On The Linear Convergence of Policy Gradient Methods

Ziniu Li
ziniuli@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

June 11, 2021

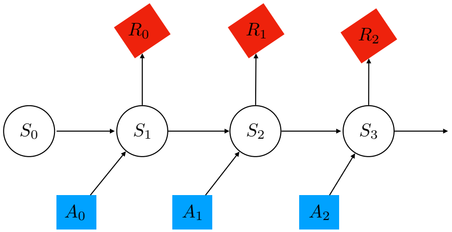
Bhandari, Jalaj, and Daniel Russo.
"On the Linear Convergence of Policy Gradient Methods for Finite MDPs." AISTATS, 2021.

Outline

Outline

Markov Decision Process

- An infinite-horizon discounted Markov Decision Process (MDP) [Puterman, 2014] is described by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma, \rho)$:
  - $\mathcal{S}$ and $\mathcal{A}$ are the finite state and action space, respectively.
  - $p(s'|s,a)$ is the transition probability matrix.
  - $r : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$ is the deterministic reward function.
  - $\gamma \in (0, 1)$ is the discount factor.
  - ρ specifies the initial state distribution.



Markov Decision Process: Policy

- To interact with MDP, we need a policy π to select actions.
  - $\pi(a|s)$ determines the probability of selecting action a at state s.
- The quality of policy π is measured by state value function V^π:

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | \pi, s_0 = s\right]. \quad (1)$$

  - $V^\pi(s)$ measures the the expected long-term discounted reward when starting from state s.
  - $V^\pi(s) \in [0, \frac{1}{1-\gamma}]$ by definition.
- To take the initial state distribution into account, we define

$$V(\pi) := V^\pi(\rho) := \mathbb{E}_{s_0 \sim \rho}\left[V^\pi(s_0)\right]. \quad (2)$$

Markov Decision Process: Value Function

- Sometimes, it is more convenient to introduce state-action value function Q^π:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | \pi, s_0 = s, a_0 = a\right]. \tag{3}$$

  - $Q^\pi(s,a)$ measures the the expected long-term discounted reward when starting from state s with action a.
  - $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[Q^\pi(s,a)\right]$ by definition.

Markov Decision Process: Discounted Stationary Distribution

- To facilitate later analysis, we introduce discounted stationary distribution d^π:

$$d^\pi_{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s|\pi, s_0). \tag{4}$$

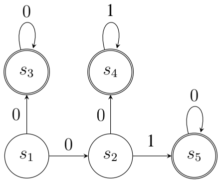
$\rightsquigarrow$ $d^\pi_{s_0}(s)$ measures the discounting probability to visit s starting from the initial state s_0.

- To take the initial state distribution into account, we define d^π_ρ as

$$d^\pi_\rho(s) = \mathbb{E}_{s_0\sim\rho}\left[d^\pi_{s_0}(s)\right]. \tag{5}$$

Markov Decision Process: Example

- Consider the following MDP example: a_1 : "up"; a_2 : "right".



$$\pi(a_2|s_1) = 1;$$

$$\pi(a_1|s_2) = 0.5,\ \pi(a_2|s_2) = 0.5.$$

$$d_\rho^\pi(\cdot) = (1-\gamma)\cdot\left(1, \gamma, 0, \frac{0.5\gamma^2}{1-\gamma}, \frac{0.5\gamma^2}{1-\gamma}\right),$$

$$V^\pi(s_1) = 0 + 0.5\left[\gamma \times 1\right] + 0.5\left[\gamma^2 \times 1 + \gamma^3 \times 1 + \cdots\right] = \frac{0.5\gamma}{1-\gamma},$$

$$Q^\pi(s_2, a_1) = 1, \quad Q^\pi(s_2, a_2) = \frac{\gamma}{1-\gamma}, \quad V^\pi(s_2) = \frac{0.5}{1-\gamma}.$$

Outline

Policy Iteration

- In this section, we consider a well-known algorithm: policy iteration.

Algorithm 1 Policy Iteration

Input: initialization $\pi^0 \in \Delta(\mathcal{A})^{|\mathcal{S}|}$.

1: **for** $t = 0, 1, \cdots,$ **do**
2: $\quad Q^{\pi_t} \leftarrow$ evaluate the state-action value function of π^t.
3: $\quad \pi^{t+1}(s) := \operatorname{argmax}_{a \in \mathcal{A}} Q^{\pi_t}(s, a)$.
4: **end for**

- The analysis of policy iteration is fundamental to policy optimization.

