Natural Policy Gradient

Carlo Alfano

November 17, 2020

Setting

A (finite) Markov Decision Process (MDP) $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho)$ is specified by:

- a finite state space $\mathcal{S}$, with cardinality $S = |\mathcal{S}|$;
- a finite action space $\mathcal{A}$, with cardinality $A = |\mathcal{A}|$;
- a transition model P, where $P(s'|s, a)$ is the probability of going from state s to state s' after taking action a;
- a reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, where $r(s, a)$ is the reward obtained in state s after taking action a;
- a discount factor $\gamma \in [0, 1)$;
- a starting state distribution ρ over $\mathcal{S}$.

Policies

- Deterministic policies: $\pi : \mathcal{S} \to \mathcal{A}$, $a_t = \pi(s_t)$.
- Stochastic policies: $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, $a_t \sim \pi(\cdot|s_t)$.

A policy induces a distribution over trajectories $\tau = (s_t, a_t, r_t)_{t=0}^{\infty}$, where s_0 is drawn from the starting state distribution ρ, and, for all subsequent time steps t, $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$.

Value Functions

The value function $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined as the discounted sum of future rewards starting at state s and executing π, i.e.

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

where the expectation is with respect to the randomness of the trajectory τ induced by π in M. Since we assume that $r(s, a) \in [0, 1]$, we have $0 \leq V^\pi(s) \leq \frac{1}{1-\gamma}$.

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

.

Action - Value Functions

The action-value (or Q-value) function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and the advantage function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are defined as:

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

$$A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s).$$

It is easy to notice that for random policies:

$$V^\pi(s) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} Q^\pi(s,a)$$

General Goal

The goal of the agent is to find a policy π that maximizes the expected value from the initial state, i.e. the optimization problem the agent seeks to solve is:

$$\max_{\pi} V^{\pi}(\rho),$$

where the max is over all policies. The famous theorem of Bellman and Dreyfus [1959] shows there exists a policy $\pi^{\star}$ which simultaneously maximizes $V^{\pi}(s_0)$, for all states $s_0 \in \mathcal{S}$.

Current Goal

This presentation regards gradient ascent methods for the optimization problem:

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\rho)$$

where $\{\pi_\theta | \theta \in \Theta\}$ is some class of parametric (stochastic) policies. We consider the softmax parametrization: for unconstrained $\theta \in \mathbb{R}^{|S||A|}$,

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$$

This policy class is complete, meaning that any stochastic policy can be represented in this class.

Utilities

Discounted state visitation distribution $d^{\pi}_{s_0}(s)$ of a policy π:

$$d^{\pi}_{s_0}(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t Pr^{\pi}(s_t = s|s_0)$$

where $Pr^{\pi}(s_t = s|s_0)$ is the state visitation probability that $s_t = s$ following policy π.

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

The policy gradient functional form is then:

$$\begin{aligned}\nabla_\theta V^{\pi_\theta}(s_0) &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s,a)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) A^{\pi_\theta}(s,a)\right]\end{aligned}$$

More Utilities

A useful lemma is the following (performance difference lemma). For all policies π, π' and states s_0,

$$V^{\pi}(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}_{s_0}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^{\pi'}(s,a)\right]$$

Natural Policy Gradient

The NPG algorithm defines a Fisher information matrix (induced by π), and performs gradient updates in the geometry induced by this matrix as follows:

$$F_\rho^\theta = \mathbb{E}_{s \sim d_{s_0}^\pi} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) (\nabla_\theta \log \pi_\theta(a|s))^\top \right]$$

$$\theta^{(t+1)} = \theta^{(t)} + \eta \left(F_\rho^{\theta^{(t)}} \right)^\dagger \nabla_\theta V^{(t)}(\rho)$$

where $M^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix M.

Simple update

Lemma

For the softmax parameterization, the NPG update takes the form:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$

$$\pi^{(t+1)}(a|s) = \pi^{(t)}(a|s) \frac{exp(\eta A^{(t)}(s,a)/(1-\gamma))}{Z_t(s)}$$

where $Z_t(s) = \sum_{a \in \mathcal{A}} \pi^{(t)}(a|s) exp(\eta A^{(t)}(s,a)/(1-\gamma))$ *is a normalizing factor.*

Proof

By the definition of the pseudoinverse, we have that $\left(F_\rho^{\theta^{(t)}}\right)^\dagger \nabla_\theta V^{(t)}(\rho) = w_\star$ if and only if

$$w_\star = \operatorname*{argmin}_w ||\nabla_\theta V^{(t)}(\rho) - F_\rho^{\theta^{(t)}} w||^2$$

