Computer Engineering Department

Policy-based Theoretical Guarantees

Mohammad Hossein Rohban, Ph.D.

Spring 2025

Courtesy: Most of slides are adopted from the RL course at Berkeley.

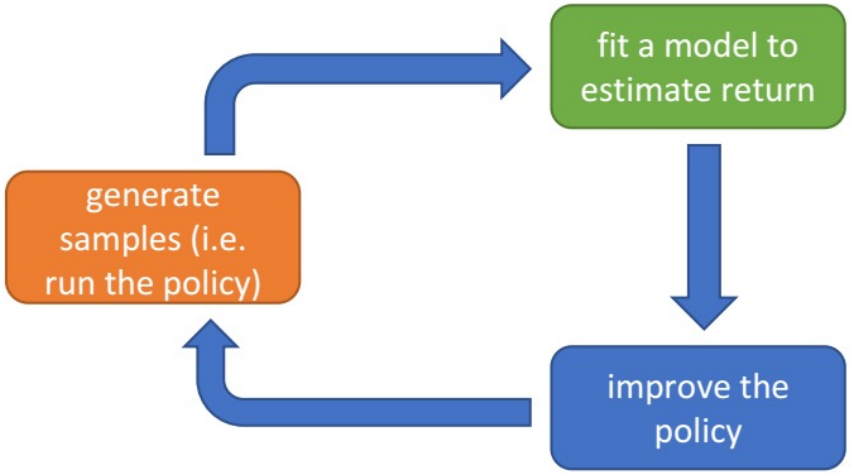
Recap: Policy Gradients

REINFORCE algorithm:

1. sample $\{\tau^i\}$ from $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \sum_i \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i) \left(\sum_{t'=t}^{T} r(\mathbf{s}_{t'}^i, \mathbf{a}_{t'}^i) \right) \right)$
3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\mathbf{a}_{i,t}|\mathbf{s}_{i,t}) \underbrace{\hat{Q}_{i,t}^{\pi}}_{\text{“reward to go”}}$$

$$\hat{Q}^\pi(\mathbf{x}_t, \mathbf{u}_t) = \sum_{t'=t}^{T} r(\mathbf{x}_{t'}, \mathbf{u}_{t'})$$



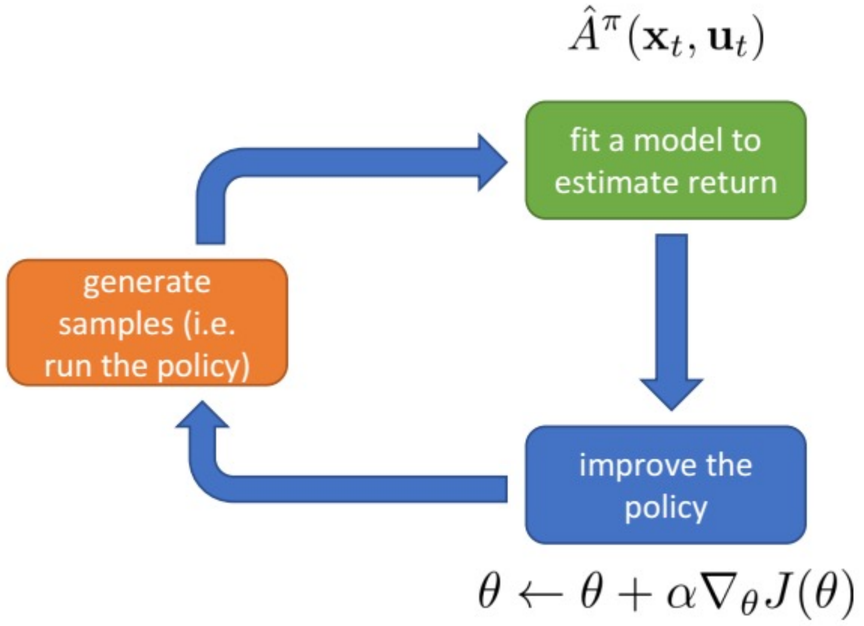
$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

Policy Gradient as Policy Iteration

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(\mathbf{a}_{i,t}|\mathbf{s}_{i,t})\hat{A}^\pi_{i,t}$$

main steps of policy gradient algorithm:

1. Estimate $\hat{A}^\pi(\mathbf{s}_t, \mathbf{a}_t)$ for current policy π
2. Use $\hat{A}^\pi(\mathbf{s}_t, \mathbf{a}_t)$ to get *improved* policy π'



Familiar to policy iteration algorithm:

1. evaluate $A^\pi(\mathbf{s}, \mathbf{a})$ $\rightarrow$ $Q(s,a) - V(s)$
2. set $\pi \leftarrow \pi'$

Policy Gradient as Policy Iteration

$$J(\theta) = E_{\tau \sim p_\theta(\tau)}\left[\sum_t \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right]$$

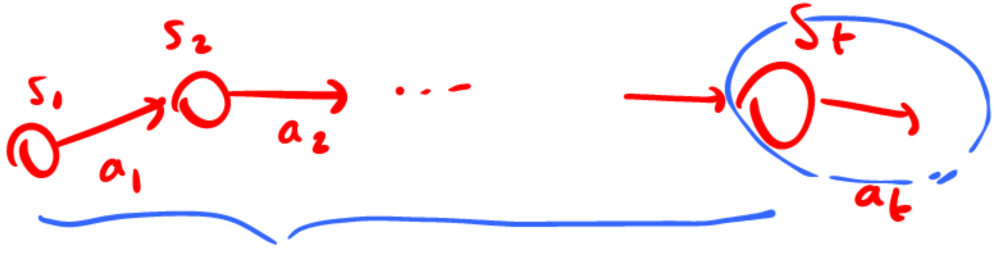
claim: $J(\theta') - J(\theta) = E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right]$

could be interpreted as policy improvement!

Policy Gradient as Policy Iteration

claim: $J(\theta') - J(\theta) = E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right]$

proof:
$$\begin{aligned}
J(\theta') - J(\theta) &= J(\theta') - E_{\mathbf{s}_0 \sim p(\mathbf{s}_0)}\left[V^{\pi_\theta}(\mathbf{s}_0)\right] \\
&= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)}\left[V^{\pi_\theta}(\mathbf{s}_0)\right] \\
&= J(\theta') - E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t V^{\pi_\theta}(\mathbf{s}_t) - \sum_{t=1}^{\infty} \gamma^t V^{\pi_\theta}(\mathbf{s}_t)\right] \\
&= J(\theta') + E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t (\gamma V^{\pi_\theta}(\mathbf{s}_{t+1}) - V^{\pi_\theta}(\mathbf{s}_t))\right] \\
&= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right] + E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t (\gamma V^{\pi_\theta}(\mathbf{s}_{t+1}) - V^{\pi_\theta}(\mathbf{s}_t))\right] \\
&= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t (r(\mathbf{s}_t, \mathbf{a}_t) + \gamma V^{\pi_\theta}(\mathbf{s}_{t+1}) - V^{\pi_\theta}(\mathbf{s}_t))\right] \\
&= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right]
\end{aligned}$$



$P(s_t)$ $\quad$ $\pi(a_t|s_t)$

$= \int P(s_1, a_1, \cdots, s_{t-1}, a_{t-1}, s_t)\, ds_1 da_1 \cdots da_{t-1}$

Policy Gradient as Policy Iteration

$$J(\theta') - J(\theta) = E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right]$$

expectation under $\pi_{\theta'}$

advantage under π_θ

importance sampling

$$E_{x \sim p(x)}[f(x)] = \int p(x) f(x) dx$$
$$= \int \frac{q(x)}{q(x)} p(x) f(x) dx$$
$$= \int q(x) \frac{p(x)}{q(x)} f(x) dx$$
$$= E_{x \sim q(x)}\left[\frac{p(x)}{q(x)} f(x)\right]$$

$$E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right] = \sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)}\left[E_{\mathbf{a}_t \sim \pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}\left[\gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right]\right]$$

$$= \sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)}\left[E_{\mathbf{a}_t \sim \underline{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}}\left[\frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t)\right]\right]$$

is it OK to use $p_\theta(\mathbf{s}_t)$ instead?

Policy Gradient as Policy Iteration

Can we ignore distribution mismatch?

