Algorithmic and Theoretical Foundations of RL

Policy Optimization II

Ke Wei
School of Data Science
Fudan University

Table of Contents

Gradient Method over Distributions

It is clear that policy optimization for RL is a special case of optimization over probability distributions:

$$\max_{\theta} J(\theta) = \mathbb{E}_{X \sim P_\theta}\left[f(X)\right].$$

The gradient ascent method for this problem is given by

$$\theta^+ = \theta + \eta \cdot \nabla J(\theta),$$

where the search direction $\Delta\theta = \nabla J(\theta)$ satisfies

$$\Delta\theta \propto \operatorname*{argmax}_{\|d\|_2 \le \alpha} \{J(\theta) + \langle \nabla J(\theta), d \rangle\}.$$

Question: Is it more natural to search over probability distribution space since $J(\theta)$ essentially relies on P_θ? YES –> Natural gradient method.

Natural Gradient over Distributions

Natural gradient method conducts search based on KL divergence between probability distributions ($F(\theta)^\dagger$ is pseudoinverse of $F(\theta)$):

$$\Delta\theta \propto \underset{\mathrm{KL}\left(P_\theta \| P_{\theta+d}\right)\le\alpha}{\operatorname{argmax}} \{J(\theta) + \langle \nabla J(\theta), d\rangle\}$$
$$\propto F(\theta)^\dagger \nabla J(\theta),$$

where $F(\theta)$ is the Fisher information matrix at θ, defined by

$$F(\theta) = \mathbb{E}_{X\sim P_\theta}\left[\nabla_\theta \log p_\theta(X)(\nabla_\theta \log p_\theta(X))^T\right].$$

This leads to natural gradient method:

$$\theta^+ = \theta + \eta \cdot F(\theta)^\dagger \nabla J(\theta),$$

which can also be viewed as preconditioned gradient method.

Derivation of Natural Gradient Direction

Given two probability distributions P and Q with pdf $p(x)$ and $q(x)$ respectively, the KL divergence is defined by

$$\mathrm{KL}(P \| Q) = \mathbb{E}_P \left[\log \frac{dP}{dQ} \right] = \mathbb{E}_P \left[\log \frac{p(X)}{q(X)} \right].$$

It follows that

$$\begin{aligned}
\mathrm{KL}(P_\theta \| P_{\theta+d}) &= \mathbb{E}_{P_\theta} \left[\log \frac{p_\theta(X)}{p_{\theta+d}(X)} \right] \\
&= -\mathbb{E}_{P_\theta} \left[\log p_{\theta+d}(X) - \log p_\theta(X) \right] \\
&\approx -d^T \underbrace{\mathbb{E}_{P_\theta} \left[\frac{\nabla_\theta p_\theta(X)}{p_\theta(X)} \right]}_{I_1 = \mathbb{E}_{P_\theta}[\nabla_\theta \log p_\theta(X)]} - \frac{1}{2} d^T \underbrace{\mathbb{E}_{P_\theta} \left[\frac{\nabla_\theta^2 p_\theta(X)}{p_\theta(X)} - \frac{\nabla_\theta p_\theta(X)(\nabla_\theta p_\theta(X))^T}{p_\theta(X)^2} \right]}_{I_2 = \mathbb{E}_{P_\theta}\left[\nabla_\theta^2 \log p_\theta(X)\right]} d.
\end{aligned}$$

Derivation of Natural Gradient Direction

For I_1, one has

$$\mathbb{E}_{P_\theta}\left[\frac{\nabla_\theta p_\theta(X)}{p_\theta(X)}\right] = \int \nabla_\theta p_\theta(X)dx = 0.$$

For I_2, one has

$$\mathbb{E}_{P_\theta}\left[\frac{\nabla_\theta^2 p_\theta(X)}{p_\theta(X)}\right] = \int \nabla_\theta^2 p_\theta(X)dx = 0$$

and

$$\mathbb{E}_{P_\theta}\left[\frac{\nabla_\theta p_\theta(X)(\nabla_\theta p_\theta(X))^T}{p_\theta(X)^2}\right] = \mathbb{E}_{P_\theta}\left[\nabla_\theta \log p_\theta(X)(\nabla_\theta \log p_\theta(X))^T\right] = F(\theta).$$

It follows that

$$\Delta\theta = \underset{\mathrm{KL}(P_\theta \| P_{\theta+d}) \le 2\alpha}{\operatorname{argmax}} \{J(\theta) + \langle \nabla J(\theta), d\rangle\} \approx \underset{d^T F(\theta) d \le 2\alpha}{\operatorname{argmax}} \{J(\theta) + \langle \nabla J(\theta), d\rangle\} \propto F(\theta)^\dagger \nabla J(\theta).$$

The pseudoinverse basically means that we won't consider the direction such $F(\theta)d = 0$ since in this case one has $\mathrm{KL}\left(P_\theta \| P_{\theta+d}\right) \approx d^T F(\theta) d = 0$ and the objective function roughly remains unchanged.

