6) Conservative Policy Iteration

Melih Kandemir

University of Southern Denmark
Department of Mathematics and Computer Science (IMADA)

Why do deep actor-critics work?

Theorem

Let V be a value function that satisfies $\|V - V^*\|_\infty = \varepsilon$ for some $\varepsilon > 0$ and the optimal value of some policy π_*. For a greedy policy $\widehat{\pi}$ with respect to V, we have

$$\|V^{\widehat{\pi}} - V^*\|_\infty \leq \frac{2\gamma\varepsilon}{1-\gamma}.$$

Furthermore, there exists some $\varepsilon_0 > 0$ such that if $\varepsilon < \varepsilon_0$ then $\widehat{\pi} = \pi_*$.

If we guarantee policy improvement at each training iteration, i.e. $V^{\widehat{\pi}_{k+1}} \geq V^{\widehat{\pi}_k}$, then there will be a moment where our critic will get so close to the optimal value function that the corresponding actor will be optimal.

Proof

$$
\begin{aligned}
\|V^{\widehat{\pi}} - V^*\|_\infty &= \|T^{\widehat{\pi}} V^{\widehat{\pi}} - V^*\|_\infty \\
&= \|T^{\widehat{\pi}} V^{\widehat{\pi}} - T^{\widehat{\pi}} V + T^{\widehat{\pi}} V - V^*\|_\infty \\
&\leq \|T^{\widehat{\pi}} V^{\widehat{\pi}} - T^{\widehat{\pi}} V\|_\infty + \|T^{\widehat{\pi}} V - V^*\|_\infty \\
&\leq \gamma \|V^{\widehat{\pi}} - V\|_\infty + \|T^{\widehat{\pi}} V - V^*\|_\infty \\
&= \gamma \|V^{\widehat{\pi}} - V\|_\infty + \|T^* V - V^*\|_\infty \\
&\leq \gamma \|V^{\widehat{\pi}} - V\|_\infty + \gamma \|V - V^*\|_\infty \\
&= \gamma \|V^{\widehat{\pi}} - V^* + V^* - V\|_\infty + \gamma \|V - V^*\|_\infty \\
&\leq \gamma \|V^{\widehat{\pi}} - V^*\|_\infty + \gamma \|V^* - V\|_\infty + \gamma \|V - V^*\|_\infty \\
&= \gamma \|V^{\widehat{\pi}} - V^*\|_\infty + 2\gamma\varepsilon
\end{aligned}
$$

Let $\delta = \min_{\pi \in \bar{\Pi}} \|V_\pi - V^*\|_\infty$ where $\bar{\Pi} := \Pi \backslash \{\pi | V_\pi = V_{\pi_*}\}$, hence the set of all non-optimal policies. If $2\gamma\varepsilon/(1-\gamma) < \delta$, that is $\varepsilon < \delta(1-\gamma)/(2\gamma) := \varepsilon_0$, then $\|V_\pi - V^*\|_\infty < \delta$ and $\widehat{\pi}$ is optimal $\square$

An ε-greedy policy update guarantees improvement

The ε-greedy policy π' wrt V satisfies $V^{\pi'} \geq V^\pi$ because[1]

$$
\begin{aligned}
T^{\pi'} V^\pi(s) &= \sum_a \pi'(a|s) Q^\pi(s,a) \\
&= \frac{\varepsilon}{|\mathcal{A}|} \sum_a Q^\pi(s,a) + (1-\varepsilon) \max_a \, Q^\pi(s,a) \\
&\geq \frac{\varepsilon}{|\mathcal{A}|} \sum_a Q^\pi(s,a) + (1-\varepsilon) \sum_a \underbrace{\frac{\pi(a|s) - \frac{\varepsilon}{|\mathcal{A}|}}{1-\varepsilon}}_{\text{sums to } 1} \, Q^\pi(s,a) \\
&= \frac{\varepsilon}{|\mathcal{A}|} \sum_a Q^\pi(s,a) + \sum_a \pi(a|s) \, Q^\pi(s,a) - \sum_a \frac{\varepsilon}{|\mathcal{A}|} Q^\pi(s,a) \\
&= \sum_a \pi(a|s) Q^\pi(s,a) = V^\pi(s).
\end{aligned}
$$

Hence $\lim_{k\to\infty} (T^{\pi'})^{(k)} V^\pi = V^{\pi'} \geq V^\pi$ due to monotonicity.

[1] Remember that $Q^\pi(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V^\pi(s')$.

The approximate policy iteration problem (words)

The convergence of policy iteration to the optimal policy depends on step-wise improvement. This improvement is guaranteed under the following conditions for all s and a:

- V^π and $p(s'|s,a)$ are known,
- $\mathbb{E}_{s' \sim p(s'|s,a)}[V^\pi(s')]$ is tractable,
- $TV_k \geq V^{\pi_k}$ for all s and a.