Policy Iteration: Linear Convergence

Theorem 1 (Linear convergence of policy iteration).

For any initialization policy π^0, *we have*

$$\|V^* - V^{\pi_t}\|_\infty \leq \frac{1}{1-\gamma}\exp(-t).$$

Policy Iteration: Proof of Theorem 1

The proof of Theorem 1 relies on the γ-contraction of the Bellman optimal operator $\mathcal{T}^*$:

$$\forall \pi, \pi', \quad \left\| \mathcal{T}^* V^\pi - \mathcal{T}^* V^{\pi'} \right\|_\infty \leq \gamma \left\| V^\pi - V^{\pi'} \right\|_\infty .$$

In particular, consider $\pi' = \pi^*$ and let $V^* := V^{\pi^*}$,

$$\forall \pi, \quad \left\| \mathcal{T}^* V^\pi - \mathcal{T}^* V^* \right\|_\infty \leq \gamma \left\| V^\pi - V^* \right\|_\infty . \tag{6}$$

Hence, performing an Bellman update can improve the value function by a γ-multiplicative factor.

$\rightsquigarrow$ The issue of policy iteration analysis is to bound the improvement of the value function of a policy (i.e., $V^{\pi^{t+1}}$) rather than an artificial value function ($\mathcal{T} V^{\pi^t}$)!

Policy Iteration: Bellman Operators

- To facilitate later analysis, we define the Bellman operator $\mathcal{T}^\pi$:

$$\mathcal{T}^\pi V(s) := \sum_{a \in \mathcal{A}} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right].$$

- The Bellman optimal operator $\mathcal{T}^*$ is

$$\mathcal{T}^* V(s) := \max_{a \in \mathcal{A}} \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right].$$

- According to fixed point theory, we have

$$V^\pi = \mathcal{T}^\pi V, \quad \forall \pi; \quad \mathcal{T}^* V^* = V^*,$$

where "$=$" holds elementwise.

Policy Iteration: Proof of Theorem 1

To facilitate analysis, let us introduce the notation π^+:

$$\pi^+(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} Q^\pi(s,a), \quad \forall s \in \mathcal{S}. \tag{7}$$

In terms of Bellman operators, this can be equivalently expressed as, $\mathcal{T}^{\pi^+} V^\pi = \mathcal{T}^* V^\pi$ with V^π being the state value functions of policy π. Our first observation is that

$$V^\pi \preceq \mathcal{T}^* V^\pi = \mathcal{T}^{\pi +} V^\pi. \tag{8}$$

The magic is that if we repeatedly apply (8), the RHS goes to $V^{\pi^{t+1}}$:

$$\boxed{V^\pi \preceq \mathcal{T}^{\pi^+} V^\pi \preceq \left(\mathcal{T}^{\pi^+}\right)^2 V^\pi \preceq \cdots \preceq \left(\mathcal{T}^{\pi^+}\right)^\infty V^\pi = V^{\pi^+},} \tag{9}$$

which implies that the improvement of $V^{pi^{t+1}}$ is always better than $\mathcal{T}^* V^{\pi^t}$ (i.e., the one obtained by value iteration).

Policy Iteration: Proof of Theorem 1

Based on previous results, we have

$$\left\|V^{\pi^+} - V^*\right\|_\infty \overset{(9)}{\leq} \left\|\mathcal{T}^{\pi^+} V^\pi - V^*\right\|_\infty \overset{(6)}{\leq} \gamma \left\|V^\pi - V^*\right\|_\infty . \tag{10}$$

⇝ For policy iteration, $\pi^+ := \pi^{t+1}$ and $Q^\pi := Q^{\pi^t}$ and $V^\pi := V^{\pi^t}$.

⇝ (10) implies

$$\left\|V^{\pi^{t+1}} - V^{\pi^t}\right\|_\infty \geq (1-\gamma) \left\|V^{\pi^t} - V^*\right\|_\infty . \tag{11}$$

- (Remark on policy optimization) Though value iteration also enjoy a linear convergence rate, the induced greedy policy (w.r.t. the ε-optimal learned value function) is $\varepsilon/(1-\gamma)$-optimal. However, policy iteration does not have such an issue by the monotonicity in (9).