Natural Policy Gradient (NPG)

Natural policy gradient is natural gradient applied to RL optimization problem:

$$\max_{\theta} V^{\pi_\theta}(\mu) = \mathbb{E}_{s_0\sim\mu}\left[V^{\pi_\theta}(s_0)\right] = \mathbb{E}_{\tau\sim P_\mu^{\pi_\theta}}\left[r(\tau)\right],$$

where given $\tau = (s_t, a_t, r_t)_{t=0}^{\infty}$,

$$P_\mu^{\pi_\theta}(\tau) = \mu(s_0)\prod_{t=0}^{\infty}\pi_\theta(a_t|s_t)P(s_{t+1}|s_t,a_t) \quad \text{and} \quad r(\tau) = \sum_{t=0}^{\infty}\gamma^t r_t.$$

Natural gradient search direction can be incorporated into different policy optimization methods (including REINFORCE, actor-critic) after MC evaluation of $F(\theta)$ (e.g., using data from an episode). We only focus on expression for $F(\theta)$.

By the definition of $F(\theta)$ and expression for $P_\mu^{\pi_\theta}$ (assuming $\pi_\theta(a|s) = 1$ for any θ),

$$\begin{aligned}
F(\theta) =& \mathbb{E}_{\tau\sim P_\mu^{\pi_\theta}}\left[\left(\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t|s_t)\right)\left(\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t|s_t)\right)^{\mathsf{T}}\right] \\
=& \mathbb{E}_{\tau\sim P_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty}\nabla_\theta \log\pi_\theta(a_t|s_t)(\nabla_\theta \log\pi_\theta(a_t|s_t))^{\mathsf{T}}\right].
\end{aligned}$$

Two Common Expressions of $F(\theta)$ to Avoid Divergence

- Average case:

$$F(\theta) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_{\tau \sim P_\mu^{\pi_\theta}} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \left(\nabla_\theta \log \pi_\theta(a_t|s_t) \right)^T \right]$$
$$= \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \left(\nabla_\theta \log \pi_\theta(a|s) \right)^T \right],$$

where $d^{\pi_\theta}(s) = \mathbb{E}_{s_0 \sim \mu} \left[\lim_{t \to \infty} P(s_t = s | s_0, \pi_\theta) \right]$ is state stationary distribution.

- Discounted case:

$$F(\theta) = (1-\gamma) \mathbb{E}_{\tau \sim P_\mu^{\pi_\theta}} \left[\sum_{t=0}^{+\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) (\nabla_\theta \log \pi_\theta(a_t|s_t))^T \right]$$
$$= \mathbb{E}_{s \sim d_\mu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) \left(\nabla_\theta \log \pi_\theta(a|s) \right)^T \right],$$

where $d_\mu^{\pi_\theta}(s) = \mathbb{E}_{s_0 \sim \mu} \left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s | s_0, \pi_\theta) \right]$ is discounted state visitation measure.

Remark

- For the discounted case, it is not difficult to verify that the natural gradient direction $F(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\mu)$ satisfies

$$F(\theta)^\dagger \nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma}\omega^*,$$

where ω^* is the (ℓ_2-minimal) solution to

$$\min_\omega L(\omega) = \mathbb{E}_{s\sim d_\mu^{\pi_\theta}, a\sim\pi_\theta(\cdot|s)}\left[\left((\nabla_\theta \log \pi_\theta(a|s))^T \omega - A^{\pi_\theta}(s,a)\right)^2\right].$$

See "On the theory of policy gradient methods: Optimality, approximation, and distribution shift" by Agarwal et al. 2021 for details.