$\pi_\theta \approx \pi_{\theta'}$

$J(\theta') - J(\theta) =$

$$\sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)}\left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}\left[\frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}\gamma^t A^{\pi_\theta}(\mathbf{s}_t,\mathbf{a}_t)\right]\right] \overset{?}{\approx} \underbrace{\sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)}\left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}\left[\frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}\gamma^t A^{\pi_\theta}(\mathbf{s}_t,\mathbf{a}_t)\right]\right]}_{\bar{A}(\theta')}$$

marg.

$\pi(a_t|s_t)$

why do we want this to be true?

$p(s_t)\, p(a_t|s_t) = p(s_t, a_t)$

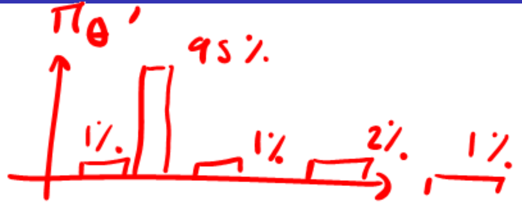
$$J(\theta') - J(\theta) \approx \bar{A}(\theta') \quad \Rightarrow \quad \theta' \leftarrow \arg\max_{\theta'} \bar{A}(\theta)$$

2. Use $\hat{A}^\pi(\mathbf{s}_t, \mathbf{a}_t)$ to get *improved* policy π'

is it true? and when?

$p_\theta(\mathbf{s}_t)$ is *close* to $p_{\theta'}(\mathbf{s}_t)$ when π_θ is *close* to $\pi_{\theta'}$

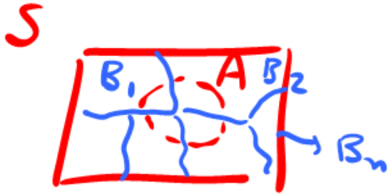
Bounding the distribution change



$\pi_\theta(s) = 3$

Claim: $p_\theta(\mathbf{s}_t)$ is *close* to $p_{\theta'}(\mathbf{s}_t)$ when π_θ is *close* to $\pi_{\theta'}$

Simple case: assume π_θ is a *deterministic* policy $\mathbf{a}_t = \pi_\theta(\mathbf{s}_t)$



$P(A) = \sum_i P(A \cap B_i)$

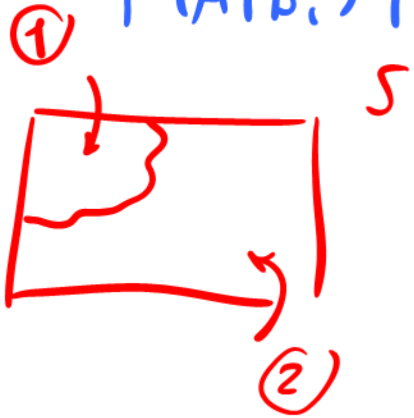
$\pi_{\theta'}$ is *close* to π_θ if $\pi_{\theta'}(\mathbf{a}_t \neq \pi_\theta(\mathbf{s}_t)|\mathbf{s}_t) \overset{=}{\cancel{\leq}} \epsilon$

$p_\theta(s_t) = (1-\epsilon)^t p_\theta(s_t) + (1-(1-\epsilon)^t)\, p_\theta(s_t)$

$$p_{\theta'}(\mathbf{s}_t) = \underbrace{(1-\epsilon)^t}_{\text{probability we made no mistakes}} p_\theta(\mathbf{s}_t) + (1-(1-\epsilon)^t)\underbrace{p_{\text{mistake}}(\mathbf{s}_t)}_{\text{some } other \text{ distribution}}$$

seem familiar?

$P(A|B_i)\, P(B_i)$



$$|p_{\theta'}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| = (1-(1-\epsilon)^t)|p_{\text{mistake}}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| \leq 2(1-(1-\epsilon)^t)$$
$$\leq 2\epsilon t$$

useful identity: $(1-\epsilon)^t \geq 1 - \epsilon t$ for $\epsilon \in [0,1]$

not a great bound, but a bound!