These conditions rarely hold. In practice, we do **approximate policy iteration** where both policy evaluation and policy improvement steps contain approximation errors. Furthermore, we maintain parametric value functions, where a parameter update does not always improve all state-action pairs simultaneously. All these errors affect how precisely the optimal policy can be identified.

The approximate policy iteration problem (math)

Assume an approximate policy iteration algorithm that generates a sequence of approximate values $(V_k)_{k=0}^{\infty}$ and the corresponding approximate greedy policies $(\pi_k)_{k=0}^{\infty}$ that incur bounded error

- $\|V_k - V^{\pi_k}\|_\infty \leq \varepsilon$ (Policy evaluation error)
- $\|T^{\pi_{k+1}} V_k - T V_k\|_\infty \leq \delta$ (Greedy policy identification error)

Above both $T^{\pi_{k+1}}$ and T are exact. Then the **policy improvement error** is:

$$V^{\pi_{k+1}} \geq V^{\pi_k} - \frac{\delta + 2\gamma\varepsilon\mathbb{1}}{1-\gamma}, \qquad k = 0, 1, \ldots$$

The resulting **optimal policy identification error** is [2]

$$\limsup_{k\to\infty} \|V^{\pi_k} - V^*\|_\infty \leq \frac{\delta + 2\gamma\varepsilon\mathbb{1}}{(1-\gamma)^2}$$

[2] Limit superior ($\limsup$) for a real-valued sequence $(x_n)_{n\in\mathbb{N}}$ is defined as below

$$\limsup_{n\to\infty} x_n := \inf_{n\in\mathbb{N}} \sup_{k\geq n} x_k = \inf\{\sup\{x_k : k \geq n\} : n \in \mathbb{N}\}$$

Policy improvement in terms of advantages

Denote by τ' a trajectory obtained by policy π', then

$$\eta(\pi') - \eta(\pi) = \mathbb{E}_{\tau' \sim \pi'} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - V^\pi(s_0) \right]$$

$$= \mathbb{E}_{\tau' \sim \pi'} \left[\sum_{t=0}^{\infty} \gamma^t [r(s_t, a_t) + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)] \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \rho_{\pi'}} \left[\sum_{t=0}^{\infty} \gamma^t [r(s, a) + \gamma V^\pi(s') - V^\pi(s)] \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \rho_{\pi'}} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{a \sim \pi'(\cdot|s)} \Big[\underbrace{r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V^\pi(s')] - V^\pi(s)}_{:= A^\pi(s,a) \quad \text{called an ''advantage function''}} \Big] \right]$$

$$= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \rho_{\pi'}} \left[\mathbb{E}_{a \sim \pi'(\cdot|s)} \Big[A^\pi(s, a) \Big] \right] := \frac{1}{1-\gamma} \mathbb{A}_\pi(\pi')$$

where $\mathbb{A}_\pi(\pi')$ is called a **policy advantage** for π' over π.

Improving mean advantage is enough

Critical implication of this advantage-based view

$$\eta(\pi') - \eta(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\rho_{\pi'}}\left[\mathbb{E}_{a\sim\pi'(\cdot|s)}\left[A^\pi(s,a)\right]\right]$$

is that keeping $A^\pi(s,a) \geq 0$ on average across the (s,a) pairs is enough for policy improvement. We can achieve this by a greedy policy update with respect to the r.h.s. of the equation above instead of an approximate Bellman backup $\approx TV^\pi$:

$$\pi' := \arg\max_\pi \mathbb{E}_{s\sim\rho_{\pi'}}[\mathbb{E}_{a\sim\pi'(\cdot|s)}[A^\pi(s,a)]].$$

The integral for $\rho_{\pi'}$ can be approximated by importance sampling with proposal distribution ρ_π. But this comes with a prohibitively high estimator variance. The following objective is much easier to learn from Monte Carlo samples as it depends on trajectories obtained from the current policy.

$$\pi' := \arg\max_\pi \mathbb{E}_{s\sim\rho_{\pi}}[\mathbb{E}_{a\sim\pi'(\cdot|s)}[A^\pi(s,a)]].$$

The price of $\rho_{\pi'} \to \rho_\pi$

Theorem

Given infinitely differentiable $f : D \to A$ with a bounded range. For any point $a \in D$ in the function domain with $f'(a) > 0$, an update $a' := a + \alpha f'(a)$ with $0 < \alpha < 1/|\mathcal{O}(a^2)|$ yields $f(a') > f(a)$.