Let us first evaluate $F_\rho^{\theta^{(t)}} w$. For the softmax policy parametrization, it is easy to see that

$$\frac{\partial \log \pi_\theta(a|s)}{\partial \theta_{s',a'}} = \mathbb{I}[s = s'] \left(\mathbb{I}[a = a'] - \pi_\theta(a'|s)\right)$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Proof continues

Then

$$w^\top \nabla_\theta \log \pi_\theta(a|s) = w_{s,a} - \sum_{a' \in \mathcal{A}} w_{s,a'} \pi_\theta(a'|s) := w_{s,a} - \overline{w}_s$$

So

$$\begin{aligned} F_\rho^\theta w &= \mathbb{E}_{s \sim d_\rho^\pi} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) (w^\top \nabla_\theta \log \pi_\theta(a|s)) \right] \\ &= \mathbb{E}_{s \sim d_{s_0}^\pi} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[\nabla_\theta \log \pi_\theta(a|s) (w_{s,a} - \overline{w}_s) \right] \\ &= \mathbb{E}_{s \sim d_{s_0}^\pi} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[w_{s,a} \nabla_\theta \log \pi_\theta(a|s) \right] \end{aligned}$$

Proof ends

Looking at a single element of the vector:

$$\left(F_\rho^\theta w\right)_{s',a'} = \mathbb{E}_{s\sim d_{s_0}^\pi}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[w_{s,a}\frac{\partial \log \pi_\theta(a|s)}{\partial_{s',a'}}\right]$$
$$= d_{s_0}^\pi(s')\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[w_{s',a}\left(\mathbb{I}[a=a'] - \pi_\theta(a'|s')\right)\right]$$
$$= d_{s_0}^\pi(s')\pi_\theta(a'|s')w_{s',a'} - d_{s_0}^\pi(s')\pi_\theta(a'|s')\mathbb{E}_{a\sim\pi_\theta(\cdot|s')}\left[w_{s',a}\right]$$
$$= d_{s_0}^\pi(s')\pi_\theta(a'|s')\left(w_{s',a'} - \overline{w}_{s'}\right)$$

Which means that

$$||\nabla_\theta V^{\pi_\theta}(\rho) - F_\rho^\theta w||^2 =$$
$$= \sum_{s,a}\Bigg(d_{s_0}^\pi(s)\pi_\theta(a|s)\cdot$$
$$\left(\frac{1}{1-\gamma}A^{\pi_\theta}(s,a) - w_{s,a} + \sum_{a'\in\mathcal{A}} w_{s,a'}\pi_\theta(a'|s)\right)\Bigg)^2$$

Global Convergence

Theorem

Suppose we run the NPG update using $\rho \in \Delta(\mathcal{S})$ and with $\theta^{(0)} = 0$. Fix $\eta > 0$. For all $T > 0$, we have:

$$V^{(T)}(\rho) \geq V^\star(\rho) - \frac{\log |\mathcal{A}|}{\eta T} - \frac{1}{(1-\gamma)^2 T}$$

If we set $\eta \geq (1-\gamma)^2 \log |\mathcal{A}|$, we see that NPG finds an ε-optimal policy in a number of iterations that is at most:

$$T \leq \frac{2}{(1-\gamma)^2 \varepsilon}$$

Improvement lower bound

Lemma

For the iterates $\pi^{(t)}$ generated by the NPG updates, we have for all starting state distributions ρ:

$$V^{(t+1)}(\rho) - V^{(t)}(\rho) \geq \frac{1-\gamma}{\eta} \mathbb{E}_{s\sim\rho} \log Z_t(s) \geq 0$$

Proof

We first show that $\log Z_t(s) \geq 0$.

$$\begin{aligned}
\log Z_t(s) &= \log \sum_{a \in \mathcal{A}} \pi^{(t)}(a|s) \exp(\eta A^{(t)}(s,a)/(1-\gamma)) \\
\text{(Jensen)} &\geq \sum_{a \in \mathcal{A}} \pi^{(t)}(a|s) \log \exp(\eta A^{(t)}(s,a)/(1-\gamma)) \\
&= \frac{\eta}{1-\gamma} \sum_{a \in \mathcal{A}} \pi^{(t)}(a|s) A^{(t)}(s,a) \\
&= 0
\end{aligned}$$