Remark

- For the softmax parameterization (i.e., $\pi_\theta(a|s) = \exp(\theta_{s,a})/(\sum_{a'} \exp(\theta_{s,a'}))$), it can be verified all the solutions to $\min_\omega L(\omega)$ has the following general form:

$$\omega^*_{s,a} = A^{\pi_\theta}(s,a) + c_s,$$

where c_s is a constant relying on s. Thus NPG in policy space is given by

$$\pi^+_{s,a} = \frac{\pi_{s,a} \cdot \exp\left(\frac{\eta}{1-\gamma} A^{\pi_\theta}(s,a)\right)}{\sum_{a'} \pi_{s,a'} \cdot \exp\left(\frac{\eta}{1-\gamma} A^{\pi_\theta}(s,a')\right)},$$

which coincides with EQA in Lecture 7 (a policy mirror ascent method).

See "On the theory of policy gradient methods: Optimality, approximation, and distribution shift" by Agarwal et al. 2021 for details.

Table of Contents

Trust Region Policy Optimization (TRPO)

Overall Idea

Given a policy π_{θ_t}, by performance difference lemma, we can rewrite $V^{\pi_\theta}(\mu)$ as

$$V^{\pi_\theta}(\mu) = V^{\pi_{\theta_t}}(\mu) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_{\theta_t}}(s,a)\right].$$

Since we do not have access to $d_\mu^{\pi_\theta}$, instead maximize the approximation:

$$\max_\theta\ V_t(\theta) = V^{\pi_{\theta_t}}(\mu) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^{\pi_{\theta_t}}(s,a)\right].$$

Trust Region Policy Optimization (TRPO)

Two Facts

- Assume $\sum_a \pi_\theta(a|s) = 1$ for any θ. It is easy to see that $V^{\pi_\theta}(\mu)$ and $V_t(\theta)$ match at θ_t up to first derivative.
- It can be shown that

$$V^{\pi_\theta}(\mu) \geq V_t(\theta) - \frac{2\gamma\varepsilon_t}{(1-\gamma)^2} \max_s \mathrm{KL}(\pi_{\theta_t}(\cdot|s) \| \pi_\theta(\cdot|s)),$$

  where $\varepsilon_t = \max_{s,a} |A^{\pi_{\theta_t}}(s,a)|$.

See "Trust region policy optimization" by Schulman et al. 2017 for derivation of second fact.

Trust Region Policy Optimization (TRPO)

TRPO is Approximately NPG Plus Line Search

The second fact suggests that we may seek a new estimator by maximizing $V_t(\theta)$ in a small neighborhood of θ_t:

$$\max_{\theta} \; V_t(\theta) \quad \text{subject to} \quad \max_{s} \mathrm{KL}(\pi_{\theta_t}(\cdot|s) \| \pi_{\theta}(\cdot|s)) \leq \delta.$$

Moreover, replace constraint by the average version and instead solve

$$\max_{\theta} \; V_t(\theta) \quad \text{subject to} \quad \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta_t}}} \left[\mathrm{KL}(\pi_{\theta_t}(\cdot|s) \| \pi_{\theta}(\cdot|s))\right] \leq \delta.$$

Trust Region Policy Optimization (TRPO)

TRPO is Approximately NPG Plus Line Search

After linear approximation to $V_t(\theta)$ and quadratic approximation to KL at θ_t,

$$V_t(\theta) \approx (\nabla_\theta V^{\pi_{\theta_t}}(\mu))^T(\theta - \theta_t),\ \mathbb{E}_{s\sim d_\mu^{\pi_{\theta_t}}}\left[\mathrm{KL}(\pi_{\theta_t}(\cdot|s)\|\pi_\theta(\cdot|s))\right] \approx \frac{1}{2}(\theta - \theta_t)^T F(\theta_t)(\theta - \theta_t),$$

we arrive at the same problem as that for NPG,

$$\max_\theta (\nabla_\theta V^{\pi_{\theta_t}}(\mu))^T(\theta - \theta_t) \quad \text{subject to} \quad \frac{1}{2}(\theta - \theta_t)^T F(\theta_t)(\theta - \theta_t) \le \delta.$$

- TRPO is NPG with adaptive line search in implementations.

Table of Contents

Proximal Policy Optimization (PPO)

Recall from last section that

$$\begin{aligned} V_t(\theta) &\propto \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_{\theta_t}}(s,a) \right] \\ &= \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a \sim \pi_{\theta_t}(\cdot|s)} \left[\frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)} A^{\pi_{\theta_t}}(s,a) \right], \end{aligned}$$

serves as a surrogate function of true target in small region around θ_t.