Hence, a small enough α guarantees a gradient-based improvement. Denote

$$L_\pi(\pi') = \eta(\pi) + \frac{1}{1-\gamma} \mathbb{E}_{s \sim \rho_\pi} \left[\mathbb{E}_{a \sim \pi'(\cdot|s)} \left[A^\pi(s, a) \right] \right]$$

and notice that $L_{\pi_\theta}(\pi_\theta) = \eta(\pi_\theta)$ and $\nabla_\theta L_{\pi_\theta}(\pi_\theta) = \nabla_\theta \eta(\pi_\theta)$. The first-order Taylor expansions[3] of L_{π_θ} and η match around π_θ. Then we can improve on η by optimizing L_{π_θ} with **small updates**! Next question is how small they should be.

[3] Any infinitely differentiable function f with bounded range can be expressed as a Taylor series:

$$f(x) = \sum_{i=1}^{\infty} \frac{1}{i!} f^{(i)}(a)(x-a)^i \text{for any } a \text{ in the function domain.}$$

Proof

Using a Taylor expansion, we have

$$\begin{aligned}
f(a') - f(a) &= f'(a)(a' - a) + \sum_{i=2}^{\infty} \frac{1}{i!} f^{(i)}(a)(x-a)^i \\
&= f'(a)(\alpha f'(a)) + \sum_{i=2}^{\infty} \frac{1}{i!} f^{(i)}(a)(\alpha f'(a))^i. \\
&= f'(a)(\alpha f'(a)) + (\alpha f'(a))^2 \sum_{i=0}^{\infty} \frac{1}{i!} f^{(i)}(a)(\alpha f'(a))^{i-2}. \\
&= \alpha (f'(a))^2 + (\alpha f'(a))^2 \mathcal{O}(a^2),
\end{aligned}$$

If $\mathcal{O}(a^2) > 0$ any $\alpha > 0$ works. Otherwise choose $\alpha < -1/\mathcal{O}(a^2)$ $\square$

Per timestep price of a mixture update

Consider acting via $\bar{\pi}(\cdot|s_t)$ that follows

$$c_t \sim \text{Bernoulli}(c_t|\alpha), \qquad a_t \sim \pi'(\cdot|s_t)^{c_t}\pi(\cdot|s_t)^{1-c_t}$$

Let $n_t = \sum_{j=0}^{t} c_j$, then $P(n_t = 0) = (1-\alpha)^t$ and $P(n_t > 0) = 1 - (1-\alpha)^t$. Since $\mathbb{E}_{a_t \sim \pi(\cdot|s_t)}\Big[A^\pi(s_t, a_t)\Big] = 0$, we have

$$\mathbb{E}_{s_t \sim p_{\bar{\pi}}^t}\Big[\mathbb{E}_{a_t \sim \bar{\pi}(\cdot|s_t)}\Big[A^\pi(s_t, a_t)\Big]\Big] = \alpha \mathbb{E}_{s_t \sim p_{\bar{\pi}}^t | c_t = 1}\Big[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\Big[A^\pi(s_t, a_t)\Big]\Big].$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}_{s_t \sim p_{\bar{\pi}}^t | c_t = 1}\Big[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\Big[A^\pi(s_t, a_t)\Big]\Big] = \\
&\quad P(n_t = 0)\mathbb{E}_{s_t \sim p_{\bar{\pi}}^t | n_t = 0}\Big[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\Big[A^\pi(s_t, a_t)\Big]\Big] \\
&\qquad + P(n_t > 0)\mathbb{E}_{s_t \sim p_{\bar{\pi}}^t | n_t > 0}\Big[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\Big[A^\pi(s_t, a_t)\Big]\Big].
\end{aligned}$$

Per timestep price of a mixture update

Given $\varepsilon := \|V^\pi\|_\infty$ then

$$\mathbb{E}_{s_t \sim p_\pi^t | c_t = 1}\left[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\left[A^\pi(s_t, a_t)\right]\right] =$$

$$\underbrace{(1-\alpha)^t \mathbb{E}_{s_t \sim p_{\pi'}^t, |n_t = 0}}_{\mathbb{E}_{s_t \sim p_\pi^t}}\left[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\left[A^\pi(s_t, a_t)\right]\right]$$

$$+ (1-(1-\alpha)^t)\underbrace{\mathbb{E}_{s_t \sim p_\pi^t | n_t > 0}\left[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\left[A^\pi(s_t, a_t)\right]\right]}_{\geq -2\varepsilon}$$

$$\geq \mathbb{E}_{s_t \sim p_\pi^t}\left[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)}\left[A^\pi(s_t, a_t)\right]\right] - 2(1-(1-\alpha)^t)\varepsilon$$