Bounding the distribution change

Claim: $p_\theta(\mathbf{s}_t)$ is *close* to $p_{\theta'}(\mathbf{s}_t)$ when π_θ is *close* to $\pi_{\theta'}$

General case: assume π_θ is an arbitrary distribution

$\pi_{\theta'}$ is *close* to π_θ if $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)| \leq \epsilon$ for all $\mathbf{s}_t$

$$\sum_y p(x,y) = P_X(x)$$

$$P_X(x) \longrightarrow p(x,y) \text{ s.t.}$$

$$P_Y(y) \qquad \sum_x p(x,y) = P_Y(y)$$

$$P(X=Y) = 1-\epsilon$$

Useful lemma: if $|p_X(x) - p_Y(x)| = \epsilon$, exists $p(x,y)$ such that $p(x) = p_X(x)$ and $p(y) = p_Y(y)$ and $p(x=y) = 1-\epsilon$

$\Rightarrow p_X(x)$ "agrees" with $p_Y(y)$ with probability ϵ

$\Rightarrow \pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)$ takes a different action than $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ with probability at most ϵ

$$|p_{\theta'}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| = (1-(1-\epsilon)^t)|p_{\text{mistake}}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| \leq 2(1-(1-\epsilon)^t)$$
$$\leq 2\epsilon t$$

Bounding the objective value

$\pi_{\theta'}$ is *close* to π_θ if $|\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t) - \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)| \leq \epsilon$ for all $\mathbf{s}_t$

$$\sum_s [P_{\theta'}(s) - P_\theta(s)] f(s)$$

$|p_{\theta'}(\mathbf{s}_t) - p_\theta(\mathbf{s}_t)| \leq 2\epsilon t$

$P_{\theta'} + P_\theta - P_\theta$

$$E_{p_{\theta'}(\mathbf{s}_t)}[f(\mathbf{s}_t)] = \sum_{\mathbf{s}_t} p_{\theta'}(\mathbf{s}_t) f(\mathbf{s}_t) \geq \sum_{\mathbf{s}_t} p_\theta(\mathbf{s}_t) f(\mathbf{s}_t) - |p_\theta(\mathbf{s}_t) - p_{\theta'}(\mathbf{s}_t)| \max_{\mathbf{s}_t} f(\mathbf{s}_t)$$

$$\geq E_{p_\theta(\mathbf{s}_t)}[f(\mathbf{s}_t)] - 2\epsilon t \max_{\mathbf{s}_t} f(\mathbf{s}_t)$$

$$\sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)} \left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \left[\frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] \geq$$

$O(Tr_{\max})$ or $O\left(\frac{r_{\max}}{1-\gamma}\right)$

$$\sum_t E_{\mathbf{s}_t \sim p_\theta(\mathbf{s}_t)} \left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \left[\frac{\pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)} \gamma^t A^{\pi_\theta}(\mathbf{s}_t, \mathbf{a}_t) \right] \right] - \sum_t 2\epsilon t C$$

maximizing this maximizes a bound on the thing we want!

Soft actor-critic

1. Q-function update

Update Q-function to evaluate current policy:

$$Q(\mathbf{s},\mathbf{a}) \leftarrow \underline{r(\mathbf{s},\mathbf{a})} + \mathbb{E}_{\mathbf{s}'\sim p_{\mathbf{s}},\ \mathbf{a}'\sim\pi}\left[\underline{Q(\mathbf{s}',\mathbf{a}')} - \underline{\log\pi(\mathbf{a}'|\mathbf{s}')}\right]$$

This converges to Q^{π}.

2. Update policy

Update the policy with gradient of information projection:

$$\pi_{\text{new}} = \arg\min_{\underset{\theta}{\pi'}} \mathrm{D_{KL}}\left(\underset{\text{NN}}{\pi'_{\theta}(\,\cdot\,|\mathbf{s})} \,\middle\|\, \underline{\frac{1}{Z}\exp Q^{\pi_{\text{old}}}(\mathbf{s},\,\cdot\,)}\right)$$

In practice, only take one gradient step on this objective

3. Interact with the world, collect more data

Haarnoja, et al. **Soft Actor-Critic Algorithms and Applications**. ‘18

Soft actor-critic

Algorithm 1 Soft Actor-Critic

Inputs: The learning rates, λ_π, λ_Q, and λ_V for functions π_θ, Q_w, and V_ψ respectively; the weighting factor τ for exponential moving average.