PPO keeps new policy close to old one through clipped objective.

PPO with Clipped Objective

Let $r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_t}(a|s)}$. Then $r(\theta_t) = 1$. The clipped objective function is given by

$$V_t^{\text{clip}}(\theta) = \mathbb{E}_{s \sim d_\mu^{\pi_{\theta_t}}} \mathbb{E}_{a \sim \pi_{\theta_t}(\cdot|s)} \left[\min\left(r(\theta) A^{\pi_{\theta_t}}(s,a), \operatorname{clip}\left(r(\theta), 1-\epsilon, 1+\epsilon\right) A^{\pi_{\theta_t}}(s,a) \right) \right],$$

where

$$\operatorname{clip}\left(r(\theta), 1-\epsilon, 1+\epsilon\right) = \begin{cases} 1+\epsilon, & r(\theta) > 1+\epsilon, \\ r(\theta), & r(\theta) \in [1-\epsilon, 1+\epsilon], \\ 1-\epsilon, & r(\theta) < 1-\epsilon. \end{cases}$$

- The $\min$ operation ensure $V_t^{\text{clip}}(\theta)$ provides a lower bound. Since a maximal point will be computed subsequently, $\min$ will not cancel the effect of clip.
- PPO policy update (in expectation): $\theta_{t+1} = \operatorname{argmax}_\theta V_t^{\text{clip}}(\theta)$.
- In flat region, gradient of $V_t^{\text{clip}}(\theta)$ is zero, thus won't move far from θ_t is using policy gradient type method to solve the sub-problem.

See "Proximal policy optimization algorithms" by Schulman et al. 2017 for details.

Table of Contents

Deterministic Policy Parameterization

Consider the case where $\mathcal{S}$ and $\mathcal{A}$ are continuous, and use π_θ to denote a deterministic policy: $a = \pi_\theta(s)$ is an action.

- Average state value:

$$V^{\pi_\theta}(\mu) = \int_{\mathcal{S}} V^{\pi_\theta}(s_0)\mu(s_0)\mathrm{d}s_0 = \mathbb{E}_{\tau \sim p_\mu^{\pi_\theta}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi_\theta(s_t), s_{t+1})\right],$$

  where given trajectory $\tau = (s_t, \pi_\theta(s_t), s_{t+1})_{t=0}^{\infty}$,

$$p_\mu^{\pi_\theta}(\tau) = \mu(s_0)\prod_{t=0}^{\infty} p(s_{t+1}|s_t, \pi_\theta(s_t))$$

  is the probability density over τ. Note that there is no probability over action space since $\pi_\theta(s)$ selects a deterministic action.

Deterministic Policy Parameterization

- Similarly, we can express $V^{\pi_\theta}(\mu)$ over state space

$$
\begin{aligned}
V^{\pi_\theta}(\mu) &= \frac{1}{1-\gamma}\int_{\mathcal{S}} d_\mu^{\pi_\theta}(s)\mathrm{d}s \int_{\mathcal{S}} p(s'|s,\pi_\theta(s))r(s,\pi_\theta(s),s')\mathrm{d}s' \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_\theta}}\mathbb{E}_{s'\sim p(\cdot|s,\pi_\theta(s))}\left[r(s,\pi_\theta(s),s')\right],
\end{aligned}
$$

where $d_\mu^{\pi_\theta}(s) = \mathbb{E}_{s_0\sim\mu}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_t(s|s_0,\pi_\theta)\right]$ is state visitation density, and $p_t(s|s_0,\pi_\theta)$ is the density over state space after transitioning t time steps. Note there is no expectation over action space since $\pi_\theta(s)$ is deterministic.