Outline

Outline

Weighted Bellman Objective

- For any policy π, let us introduce weighted policy iteration or weighted Bellman objective, defined as

$$\mathcal{B}(\overline{\pi}|d^{\pi}, Q^{\pi}) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\pi}(s)Q^{\pi}(s,a)\overline{\pi}(a|s) = \langle Q^{\pi}, \overline{\pi}\rangle_{d^{\pi}\times 1}, \tag{12}$$

  where $\langle v, u\rangle_W = \sum_i \sum_j v(i,j)u(i,j)W(i,j)$ and $d^{\pi} \times 1$ denotes a weight matrix that places $d^{\pi}(s)$ on any state-action pair $(s, \cdot)$.

- Our objective is to maximize such defined weighted Bellman objective,

$$\pi^{+} = \operatorname*{argmax}_{\overline{\pi}\in\Pi} \mathcal{B}(\overline{\pi}|d^{\pi}, Q^{\pi}).$$

- Now let us check the gradient of $\mathcal{B}(\overline{\pi}|d^{\pi}, Q^{\pi})$.

$$\frac{\partial \mathcal{B}(\overline{\pi}|d^{\pi}, Q^{\pi})}{\partial \overline{\pi}(a|s)} = d^{\pi}(s)Q^{\pi}(s,a).$$

Weighted Policy Gradient

- Consider the weighted objective function:

$$\ell(\pi) = (1-\gamma)\sum_{s\sim\rho} \rho(s) V^{\pi}(s). \tag{13}$$

- Recall the policy gradient theorem states that

$$\frac{\partial \ell(\pi)}{\partial \pi(a|s)} = d^{\pi}(s) Q^{\pi}(s,a).$$

Theorem 2 (Policy Gradient Theorem).

For the direct parameterization and any initial state distribution μ, we have

$$\frac{\partial V^{\pi}(\mu)}{\partial \pi(a|s)} = \frac{1}{1-\gamma} d_{\mu}^{\pi}(s) Q^{\pi}(s,a).$$

Connection between Policy Iteration and Policy Gradient

- We see that the gradient of weighted Bellman objective is identical to the gradient of expected return!

$$\frac{\partial \langle Q^\pi, \overline{\pi} \rangle_{d^\pi \times 1}}{\partial \overline{\pi}(a|s)} = \frac{\partial \ell(\pi)}{\partial \pi(a|s)} = d^\pi(s) Q^\pi(s,a).$$

- Importantly, we see that the solution of weighted Bellman objective corresponds to a policy iteration update:

$$\pi^+ \in \underset{\overline{\pi}}{\operatorname{argmax}} \langle Q^\pi, \overline{\pi} \rangle_{d^\pi \times 1},$$

  where π^+ is defined as in (7) for policy iteration. Hence, we design policy-gradient algorithms and analyze them in terms of policy iteration update (i.e., Bellman update).
- $\rho(s) > 0$ for any $s \in \mathcal{S}$ is indispensable to ensure the connection is valid.

Outline

Policy Gradient Algorithms

- **Frank-wolfe.** The key idea of frank-wolfe is to optimize the linearized objective over the constrained set and then to make a convex combination. More precisely, define

$$\pi^{+} = \operatorname*{argmax}_{\overline{\pi}\in\Pi}\langle\nabla\ell(\pi),\overline{\pi}\rangle = \operatorname*{argmax}_{\overline{\pi}\in\Pi}\langle Q^{\pi},\overline{\pi}\rangle_{d^{\pi}\times 1}; \tag{14}$$

  then we update the policy to $\pi' = (1-\eta)\pi + \eta\pi^{+}$ for some $\eta \in [0,1]$.

- **Projected Gradient Ascent.** The core of projected gradient descent is more simple: we first take a gradient descent update then project the updated policy into the constrained set:

$$\begin{aligned}\pi' &= \operatorname*{argmax}_{\overline{\pi}\in\Pi}\left\{\langle\nabla\ell(\pi),\overline{\pi}\rangle - \frac{1}{2\eta}\left\|\overline{\pi}-\pi\right\|_2^2\right\}\\ &= \operatorname*{argmax}_{\overline{\pi}\in\Pi}\left\{\langle Q^{\pi},\overline{\pi}\rangle_{d^{\pi}\times 1} - \frac{1}{2\eta}\left\|\overline{\pi}-\pi\right\|_2^2\right\}\end{aligned}$$

  We see that as $\eta\to\infty$ (i.e., there is no regularization), π' converges to the solution of (14).