Initialize parameters θ, w, ψ, and $\bar{\psi}$.

for each iteration **do**

    (In practice, a combination of a single environment step and multiple gradient steps is found to work best.)

    for each environment setup **do**

        $a_t \sim \pi_\theta(a_t|s_t)$

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

        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r(s_t, a_t), s_{t+1}\}$

    for each gradient update step **do**

        $\psi \leftarrow \psi - \lambda_V \nabla_\psi J_V(\psi)$.

        $w \leftarrow w - \lambda_Q \nabla_w J_Q(w)$.

        $\theta \leftarrow \theta - \lambda_\pi \nabla_\theta J_\pi(\theta)$.

        $\bar{\psi} \leftarrow \tau\psi + (1 - \tau)\bar{\psi})$.

Loss functions

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\frac{1}{2} \left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\phi} \left[Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\phi(\mathbf{a}_t|\mathbf{s}_t) \right] \right)^2 \right] \tag{5}$$

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t) \right)^2 \right], \tag{7}$$

$$\sum_{s_t \in D} \left[\sum_{a_t} \pi_\phi(a_t | s_t) \underbrace{\log Z_\theta(s_t)} \ \cdots \right]$$

with

$$p(s_{t+1} / s_t, a_t)$$

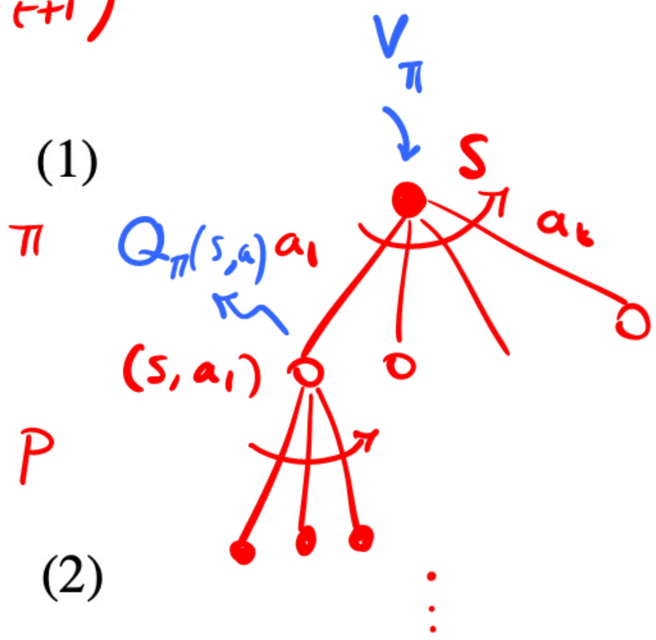
$$\hat{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \, \mathbb{E}_{\mathbf{s}_{t+1} \sim p} \left[V_{\bar{\psi}}(\mathbf{s}_{t+1}) \right], \tag{8}$$

$$J_\pi(\phi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}} \left(\pi_\phi(\cdot|\mathbf{s}_t) \middle\| \frac{\exp(Q_\theta(\mathbf{s}_t, \cdot))}{Z_\theta(\mathbf{s}_t)} \right) \right]. \tag{10}$$

Soft Actor Critic

$\sum_{s_{t+1}, a_{t+1}} \log \pi(a_{t+1} | s_{t+1}) \cdot P(s_{t+1} | s_t, a_t)\, \pi(a_{t+1} | s_{t+1})$

$$J(\pi) = \underbrace{\sum_{t=0}^{T} \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \rho_\pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\,\cdot\,|\mathbf{s}_t)) \right]}. \quad (1)$$



$r'(s_t, a_t)$

$$\mathcal{T}^\pi Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq \underline{r(\mathbf{s}_t, \mathbf{a}_t)} + \underbrace{\gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p} \left[V(\mathbf{s}_{t+1}) \right]}_{\mathbb{E}_{s_{t+1}, \pi} [Q(s_{t+1}, a_{t+1}) - \log \pi(a_{t+1} | s_{t+1})]}, \quad (2)$$

where

$\underline{P(s_{t+1} | s_t, a_t) \cdot \pi(a_{t+1} | s_{t+1})}$

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi} \left[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t | \mathbf{s}_t) \right] \quad (3)$$