Deterministic Policy Gradient Theorem

Theorem 1 (Deterministic Policy Gradient Theorem)

Suppose that $\nabla_\theta \pi_\theta(s)$ *and* $\nabla_a Q^{\pi_\theta}(s, a)$ *exist. Then,*

$$\nabla_\theta V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi_\theta}(s)} \left[\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s, a)|_{a=\pi_\theta(s)} \right].$$

Proof of Theorem 1

First note that

$$\begin{aligned} V^{\pi_\theta}(s_0) &= Q^{\pi_\theta}(s_0, \pi_\theta(s_0)) \\ &= \int_{\mathcal{S}} \big(r(s_0, \pi_\theta(s_0), s_1) + \gamma V^{\pi_\theta}(s_1)\big) p(s_1|s_0, \pi_\theta(s_0)) \mathrm{d}s_1. \end{aligned}$$

Therefore, one has

$$\begin{aligned} \nabla_\theta V^{\pi_\theta}(s_0) &= \int_{\mathcal{S}} \nabla_a r(s_0, a, s_0)|_{a=\pi_\theta(s_0)} \nabla_\theta \pi_\theta(s_0) p(s_1|s_0, \pi_\theta(s_0)) \mathrm{d}s_1 \\ &+ \int_{\mathcal{S}} r(s_0, \pi_\theta(s_0), s_1)\ \nabla p(s_1|s_0, a)|_{a=\pi_\theta(s_0)} \nabla_\theta \pi_\theta(s_0) \mathrm{d}s_1 \\ &+ \gamma \int_{\mathcal{S}} V^{\pi_\theta}(s_1)\ \nabla p(s_1|s_0, a)|_{a=\pi_\theta(s_0)} \nabla_\theta \pi_\theta(s_0) \mathrm{d}s_1 \\ &+ \gamma \int_{\mathcal{S}} \nabla_\theta V^{\pi_\theta}(s_1) p(s_1|s_0, \pi_\theta(s_0)) \mathrm{d}s_1. \end{aligned}$$

Proof of Theorem 1 (Cont'd)

Moreover, it is easy to verify that the sum of the first three terms is equal to

$$\nabla_\theta \pi_\theta(s_0)\ \nabla_a Q^{\pi_\theta}(s,a)\big|_{a=\pi_\theta(s_0)}.$$

Therefore,

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta}(s_0) &= \nabla_\theta \pi_\theta(s_0)\ \nabla_a Q^{\pi_\theta}(s,a)\big|_{a=\pi_\theta(s_0)} + \gamma \int_{\mathcal{S}} \nabla_\theta V^{\pi_\theta}(s_1) p(s_1|s_0,\pi_\theta(s_0))\mathrm{d}s_1 \\
&= \ldots \\
&= \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \pi_\theta(s_t) \nabla_a Q^{\pi_\theta}(s_t,a)|_{a=\pi_\theta(s_t)} | s_0, \pi_\theta\Big] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s\sim d_{s_0}^{\pi_\theta}} \left[\nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s,a)|_{a=\pi_\theta(s)}\right].
\end{aligned}$$

Averaging over all s_0 completes the proof of Theorem 1.

Deep Deterministic Policy Gradient (DDPG)

- DDPG is a policy gradient method which learns a deterministic policy π_θ and an action value function $Q^\omega(s, a) \approx Q^{\pi_\theta}(s, a)$. It is an actor-critic algorithm.
- Policy of DDPG is deterministic, need to add random noisy when collecting data; experience replay buffer is also used to break statistical dependence.
- Update of ω for action value function is overall the same to Fitted Q-learning.

See "Continuous control with deep reinforcement learning" by Lillicrap et al. 2016 for details.

Table of Contents

Motivation: Enhance exploration by entropy regularization

Given a policy π, entropy regularized objective function is define by

$$\begin{aligned} V_\lambda^\pi(\mu) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^\pi}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s')\right] + \lambda H(\pi(\cdot|s))\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^\pi}\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') - \lambda\log\pi(a|s)\right], \end{aligned}$$

where $H(\pi(\cdot|s))$ denotes the entropy of the probability distribution $\pi(\cdot|s)$:

$$H(\pi(\cdot|s)) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\log\frac{1}{\pi(a|s)}\right].$$

We can rewrite $V_\lambda^\pi(\mu)$ in terms of state values based on a regularized reward

$$V_\lambda^\pi(\mu) = \mathbb{E}_{s\sim\mu}\left[V_\lambda^\pi(s)\right],$$

where $V_\lambda^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_\lambda(s_t,a_t,s_{t+1})|s_0 = s,\pi\right]$ with

$$r_\lambda(s,a,s') = r(s,a,s') - \lambda\log\pi(a|s).$$

- Note that $r_\lambda(s,a,s')$ is not a fixed reward but varies from π to π.

