathfrak{a}^{\tilde{\pi}}(s,a)}{1-\gamma} - \frac{4\gamma\ \max_s \mathrm{KL}(\tilde{\pi}(\cdot|s)||\pi(\cdot|s))}{(1-\gamma)^3}\right]$$

(By Pinsker's inequality, $2\,\mathrm{KL}(\tilde{\pi}(\cdot|s)||\pi(\cdot|s)) \geq \|\pi(\cdot|s)-\tilde{\pi}(\cdot|s)\|_1^2$)

$$\implies v^{\pi}(\rho) \geq h(\pi) \quad \square$$

- For the tabular policy parameterization, a variant of TRPO that uses $\mathrm{KL}(\pi||\pi_t)$ (instead of $\mathrm{KL}(\pi_t||\pi)$) can be shown to converge to the optimal policy at an $O(1/\sqrt{T})$ rate [SEM20, Theorem 16]. However, the rate still involves the distribution mismatch ratio.

Proximal Policy Optimization

- Proximal Policy Optimization (PPO) is an alternative to TRPO. It is computationally more efficient, typically results in better performance, and is hence widely used in practice.
- PPO maximizes the following surrogate function (ϵ is a parameter) at iteration t:

$$\pi_{t+1} = \arg\max_{\pi \in \Pi_\theta} \left\{ \mathbb{E}_{s \sim d^{\pi_t}} \mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[\mathfrak{a}^{\pi_t}(s,a) \min \left\{ \frac{\pi(a|s)}{\pi_t(a|s)}, \text{clip}\left(\frac{\pi(a|s)}{\pi_t(a|s)}, 1-\epsilon, 1+\epsilon \right) \right\} \right] \right\}$$

  where $\text{clip}(x,a,b) = \min\{\max\{x,a\},b\}$ projects x onto the $[a,b]$ interval.
- Compared to the TRPO surrogate: $\mathbb{E}_{s \sim d^{\pi_t}} \mathbb{E}_{a \sim \pi_t(\cdot|s)} \frac{\pi(a|s)}{\pi_t(a|s)} \mathfrak{a}^{\pi_t}(s,a)$, s.t $\mathbb{E}_{s \sim d^{\pi_t}} \text{KL}(\pi_t(\cdot|s)||\pi(\cdot|s)) \leq \delta$ which ensures that the importance sampling ratio $\frac{\pi(a|s)}{\pi_t(a|s)}$ does not become too large by controlling $\text{KL}(\pi_t(\cdot|s)||\pi(\cdot|s))$, PPO directly ensures that the importance sampling ratio is bounded by clipping it.
- There is no theoretical justification for the clipped PPO surrogate (even with tabular policy parameterization). In fact, PPO can fail on simple problems [HMDH20].
- Recent literature [EIS+20] suggests that code-level implementation details are responsible for most of PPO's gain over TRPO.

References i

- Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry, *Implementation matters in deep policy gradients: A case study on ppo and trpo*, arXiv preprint arXiv:2005.12729 (2020).
- Chloe Ching-Yun Hsu, Celestine Mendler-Dünner, and Moritz Hardt, *Revisiting design choices in proximal policy optimization*, arXiv preprint arXiv:2009.10897 (2020).
- Emmeran Johnson, Ciara Pike-Burke, and Patrick Rebeschini, *Optimal convergence rate for exact policy mirror descent in discounted markov decision processes*, arXiv preprint arXiv:2302.11381 (2023).
- Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans, *Understanding the effect of stochasticity in policy optimization*, Advances in Neural Information Processing Systems **34** (2021), 19339–19351.

References ii

- Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans, *Stochastic gradient succeeds for bandits*.
- Lior Shani, Yonathan Efroni, and Shie Mannor, *Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps*, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 5668–5675.