Policy Gradient Algorithms

- **Mirror-descent.** The mirror descent method adapts to the geometry of the probability simplex by using a non-Euclidean regularizer. We focus on using the Kullback-Leibler(KL) divergence, under which an iteration of mirror descend updates policy π to π' as
$$\pi' = \operatorname*{argmax}_{\pi \in \Pi} \left\{ \langle \nabla \ell(\pi), \bar{\pi} \rangle - \frac{1}{\eta} D_{\mathrm{KL}}(\bar{\pi} \| \pi) \right\}, \tag{15}$$
where $D_{\mathrm{KL}}(\bar{\pi}\|\pi) = \sum_{s \in \mathcal{S}} D_{\mathrm{KL}}\left(\pi(\cdot|s) \| \bar{\pi}(\cdot|s)\right)$, and
$D_{\mathrm{KL}}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log\left(p(x)/q(x)\right)$ for two probability distributions p and q.
- It is well know that the solution to (15) is the exponentiated gradient update [Bubeck, 2015, Section 6.3],
$$\pi'(a|s) = \frac{\pi(a|s) \exp\left(\eta d^{\pi}(s) Q^{\pi}(s,a)\right)}{\sum_{a \in \mathcal{A}} \pi(a|s) \exp\left(\eta d^{\pi}(s) Q^{\pi}(s,a)\right)}. \tag{16}$$
Again, we see that as $\eta \to \infty$, π' converges to a policy iteration update.

Policy Gradient Algorithms

- **Natural policy gradient.** We focus on NPG applied to the softmax parameterization for which it is actually an instance of mirror descent with a specific regularizer. In particular, we have

$$\pi' = \operatorname*{argmax}_{\pi \in \Pi} \left\{ \langle \nabla \ell(\pi), \overline{\pi} \rangle - \frac{1}{\eta} D_{\mathrm{KL}}^{d^{\pi}}(\overline{\pi} \| \pi) \right\}, \tag{17}$$

where $D_{\mathrm{KL}}^{d^{\pi}}(\overline{\pi} \| \pi) = \sum_{s \in \mathcal{S}} d^{\pi}(s) D_{\mathrm{KL}}\left(\pi(\cdot|s) \| \overline{\pi}(\cdot|s)\right)$ is a weighted regularizer.

- Again, (17) corresponds to a exponentiated policy update:

$$\pi'(a|s) = \frac{\pi(a|s) \exp\left(\eta Q^{\pi}(s,a)\right)}{\sum_{a \in \mathcal{A}} \pi(a|s) \exp\left(\eta Q^{\pi}(s,a)\right)}. \tag{18}$$

Note that this update is independent of state distribution d^{π}.

Outline

Stepsize Choice

- In this part, we tackle the stepsize issue. Our main focus is exact line search.
- Exact line search will find the "optimal" stepsize by line search; more precisely, $\pi^{t+1} = \pi_{\eta^*}^{t+1}$, where $\eta^* = \mathrm{argmax}_\eta\, \ell(\pi_\eta^{t+1})$ whenever this maximizer exists. More generally, we define
$$\pi^{t+1} = \operatorname*{argmax}_{\pi \in \Pi^{t+1}} \ell(\pi), \tag{19}$$
where $\Pi^{t+1} = \mathrm{Closure}(\{\pi_\eta^{t+1}\})$ denotes the close curve of policies traced out by varying stepsize η.
- For example, $\Pi^{t+1} = \{\eta\pi^t + (1-\eta)\pi_+^t : \eta \in [0,1]\}$ is the line segment connecting the current policy π^{t+} and its policy iteration update π_+^t. For NPG, $\Pi^{t+1} = \{\pi_\eta^{t+1}\}$ is a curve where $\pi_0^{t+1} = \pi^t$ and $\pi_\eta^{t+1} \to \pi_+^t$ as $\eta \to \infty$. Since π_+^t is to attainable under any fixed η, this curve is not closed. By taking the closure, and define line search via (19), certain formulas become cleaner.

Stepsize Remark

- (Policy parameterization and infima vs minima). The class of softmax policies can approximately any stochastic policy to arbitrary precision, however, this is nearly the same as optimizing over Π.
- (Policy optimization vs parameter optimization) The above results do not apply to more naive gradient methods that directly linearize $\ell(\pi_\theta)$ with respect to θ. In that case, a gradient update to θ may not approximate a policy iteration update, no matter how large the stepsize is chosen to be.