Soft Bellman Equation

- Soft state value V_λ^π :

$$V_\lambda^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_\lambda(s_t, a_t, s_{t+1}) | s_0 = s, \pi\right].$$

- Soft action value $Q_\lambda^\pi(s, a)$: [a_0 is chosen, thus entropy equal to 0]

$$Q_\lambda^\pi(s, a) = \mathbb{E}\left[r(s_0, a_0, s_1) + \sum_{t=1}^{\infty} \gamma^t r_\lambda(s_t, a_t, s_{t+1}) | s_0 = s, a_0 = a, \pi\right].$$

- Relation between Q_λ^π and V_λ^π:

$$Q_\lambda^\pi(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}[r(s, a, s') + \gamma V_\lambda^\pi(s')],$$
$$V_\lambda^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[-\lambda \log \pi(a|s) + Q_\lambda^\pi(s, a)].$$

- Soft Bellman equation:

$$V_\lambda^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r_\lambda(s, a, s') + \gamma V_\lambda^\pi(s')\right],$$
$$Q_\lambda^\pi(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[r(s, a, s') + \gamma \mathbb{E}_{a' \sim \pi(\cdot|s')}[Q_\lambda^\pi(s', a') - \lambda \log \pi(a'|s')]\right].$$

Soft Bellman Operator

- For state value, soft Bellman operator $\mathcal{T}_\lambda^\pi$ under a policy π is defined by

$$[\mathcal{T}_\lambda^\pi V_\lambda](s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim P(\cdot|s,a)}[r_\lambda(s,a,s') + \gamma V_\lambda(s')].$$

  - $\mathcal{T}_\lambda^\pi$ is γ-contraction with respect to ℓ_∞-norm and V_λ^π is unique fixed point.
- For action value, soft Bellman operator $\mathcal{F}_\lambda^\pi$ under a policy π is defined by

$$[\mathcal{F}_\lambda^\pi Q_\lambda](s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}\left[Q_\lambda(s',a') - \lambda\log\pi(a'|s')\right]\right],$$

  - $\mathcal{F}_\lambda^\pi$ is γ-contraction with respect to ℓ_∞-norm and Q_λ^π is unique fixed point.

Soft Bellman Optimality Equation: State Value

For any $V_\lambda \in \mathbb{R}^{|\mathcal{S}|}$, the soft Bellman optimality operator $\mathcal{T}_\lambda$ is defined by

$$
\begin{aligned}
[\mathcal{T}_\lambda V_\lambda](s) &= \max_\pi \mathbb{E}_{a\sim\pi(\cdot|s)} \mathbb{E}_{s'\sim P(\cdot|s,a)} [r_\lambda(s,a,s') + \gamma V(s')] \\
&= \max_\pi \mathbb{E}_{a\sim\pi(\cdot|s)} \Big[\underbrace{\mathbb{E}_{s'\sim P(\cdot|s,a)} [r(s,a,s') + \gamma V(s')]}_{:=Q_\lambda(s,a)} - \lambda \log \pi(a|s) \Big] \\
&= \lambda \log \left(\left\| \exp\left(Q_\lambda(s,\cdot)/\lambda \right) \right\|_1 \right),
\end{aligned}
$$

where maximum value is attained at (i.e, extracted policy)

$$
\begin{aligned}
\pi_\lambda(a|s) &= \frac{\exp(Q_\lambda(s,a)/\lambda)}{\left\| \exp\left(Q_\lambda(s,\cdot)/\lambda \right) \right\|_1} \\
&= \frac{\exp(Q_\lambda(s,a)/\lambda)}{\exp([\mathcal{T}_\lambda V_\lambda](s)/\lambda)}.
\end{aligned}
$$

following Lemma 5 of Lecture 7.

Remark

- Entropy regularization moves the maxima to the interior so that it has an explicit solution in terms of softmax representation.
- Also by Lemma 5 of Lecture 7, one has for any $a \neq a'$,

$$Q_\lambda(s,a) - \lambda \log \pi_\lambda(a|s) = Q_\lambda(s,a') - \lambda \log \pi_\lambda(a'|s)$$

  at optimal π (adding entropy tends to average something). Thus,

$$[\mathcal{T}_\lambda V_\lambda](s) = Q_\lambda(s,a) - \lambda \log \pi_\lambda(a|s), \quad \forall a.$$

- $\mathcal{T}_\lambda$ is γ-contraction with respect to ℓ_∞-norm.