Proof ends

By the performance difference lemma:

$$
\begin{aligned}
V^{(t+1)}(\rho) - V^{(t)}(\rho) = \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) A^{(t)}(s,a) \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d_\rho^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) \log \frac{\pi^{(t+1)}(a|s) Z_t(s)}{\pi^{(t)}(a|s)} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d_\rho^{(t+1)}} KL(\pi^{(t+1)}||\pi^{(t)}) + \frac{1}{\eta}\mathbb{E}_{s\sim d_{s_0}^{(t+1)}} \log Z_t(s) \\
&\geq \frac{1}{\eta}\mathbb{E}_{s\sim d_{s_0}^{(t+1)}} \log Z_t(s) \geq \frac{1-\gamma}{\eta}\mathbb{E}_{s\sim\rho} \log Z_t(s)
\end{aligned}
$$

where we used that $d_\rho^{(t+1)} \geq (1-\gamma)\rho$ and that $\log Z_t(s) \geq 0$

Proof of global convergence

$$
\begin{aligned}
V^\star(\rho) - V^{(t)}(\rho) = \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star}\sum_a \pi^\star(a|s)A^{(t)}(s,a) \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\sum_a \pi^\star(a|s)\log\frac{\pi^{(t+1)}(a|s)Z_t(s)}{\pi^{(t)}(a|s)} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star}\left(KL(\pi^\star||\pi^{(t)}) - KL(\pi^\star||\pi^{(t+1)}) + \log Z_t(s)\right)
\end{aligned}
$$

where $d^\star = d_\rho^{\pi^\star}$.

Proof of global convergence continues

Using the previous lemma:

$$\frac{1}{\eta}\mathbb{E}_{s\sim d^\star} \log Z_t(s) \leq \frac{1}{1-\gamma}\left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\right)$$

So

$$V^{\pi^\star}(\rho) - V^{(T-1)}(\rho) \leq \frac{1}{T}\sum_{t=0}^{T-1}\left(V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right)$$

$$= \frac{1}{\eta T}\sum_{t=0}^{T-1}\mathbb{E}_{s\sim d^\star}\left(KL(\pi^\star||\pi^{(t)}) - KL(\pi^\star||\pi^{(t+1)}) + \log Z_t(s)\right)$$

$$\leq \frac{\mathbb{E}_{s\sim d^\star} KL(\pi^\star||\pi^{(0)})}{\eta T} + \frac{1}{(1-\gamma)T}\sum_{t=0}^{T-1}\left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\right)$$

Proof of global convergence ends

$$
\begin{aligned}
&= \frac{\mathbb{E}_{s\sim d^\star} KL(\pi^\star || \pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T} \\
&\leq \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T}
\end{aligned}
$$

Estimated gradients

So far we have used exact gradients. What happens if we use an estimate for gradients? Let's look at the update:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma} A^{(t)}$$

$$\pi^{(t+1)}(a|s) = \pi^{(t)}(a|s) \frac{\exp(\eta A^{(t)}(s,a)/(1-\gamma))}{Z_t(s)}$$

where $Z_t(s) = \sum_{a \in \mathcal{A}} \pi^{(t)}(a|s) \exp(\eta A^{(t)}(s,a)/(1-\gamma))$.

Estimated gradients

So far we have used exact gradients. What happens if we use an estimate for gradients? Let's look at the update:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{\eta}{1-\gamma}\widehat{A}^{(t)}$$

$$\pi^{(t+1)}(a|s) = \pi^{(t)}(a|s)\frac{\exp(\eta\widehat{A}^{(t)}(s,a)/(1-\gamma))}{\widehat{Z}_t(s)}$$

where $\widehat{Z}_t(s) = \sum_{a\in\mathcal{A}} \pi^{(t)}(a|s)\exp(\eta\widehat{A}^{(t)}(s,a)/(1-\gamma))$.

Sampler

Suppose we have access to a simulator of the MDP. Then we can define an unbiased sampler for A^π. Remember that:

$$\begin{aligned} A^\pi(s,a) &= Q^\pi(s,a) - V^\pi(s) \\ &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \middle| \pi, s_0 = s, a_0 = a\right] \\ &- \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t) \middle| \pi, s_0 = s\right] \end{aligned}$$

Sampler

Define:

$$\widehat{Q}^{\pi}(s,a) := \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \qquad \text{with } s_0 = s, a_0 = a$$

$$\widehat{V}^{\pi}(s) := \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \qquad \text{with } s_0 = s$$

$$\widehat{A}^{\pi}(s,a) := \widehat{Q}^{\pi}(s,a) - \widehat{V}^{\pi}(s)$$

where $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. Then $\widehat{A}^{\pi}(s,a)$ is by definition an unbiased estimate of $A^{\pi}(s,a)$.