Full price of a mixture update

$$
\begin{aligned}
&\eta(\pi') - \eta(\pi) \\
&\geq \frac{\alpha}{1-\gamma} \sum_{t=0}^{\infty} \gamma^t \left(\mathbb{E}_{s_t \sim p_\pi^t} \left[\mathbb{E}_{a_t \sim \pi'(\cdot|s_t)} \left[A^\pi(s_t, a_t) \right] \right] - 2\varepsilon(1 - (1-\alpha)^t) \right) \\
&\geq \frac{\alpha}{1-\gamma} \mathbb{E}_{s \sim \rho_\pi} \left[\mathbb{E}_{a \sim \pi'(\cdot|s)} \left[A^\pi(s, a) \right] \right] - \frac{2\varepsilon\alpha}{1-\gamma} \left(\frac{1}{1-\gamma} - \frac{1}{1-\gamma(1-\alpha)} \right)
\end{aligned}
$$

Hence π' improves over π if the r.h.s. is greater than zero, in other words

$$
\mathbb{E}_{s \sim \rho_\pi} \left[\mathbb{E}_{a \sim \pi'(\cdot|s)} \left[A^\pi(s, a) \right] \right] \geq \frac{2\varepsilon\alpha\gamma}{(1-\gamma)(1-\gamma(1-\alpha))} \geq \frac{2\varepsilon\alpha\gamma}{1-\gamma}
$$

Therefore the updates should be slower than the following mixture rate:

$$
\alpha \leq \frac{1-\gamma}{2\varepsilon\gamma} \mathbb{E}_{s \sim \rho_\pi} \left[\mathbb{E}_{a \sim \pi'(\cdot|s)} \left[A^\pi(s, a) \right] \right].
$$

to guarantee an improvement based on a greedy update wrt ρ_π. This method is called **Conservative Policy Iteration**.

Recipe for Conservative Policy Iteration

- Do importance sampling on the advantage estimate
- Approximate the action-values by Monte Carlo sampling
- Learn a value function approximator V_θ
- Limit policy update speed

$$\mathbb{E}_{s\sim\rho_\pi}\Big[\mathbb{E}_{a\sim\pi'(\cdot|s)}\Big[A^\pi(s,a)\Big]\Big] = \mathbb{E}_{s\sim\rho_\pi}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\left[\frac{\pi'(a|s)}{\pi(a|s)}A^\pi(s,a)\right]\right]$$

$$= \mathbb{E}_{s\sim\rho_\pi}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\left[\frac{\pi'(a|s)}{\pi(a|s)}\Big(Q^\pi(s,a) - V^\pi(s)\Big)\right]\right]$$

$$\approx \frac{1}{T}\sum_{t=0}^{T-1}\frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)}\left(\sum_{j=t}^{T-1}\gamma^{j-t}r_j - V_\theta(s_t)\right)$$

Recipe for Conservative Policy Iteration

Limit policy update speed by clipping $[x]_{1-\epsilon}^{1+\epsilon} = \max(\min(x, 1+\epsilon), 1-\epsilon)$ and reduce advantage estimator variance using λ-returns, called **Generalized Advantage Estimate (GAE)** [Schulman et al., 2016]):

$$\arg\max_{\pi'} \frac{1}{T} \sum_{t=0}^{T-1} \left[\frac{\pi'(a_t|s_t)}{\pi(a_t|s_t)} \right]_{1-\epsilon}^{1+\epsilon} \left(\underbrace{(1-\lambda) \sum_{j=t}^{T-1} (\lambda\gamma)^{j-t} \delta_j}_{\widehat{A}_\pi^\lambda(s_t, a_t): \text{GAE}} \right)$$

with $\delta_j := r_j + \gamma V_\theta(s_{j+1}) - V_\theta(s_j)$.

The Proximal Policy Optimization (PPO) Algorithm

repeat
    $\mathcal{D} := \emptyset$ ▷ Erase replay buffer
    $s := \texttt{env.reset}()$
    repeat
        $a \sim \pi(\cdot|s)$
        $r, s' := \texttt{env.step(s,a)}$
        $\mathcal{D} := \mathcal{D} \cup (s, a, r, s')$
        $s = s'$
    until episode end
    repeat
        Compute $\widehat{A}_\pi^\lambda(s, a)$ for all $(s, a) \in \widetilde{D}$ for minibatch $\widetilde{\mathcal{D}} \sim \mathcal{D}$
        $\pi' := \arg\max_{\pi'} \frac{1}{|\widetilde{\mathcal{D}}|} \sum_{(s,a)\in\widetilde{\mathcal{D}}} \left[\frac{\pi'(a|s)}{\pi(a|s)}\right]_{1-\epsilon}^{1+\epsilon} \widehat{A}_\pi^\lambda(s, a)$
        Update V_θ using TD(λ) on $\widetilde{\mathcal{D}}$
    until convergence
    $\pi := \pi'$
until convergence