Soft Bellman Optimality Equation: Action Value

For $Q_\lambda \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, the soft Bellman optimality operator $\mathcal{F}_\lambda$ is defined by

$$
\begin{aligned}
[\mathcal{F}_\lambda Q_\lambda](s,a) &= \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma \max_{\pi} \mathbb{E}_{a'\sim\pi(\cdot|s')}\left[Q_\lambda(s',a') - \lambda\log\pi(a'|s')\right]\right] \\
&= \mathbb{E}_{s'\sim P(\cdot|s,a)}\left[r(s,a,s') + \gamma\left[\lambda\log\left(\left\|\exp\left(Q_\lambda\left(s',\cdot\right)/\lambda\right)\right\|_1\right)\right]\right],
\end{aligned}
$$

where maximum value is attained at $\pi_\lambda(\cdot|s') \propto \exp(Q_\lambda(s',\cdot)/\lambda)$.

- $\mathcal{F}_\lambda$ is γ-contraction with respect to ℓ_∞-norm.

Optimal Policy

Theorem 2
Let V_λ^* and Q_λ^* be the fixed points of $\mathcal{T}_\lambda$ and $\mathcal{F}_\lambda$, respectively. One has

$$V_\lambda^*(s) = \max_\pi V_\lambda^\pi(s), \ \forall s \quad \textit{and} \quad Q_\lambda^*(s,a) = \max_\pi Q_\lambda^\pi(s,a), \ \forall s,a.$$

The equality is achieved by the optimal policy given by

$$\pi_\lambda^*(a|s) = \frac{\exp(Q_\lambda^*(s,a)/\lambda)}{\|\exp(Q_\lambda^*(s,\cdot)/\lambda)\|_1}.$$

Moreover, $Q_\lambda^*(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a)}[r(s,a,s') + \gamma V_\lambda^*(s')]$ *and*

$$V_\lambda^* = \lambda \log\left(\left\|\exp\left(Q_\lambda^*\left(s,\cdot\right)/\lambda\right)\right\|_1\right) = Q_\lambda^*(s,a) - \lambda \log \pi_\lambda^*(a|s), \quad \forall a.$$

See “Bridging the gap between value and policy based reinforcement learning” by Nachum et al. 2017 for details.

Remark

- Theorem 2 implies that optimal policy is unique with entropy regularlization.
- It is evident that as $\lambda \to 0$, $\pi_\lambda^*(a|s) \to 0$ for $a \notin \operatorname{argmax} Q^*(s,a)$.
- Since one has

$$\max_a Q_\lambda^*(s,a) \le \lambda \log \left(\left\| \exp \left(Q_\lambda^* \left(s, \cdot \right) / \lambda \right) \right\|_1 \right) \le \lambda \log |\mathcal{A}| + \max_a Q_\lambda^*(s,a),$$

  it is easy to see that $V_\lambda^*(s) \to \max_a Q^*(s,a) = V^*(s)$ as $\lambda \to 0$.

Soft Policy Iteration

- Soft policy evaluation:

$$Q_\lambda^{\pi_k} = \mathcal{F}_\lambda^{\pi} Q_\lambda^{\pi_k} = \mathbb{E}_{s' \sim P(\cdot|s,a)}[r(s,a,s') + \gamma V_\lambda^{\pi_k}(s')].$$

- Soft policy improvement (soft greedy, use softmax to approximate max):

$$\pi_{k+1} = \frac{\exp(Q_\lambda^{\pi_k}(s,\cdot)/\lambda)}{\|\exp(Q_\lambda^{\pi_k}(s,\cdot)/\lambda\|_1}.$$

Theorem 3 (Informal)
It can be shown that π_{k+1} is an improved policy compared to π_k and the γ-rate convergence of soft PI can also be established.

See "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor" by Haarnoja et al. 2018 for details.

Soft Actor Critic (SAC)

SAC is a policy based or actor-critic method for solving

$$\max_{\theta} V_{\lambda}^{\pi_\theta}(\mu) = \mathbb{E}_{s\sim\mu}\left[V_{\lambda}^{\pi_\theta}(s)\right].$$

In addition to typical ways for updating value function and policy parameters,

- Reparametrizarion trick is used in the computation of policy gradient;
- Both state and action values have been parametrized for stable training.

See "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor" by Haarnoja et al. 2018 for details.

Questions?