Error bound

The following Bernstein type bounds hold:

$$P\left(\frac{1}{N}\sum_{n=1}^{N}\widehat{A}_n^{\pi}(s,a) - A^{\pi}(s,a) \geq -\varepsilon\right) \geq 1 - \exp\left(-\frac{3N\varepsilon^2(1-\gamma)}{6+4\varepsilon}\right)$$

$$P\left(A^{\pi}(s,a) - \frac{1}{N}\sum_{n=1}^{N}\widehat{A}_n^{\pi}(s,a) \geq -\varepsilon\right) \geq 1 - \exp\left(-\frac{3N\varepsilon^2(1-\gamma)}{6+4\varepsilon}\right)$$

Proof

We have that if a random variable X satisfies the Bernstein one-sided condition:

$$\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq \exp\left(\frac{\lambda^2 \mathrm{Var} X}{2(1-b\lambda)}\right) \qquad \forall \lambda \in [0, 1/b)$$

then, for $X_1, \ldots, X_N \sim X$ i.i.d. and for any $\varepsilon \geq 0$:

$$P\left(\frac{1}{N}\sum_{n=1}^{N} X_n - \mathbb{E}X \geq \varepsilon\right) \leq \exp\left(-\frac{N\varepsilon^2}{2(\mathrm{Var} X + b\varepsilon)}\right)$$

Proof continues

But if $X - \mathbb{E}X \leq c$ a.s. for a given $c > 0$, then X satisfies the one-sided Bernstein condition with parameter $b = c/3$. So if we consider the random variable $-\widehat{A}^{(t)}(s,a)$, we have $-\widehat{A}^{\pi}(s,a) - (-A^{\pi}(s,a)) \leq \frac{2}{1-\gamma}$ and that $-\widehat{A}^{(t)}(s,a)$ satisfies the one-sided Bernstein condition with parameter $b = \frac{2}{3(1-\gamma)}$.

We can also bound the variance of $-\widehat{A}^{\pi}(s,a)$ using Popoviciu's inequality: $\mathrm{Var}(-\widehat{A}^{(t)}(s,a)) \leq \frac{1}{1-\gamma}$.

Proof continues

If $\widehat{A}^{\pi}_1(s,a),\ldots,\widehat{A}^{\pi}_N(s,a) \sim \widehat{A}^{\pi}(s,a)$, then for $\varepsilon \geq 0$

$$
\begin{aligned}
& P\left(\frac{1}{N}\sum_{n=1}^{N}\left(-\widehat{A}^{\pi}_n(s,a)\right) - (-A^{\pi}(s,a)) \geq \varepsilon\right) \\
& \leq \exp\left(-\frac{N\varepsilon^2}{2(\mathrm{Var}X + \frac{2}{3(1-\gamma)}\varepsilon)}\right) \\
& \leq \exp\left(-\frac{N\varepsilon^2}{2(\frac{1}{1-\gamma} + \frac{2}{3(1-\gamma)}\varepsilon)}\right) \\
& = \exp\left(-\frac{3N\varepsilon^2(1-\gamma)}{6+4\varepsilon}\right)
\end{aligned}
$$

Proof ends

So

$$
\begin{aligned}
&P\left(\frac{1}{N}\sum_{n=1}^{N}\widehat{A}_n^{\pi}(s,a)-A^{\pi}(s,a)\geq -\varepsilon\right)\\
&=P\left(\frac{1}{N}\sum_{n=1}^{N}\left(-\widehat{A}_n^{\pi}(s,a)\right)-(-A^{\pi}(s,a))\leq \varepsilon\right)\\
&=1-P\left(\frac{1}{N}\sum_{n=1}^{N}\left(-\widehat{A}_n^{\pi}(s,a)\right)-(-A^{\pi}(s,a))\geq \varepsilon\right)\\
&\geq 1-\exp\left(-\frac{3N\varepsilon^2(1-\gamma)}{6+4\varepsilon}\right)
\end{aligned}
$$