$\mathbb{E}_{\underline{s_{t+1}}, \pi} [Q(s_t, a_t) - \overbrace{\underline{\text{const}}}^{g(s_t, a_t)}(s_{t+1}, a_{t+1})$

Soft Policy Evaluation

Lemma 1 (Soft Policy Evaluation)**.** *Consider the soft Bellman backup operator $\mathcal{T}^\pi$ in Equation 2 and a mapping $Q^0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $|\mathcal{A}| < \infty$, and define $Q^{k+1} = \mathcal{T}^\pi Q^k$. Then the sequence Q^k will converge to the soft Q-value of π as $k \to \infty$.*

Soft Policy Evaluation

Lemma 1 (Soft Policy Evaluation). *Consider the soft Bellman backup operator $\mathcal{T}^\pi$ in Equation 2 and a mapping $Q^0 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $|\mathcal{A}| < \infty$, and define $Q^{k+1} = \mathcal{T}^\pi Q^k$. Then the sequence Q^k will converge to the soft Q-value of π as $k \to \infty$.*

Proof. Define the entropy augmented reward as $r_\pi(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \mathbb{E}_{\mathbf{s}_{t+1} \sim p}\left[\mathcal{H}\left(\pi(\,\cdot\,|\mathbf{s}_{t+1})\right)\right]$ and rewrite the update rule as

$$Q(\mathbf{s}_t, \mathbf{a}_t) \leftarrow r_\pi(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1} \sim p, \mathbf{a}_{t+1} \sim \pi}\left[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})\right] \tag{15}$$

and apply the standard convergence results for policy evaluation (Sutton & Barto, 1998). The assumption $|\mathcal{A}| < \infty$ is required to guarantee that the entropy augmented reward is bounded. $\square$

Soft Policy Improvement

$$\pi_{\text{new}} = \arg\min_{\pi' \in \Pi} \mathrm{D}_{\mathrm{KL}} \left(\pi'(\,\cdot\,|\mathbf{s}_t) \;\middle\|\; \frac{\exp\left(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \cdot\,)\right)}{Z^{\pi_{\text{old}}}(\mathbf{s}_t)} \right). \tag{4}$$

Lemma 2 (Soft Policy Improvement). *Let* $\pi_{\text{old}} \in \Pi$ *and let* π_{new} *be the optimizer of the minimization problem defined in Equation 4. Then* $Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t)$ *for all* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$ *with* $|\mathcal{A}| < \infty$.

Soft Policy Improvement

$D_{KL}(p \| q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \sum p(x) \log p(x) - p(x) \log q(x)$

Lemma 2 (Soft Policy Improvement). *Let $\pi_{\text{old}} \in \Pi$ and let π_{new} be the optimizer of the minimization problem defined in Equation 4. Then $Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t)$ for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$ with $|\mathcal{A}| < \infty$.*

Proof. Let $\pi_{\text{old}} \in \Pi$ and let $Q^{\pi_{\text{old}}}$ and $V^{\pi_{\text{old}}}$ be the corresponding soft state-action value and soft state value, and let π_{new} be defined as

$\sum_{a_t} \pi'(a_t|s_t) \log \pi' - \pi' \log \exp(\dots)$

J

$$\pi_{\text{new}}(\cdot|\mathbf{s}_t) = \arg\min_{\pi' \in \Pi} \mathrm{D_{KL}}\left(\pi'(\cdot|\mathbf{s}_t) \,\|\, \exp\left(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \cdot) - \log Z^{\pi_{\text{old}}}(\mathbf{s}_t)\right)\right)$$
$$= \arg\min_{\pi' \in \Pi} J_{\pi_{\text{old}}}(\pi'(\cdot|\mathbf{s}_t)). \quad (16)$$

$\mathbb{E}_{a\sim\pi'}[\log \pi' - Q^{\pi_{old}} + \log Z]$

$\sum \pi' \log \pi' - \pi'[Q^{\pi_{old}} - \log Z]$

It must be the case that $J_{\pi_{\text{old}}}(\pi_{\text{new}}(\cdot|\mathbf{s}_t)) \leq J_{\pi_{\text{old}}}(\pi_{\text{old}}(\cdot|\mathbf{s}_t))$, since we can always choose $\pi_{\text{new}} = \pi_{\text{old}} \in \Pi$. Hence