Outline

Linear Convergence of Policy Optimization I

Suppose one of the first-order algorithms introduced in Section 2 is applied to maximize $\ell(\pi)$ over $\pi \in \Pi$ with stepsizes $\{\eta_t\}_{t\geq 0}$. Let π^0 be the initial policy and $\{\pi^t\}_{t\geq 0}$. Let π^0 denote the initial policy and $\{\pi^t\}_{t\geq 0}$ denote the sequence of iterates. The following bounds apply [Bhandari and Russo, 2021].

- **Exact line search.** If either Frank-Wolfe, projected gradient descent, mirror descent, or NPG is applied with stepsizes chosen by exact line search in (19), then

$$\left\| V^{\pi^t} - V^* \right\|_\infty \leq \left(1 - \min_{s\in\mathcal{S}} \rho(s)(1-\gamma)\right)^t \frac{\left\| V^{\pi^0} - V^* \right\|_\infty}{\min_{s\in\mathcal{S}} \rho(s)}.$$

- **Constant stepsize Frank-Wolfe.** Under Frank-Wolfe with constant stepszie $\eta \in (0,1]$,

$$\left\| V^{\pi_t} - V^* \right\|_\infty \leq (1 - \eta(1-\gamma))^t \left\| V^{\pi^0} - V^* \right\|_\infty .$$

Linear Convergence of Policy Optimization II

- **Natural policy gradient with softmax policies and adaptive stepsize.** Fix any $\varepsilon > 0$. Let $a_t^* = \text{argmax}_a Q^{\pi^t}(s,a)$. Suppose NPG is performed with an adaptive step-size sequence,

$$\eta_t(s) \geq \frac{2}{(1-\gamma)\varepsilon} \log\left(\frac{2}{\pi^t(s,a_t^*)}\right).$$

Then,

$$\left\|V^{\pi^t} - V^*\right\|_\infty \leq \left(\frac{1+\gamma}{2}\right)^t \left\|V^{\pi^0} - V^*\right\|_\infty + \varepsilon.$$

Analysis For Linear Convergence: Warm-up

- How to prove the linear convergence for a sequence $\{f(x_k)\}_k$? (i.e., what are key steps?)
- One of key step in previous analysis (for policy iteration) is

$$\text{(Type I):} \qquad f(x_{k+1}) - f^* \leq \gamma \left(f(x_k) - f^*\right). \tag{20}$$

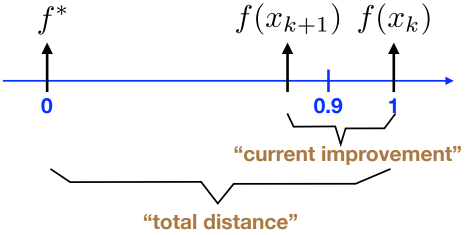
  with $\gamma \in (0, 1)$.
- What if (20) is hard to verify? We move to the following step:

$$\text{(Type II):} \quad f(x_k) - f(x_{k+1}) \geq (1-\gamma)\left(f(x_k) - f^*\right) \tag{21}$$

  $\rightsquigarrow$ (21) implies (20).

Analysis For Linear Convergence: Warm-up

(Type II): $f(x_k) - f(x_{k+1}) \geq (1-\gamma)\,(f(x_k) - f^*)$



Message: "current improvement" is at least $(1-\gamma)$ times of "current distance".

Proof of Exact Line Search I

For each algorithm at iteration t, the policy iteration update π_+^t is contained in Π^{t+1} introduced as in (19). Therefore, for each algorithm,

$$\ell(\pi^{t+1}) = \max_{\pi \in \Pi^{t+1}} \ell(\pi) \geq \ell(\pi_+^t).$$

Therefore, PG with exact line search is never worse than a policy iteration update. The remaining step is to monitor the progress in terms of expected return by the linear convergence of policy iteration that is bounded by ℓ_∞-norm.

$$\begin{aligned}
\ell(\pi^{t+1}) - \ell(\pi^t) &\geq \ell(\pi_+^t) - \ell(\pi^t) \\
&= (1-\gamma) \sum_{s \in \mathcal{S}} \rho(s) \left(V^{\pi_+^t}(s) - V^{\pi^t}(s) \right) \\
&\geq (1-\gamma) \rho_{\min} \sum_{s \in \mathcal{S}} \left(V^{\pi_+^t}(s) - V^{\pi^t}(s) \right)
\end{aligned}$$

Proof of Exact Line Search II

$$\begin{aligned}
&\overset{(9)}{\geq} (1-\gamma)\rho_{\min} \left\| V^{\pi_+^t} - V^{\pi^t} \right\|_\infty && (V^{\pi_+^t} \succeq V^{\pi^t}) \\
&\overset{(11)}{\geq} (1-\gamma)\rho_{\min} \left[(1-\gamma) \left\| V^* - V^{\pi^t} \right\|_\infty \right] \\
&\geq (1-\gamma)\rho_{\min} \left[(1-\gamma) \sum_{s\in\mathcal{S}} \rho(s) \left(V^*(s) - V^{\pi^t}(s) \right) \right] && (V^* \succeq V^{\pi^t}) \\
&= (1-\gamma)\rho_{\min} \left(\ell(\pi^*) - \ell(\pi^t) \right)
\end{aligned}$$

Rearranging, we obtain that

$$\ell(\pi^*) - \ell(\pi^{t+1}) \leq (1 - (1-\gamma)\rho_{\min}) \left(\ell(\pi^*) - \ell(\pi^t) \right).$$

To obtain the guarantee for $V^* - V(\pi^{t+1})$ instead of $\ell(\pi^*) - \ell(\pi^{t+1})$, we note that

$$\left\| V^* - V(\pi^{t+1}) \right\|_\infty \leq \frac{1}{(1-\gamma)\rho_{\min}} \left(\ell(\pi^*) - \ell(\pi^{t+1}) \right)$$

Proof of Exact Line Search III

$$\begin{aligned} &\leq \frac{(1-(1-\gamma)\rho_{\min})}{(1-\gamma)\rho_{\min}} \left(\ell(\pi^*) - \ell(\pi^t)\right) \\ &\leq \frac{(1-(1-\gamma)\rho_{\min})^{t+1}}{\rho_{\min}} \left\| V^* - V^{\pi^0} \right\|_\infty, \end{aligned}$$

where the last step follows $\sum_{s\in\mathcal{S}} \rho(s) \left(V^*(s) - V^{\pi^0}(s)\right) \leq \left\| V^* - V^{\pi^0} \right\|_\infty$ due to ρ is a probability simplex.

Proof of Constant Stepsize Frank-Wolfe I

Recall that a Frank-Wolfe update amounts to a soft policy iteration update:

$$\pi^{t+1}(s) = (1-\eta)\pi^t(s) + \eta\pi_+^t(s),$$

where π_+^t is the policy iteration update to π^t. By linearity, we have that for any state s,

$$\begin{aligned} \mathcal{T}^{\pi^{t+1}} V^{\pi^t}(s) &= (1-\eta)\mathcal{T}^{\pi^t} V^{\pi^t}(s) + \eta \mathcal{T}^{\pi_+^t} V^{\pi^t}(s) \\ &= (1-\eta) V^{\pi^t}(s) + \eta \mathcal{T}^* V^{\pi^t}(s). \end{aligned} \tag{22}$$

Since we have $V^{\pi^t} \preceq \mathcal{T}^* V^{\pi^t}$, we obtain

$$\mathcal{T}^{\pi^{t+1}} V^{\pi^t} \succeq (1-\eta)\mathcal{T}^{\pi^t} V^{\pi^t} + \eta \mathcal{T}^{\pi^t} V^{\pi^t} \succeq V^{\pi^t}.$$

By monotonicity of $\mathcal{T}^{\pi^{t+1}}$, we repeatedly apply $\mathcal{T}^{\pi^{t+1}}$ on both sides:

$$V^{\pi^{t+1}} = \lim_{k\to\infty} (\mathcal{T}^{\pi^{t+1}} V^{\pi^t})^k \succeq V^{\pi^t}.$$