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{new}}}\left[\log \pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t) - Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + \cancel{\log Z^{\pi_{\text{old}}}(\mathbf{s}_t)}\right] \leq \underbrace{\mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{old}}}\left[\log \pi_{\text{old}}(\mathbf{a}_t|\mathbf{s}_t) - Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t)\right.}_{-V^{\pi_{old}}(s_t)} \left. + \cancel{\log Z^{\pi_{\text{old}}}(\mathbf{s}_t)}\right], \quad (17)$$

Soft Policy Improvement

$\mathbb{E}_{a_{t+1}\sim\pi_{new}}[Q^{\pi_{old}}(s_{t+1},a_{t+1}) - \log\pi_{new}(a_{t+1}|s_{t+1})] \geq V^{\pi_{old}}(s_{t+1})$

$X \leq Y \longrightarrow \mathbb{E}(X) \leq \mathbb{E}(Y)$

and since partition function $Z^{\pi_{\text{old}}}$ depends only on the state, the inequality reduces to

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_t,\mathbf{a}_t) - \log\pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t)\right] \geq V^{\pi_{\text{old}}}(\mathbf{s}_t). \tag{18}$$

Next, consider the soft Bellman equation:

$\pi^{old} \longrightarrow$

$$\begin{aligned}
Q^{\pi_{\text{old}}}(\mathbf{s}_t,\mathbf{a}_t) &= r(\mathbf{s}_t,\mathbf{a}_t) + \gamma\,\mathbb{E}_{\mathbf{s}_{t+1}\sim p}\left[V^{\pi_{\text{old}}}(\mathbf{s}_{t+1})\right] \\
&\leq r(\mathbf{s}_t,\mathbf{a}_t) + \gamma\,\mathbb{E}_{\mathbf{s}_{t+1}\sim p}\left[\mathbb{E}_{\mathbf{a}_{t+1}\sim\pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}(\mathbf{s}_{t+1},\mathbf{a}_{t+1}) - \log\pi_{\text{new}}(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})\right]\right] \\
&\;\;\vdots \\
&\leq Q^{\pi_{\text{new}}}(\mathbf{s}_t,\mathbf{a}_t),
\end{aligned} \tag{19}$$

where we have repeatedly expanded $Q^{\pi_{\text{old}}}$ on the RHS by applying the soft Bellman equation and the bound in Equation 18. Convergence to $Q^{\pi_{\text{new}}}$ follows from Lemma 1. □

Soft Policy Iteration

Theorem 1 (Soft Policy Iteration). *Repeated application of soft policy evaluation and soft policy improvement to any $\pi \in \Pi$ converges to a policy π^* such that $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ for all $\pi \in \Pi$ and $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$, assuming $|\mathcal{A}| < \infty$.*

Proof. Let π_i be the policy at iteration i. By Lemma 2, the sequence Q^{π_i} is monotonically increasing. Since Q^{π} is bounded above for $\pi \in \Pi$ (both the reward and entropy are bounded), the sequence converges to some π^*. We will still need to show that π^* is indeed optimal. At convergence, it must be case that $J_{\pi^*}(\pi^*(\cdot|\mathbf{s}_t)) < J_{\pi^*}(\pi(\cdot|\mathbf{s}_t))$ for all $\pi \in \Pi$, $\pi \neq \pi^*$. Using the same iterative argument as in the proof of Lemma 2, we get $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) > Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ for all $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$, that is, the soft value of any other policy in Π is lower than that of the converged policy. Hence π^* is optimal in Π. $\square$

Policy @ Convergence $\rightarrow \underbrace{\pi^*}_{\pi_{new}} = \arg\min_{\pi' \in \Pi} J_{\underbrace{\pi^*}_{\pi_{old}}}(\pi') \longrightarrow J_{\pi^*}(\pi^*) \leq J_{\pi^*}(\pi)$

$$\underbrace{\mathbb{E}_{\pi^*}\left[Q^{\pi^*}(s_t, a_t) - \log \pi_*(a_t|s_t)\right]}_{V^{\pi_*}(s_t)} \geq \mathbb{E}_{\pi}\left[Q^{\pi^*}(s_t, a_t) - \log \pi\right]$$