Proof of Constant Stepsize Frank-Wolfe II

Therefore, from (22), we get

$$V^{\pi^{t+1}} \succeq (1-\eta)V^{\pi^t} + \eta\mathcal{T}^* V^{\pi^t}.$$

To show the linear convergence, we turn to the key step (i.e., the improvement is at least proportional to current distance):

$$\begin{aligned} V^{\pi^{t+1}} - V^{\pi^t} &\succeq \eta\left(\mathcal{T}^* V^{\pi^t} - V^{\pi^t}\right) \\ &= \eta\left(\mathcal{T}^* V^{\pi^t} - V^* + V^* - V^{\pi^t}\right) \\ &\succeq \eta\left(-\gamma(V^* - V^{\pi^t}) + V^* - V^{\pi^t}\right) \\ &= \eta(1-\gamma)(V^* - V^{\pi^t}). \end{aligned} \tag{23}$$

By the previous reasoning, we conclude that

$$\left\|V^* - V(\pi^{t+1})\right\|_\infty \le (1-\eta(1-\gamma))\left\|V^* - V^{\pi^t}\right\|_\infty.$$

Proof of NPG with Adaptive Stepsize I

Recall the natural policy gradient (NPG) update (see (18)) with an adaptive stepsize takes the form:

$$\pi^{t+1}(a|s) = \frac{\pi^t(a|s)\exp\left(\eta^t(s)Q^{\pi^t}(s,a)\right)}{\sum_{a\in\mathcal{A}}\pi^t(a|s)\exp\left(\eta^t(s)Q^{\pi^t}(s,a)\right)}.$$

For simplicity, we let $c := 2(1-\gamma)^{-1}$, which implies $\eta_t(s) \geq \frac{c}{\varepsilon}\log\left(\frac{2}{\pi^t(a_t^*|s)}\right)$, where $a_t^* = \mathrm{argmax}_a\, Q^{\pi^t}(s,a)$.

$\rightsquigarrow$ If we can use an infinitely large stepsize, we see that $\pi^{t+1} \to \pi_+^t$, which puts the probability 1 for the optimal action and the probability 0 for sub-optimal actions.

$\rightsquigarrow$ To guarantee a "minimal improvement", we need to control probabilities of sub-optimal actions decrease by a certain factor $\lambda \in (0,1)$ with a finite stepsize.

Proof of NPG with Adaptive Stepsize II

$$\frac{\pi^{t+1}(a|s)}{\pi^t(a|s)} = \frac{\exp(\eta^t(s)Q^t(s,a))}{\sum_{a'} \pi^t(a'|s)\exp(\eta^t(s)Q^t(s,a'))} = \frac{\exp(\eta^t(s)Q^t(s,a))}{Z_t} \le \lambda \in (0,1).$$

$$\overset{\text{s.p.}}{\Longrightarrow} \quad \eta^t(s)Q^t(s,a) \le \log(\lambda Z_t)$$

$$\overset{\text{s.p.}}{\Longrightarrow} \quad \eta^t(s)Q^t(s,a) \le \log\left(\lambda \pi^t(a_t^*|s)\exp(\eta^t(s)Q^t(s,a_t^*))\right) \le \lambda \log Z_t$$

$$\overset{\text{s.p.}}{\Longrightarrow} \quad \eta^t(s)Q^t(s,a) \le \log(\lambda \pi^t(a_t^*|s)) + \eta^t(s)Q^t(s,a_t^*)$$

$$\overset{\text{s.p.}}{\Longrightarrow} \quad \log\left(\frac{1}{\lambda \pi^t(a_t^*|s)}\right) \le \eta^t(s)(Q^t(s,a_t^*) - Q^t(s,a))$$

In particular, if $Q^t(s,a_t^*) - Q^t(s,a) > \delta$, it suffices to set

$$\eta^t(s) \ge \frac{1}{\delta}\log\left(\frac{1}{\lambda \pi^t(a_t^*|s)}\right).$$

Proof of NPG with Adaptive Stepsize III

Step 1: NPG update for sub-optimal actions: Fix some state $s \in \mathcal{S}$. Without loss of generality, we assume the following ordering on the Q-values: $Q^{\pi^t}(s,1) > Q^{\pi^t}(s,2) > \cdots > Q^{\pi^t}(s,|\mathcal{A}|)$, which implies the action 1 is optimal in state s under policy π^t. For error tolerance $\varepsilon > 0$, define $O_t^-(s)$ and $O_t^+(s)$ as

$$O_t^-(s) := \left\{ a | Q^{\pi^t}(s,1) - Q^{\pi^t}(s,a) \geq \frac{\varepsilon}{c} \right\},$$

$$O_t^+(s) := \left\{ a | Q^{\pi^t}(s,1) - Q^{\pi^t}(s,a) < \frac{\varepsilon}{c} \right\}.$$

Lemma 1.

For any state, $\frac{\pi^{t+1}(s,a)}{\pi^t(s,a)} \leq \frac{1}{2}, \quad \forall i \in O_t^-(s)$.

Proof of NPG with Adaptive Stepsize IV

Step 2: NPG updates as soft policy iteration: Lemma 1 shows how an NPG update with appropriate stepsize decays the probabilities of sub-optimal actions by a multiplicative factor instead of zeroing them out. This resembles a soft-policy iteration update for the set of actions $O_t^-(s)$.

Lemma 2.
Let $V^{\pi^t}(s)$ denote the state-value function for policy π^t from any starting state $s \in \mathcal{S}$. Then,

$$\mathcal{T}^{\pi^{t+1}} V^{\pi^t}(s) - V^{\pi^t}(s) \geq \frac{1}{2}\Big(\mathcal{T}^* V^{\pi^t}(s) - V^{\pi^t}(s)\Big) - \frac{\varepsilon}{c}. \tag{24}$$

Proof of NPG with Adaptive Stepsize V

Step 3: Completing the proof: Lemma 2 clearly quantifies the relationship between an NPG update with step-size α_t and a soft policy iteration update with an additive error $\frac{\varepsilon}{c}$.
$\rightsquigarrow$ It remains to prove that $\mathcal{T}^{\pi^{t+1}} V^{\pi^t} \succeq V^{\pi^t}$ so that we can repeatedly apply this relation to obtain that $V^{\pi^{t+1}} \succeq \mathcal{T}^{\pi^{t+1}} V^{\pi^t}$. To this end, we recall that

$$\pi^{t+1}(s) = \operatorname*{argmax}_{a \in \Delta(\mathcal{A})} \left[Q^{\pi^t}(s,a) - \frac{d^{\pi^t}(s)}{\eta(s)} D_{\mathrm{KL}}(a || \pi^t(s)) \right].$$

Since $a = \pi^t$ is a feasible solution, we have

$$\mathcal{T}^{\pi^{t+1}} V^{\pi^t}(s) = Q^{\pi^t}(s, \pi^{t+1}(s)) \geq Q^{\pi^t}(s, \pi^t(s)) = V^{\pi^t}(s).$$

Hence, we conclude that

$$\mathcal{T}^{\pi^{t+1}} V^{\pi^t} \succeq V^{\pi^t} \quad \Longrightarrow \quad V^{\pi^{t+1}} \succeq \mathcal{T}^{\pi^{t+1}} V^{\pi^t}.$$

Proof of NPG with Adaptive Stepsize VI

Therefore, by Lemma 2 we get

$$\begin{aligned} V^{\pi^{t+1}}(s) - V^{\pi^t}(s) &\geq \frac{1}{2}\left(\mathcal{T}^* V^{\pi^t}(s) - V^{\pi^t}(s)\right) + \frac{\varepsilon}{c} \\ &\geq \frac{1}{2}\left(\mathcal{T}^* V^{\pi^t}(s) - V^*(s) + V^*(s) - V^{\pi^t}(s)\right) + \frac{\varepsilon}{c} \\ &\geq \frac{1}{2}(1-\gamma)\left(V^*(s) - V^{\pi^t}(s)\right) + \frac{\varepsilon}{c}. \end{aligned}$$

This implies

$$\begin{aligned} \left\| V^* - V^{\pi^t} \right\|_\infty &\leq \left(\frac{1+\gamma}{2}\right)^t \left\| V^* - V^{\pi^0} \right\|_\infty + \sum_{\ell=1}^{t} \left(\frac{\varepsilon}{c}\right)^\ell \\ &= \left(\frac{1+\gamma}{2}\right)^t \left\| V^* - V^{\pi^t} \right\|_\infty + \varepsilon, \end{aligned}$$

where the last step follows our definition that $c = 2(1-\gamma)^{-1}$.

References I

J. Bhandari and D. Russo. On the linear convergence of policy gradient methods for finite mdps. In Proceedings of the 24th International Conference on Artificial Intelligence and Statistics, pages 2386–2394, 2021.

S. Bubeck. Convex optimization: Algorithms and complexity. Foundations and Trends in Machine Learning, 8(3-4):231–357, 2015.

M. L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 2014.