The proof for the second inequality is the same

Back to global convegence

Let's see the proof again and modify it. Denote:

$$\delta(\varepsilon) = 1 - \exp\left(-\frac{3N\varepsilon^2(1-\gamma)}{6+4\varepsilon}\right)$$

and

$$\widehat{A}^{\pi}(s,a) = \frac{1}{N}\sum_{n=1}^{N} \widehat{A}_n^{\pi}(s,a)$$

Back to Improvement lower bound

Lemma

For the iterates $\pi^{(t)}$ generated by the NPG updates, we have for all starting state distributions ρ and with probability at least $\delta(\varepsilon)$:

$$V^{(t+1)}(\rho) - V^{(t)}(\rho) \geq \frac{1-\gamma}{\eta} \mathbb{E}_{s \sim \rho} \log \widehat{Z}_t(s) - \frac{\varepsilon}{1-\gamma} \geq -\frac{2-\gamma}{1-\gamma} \varepsilon$$

Proof

We first show that $\log \widehat{Z}_t(s) \geq -\frac{\eta\varepsilon}{1-\gamma}$.

$$
\begin{aligned}
\log \widehat{Z}_t(s) &= \log \sum_{a\in\mathcal{A}} \pi^{(t)}(a|s)\exp(\eta \widehat{A}^{(t)}(s,a)/(1-\gamma)) \\
(\text{Jensen}) &\geq \sum_{a\in\mathcal{A}} \pi^{(t)}(a|s) \log \exp(\eta \widehat{A}^{(t)}(s,a)/(1-\gamma)) \\
&= \frac{\eta}{1-\gamma} \sum_{a\in\mathcal{A}} \pi^{(t)}(a|s) \widehat{A}^{(t)}(s,a) \\
&= \frac{\eta}{1-\gamma} \sum_{a\in\mathcal{A}} \pi^{(t)}(a|s) A^{(t)}(s,a) + \\
&+ \frac{\eta}{1-\gamma} \sum_{a\in\mathcal{A}} \pi^{(t)}(a|s) \left(\widehat{A}^{(t)}(s,a) - A^{(t)}(s,a) \right) \\
&\geq -\frac{\eta\varepsilon}{1-\gamma}
\end{aligned}
$$

Proof continues

By the performance difference lemma:

$$
\begin{aligned}
&V^{(t+1)}(\rho) - V^{(t)}(\rho) = \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) A^{(t)}(s,a) \\
&\geq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) \widehat{A}^{(t)}(s,a) - \frac{\varepsilon}{1-\gamma} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d_\rho^{(t+1)}} \sum_a \pi^{(t+1)}(a|s) \log \frac{\pi^{(t+1)}(a|s)\widehat{Z}_t(s)}{\pi^{(t)}(a|s)} - \frac{\varepsilon}{1-\gamma} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d_\rho^{(t+1)}} KL(\pi^{(t+1)}||\pi^{(t)}) + \frac{1}{\eta}\mathbb{E}_{s\sim d_{s_0}^{(t+1)}} \log \widehat{Z}_t(s) - \frac{\varepsilon}{1-\gamma}
\end{aligned}
$$

Proof ends

$$
\begin{aligned}
&= \frac{1}{\eta}\mathbb{E}_{s\sim d_\rho^{(t+1)}} KL(\pi^{(t+1)}||\pi^{(t)}) + \frac{1}{\eta}\mathbb{E}_{s\sim d_{s_0}^{(t+1)}} \log \widehat{Z}_t(s) - \frac{\varepsilon}{1-\gamma} \\
&\geq \frac{1}{\eta}\mathbb{E}_{s\sim d_{s_0}^{(t+1)}} \log \widehat{Z}_t(s) - \frac{\varepsilon}{1-\gamma} \\
&\geq \frac{1-\gamma}{\eta}\mathbb{E}_{s\sim\rho} \log \widehat{Z}_t(s) - \frac{\varepsilon}{1-\gamma} \\
&\geq -\frac{1-\gamma}{\eta}\frac{\eta\varepsilon}{1-\gamma} - \frac{\varepsilon}{1-\gamma} \\
&= -\frac{2-\gamma}{1-\gamma}\varepsilon
\end{aligned}
$$

where we used that $d_\rho^{(t+1)} \geq (1-\gamma)\rho$.

Proof of global convergence

$$
\begin{aligned}
&V^\star(\rho) - V^{(t)}(\rho) = \\
&= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\star} \sum_a \pi^\star(a|s) A^{(t)}(s,a) \\
&\leq \frac{1}{\eta}\mathbb{E}_{s\sim d^\star} \sum_a \pi^\star(a|s) \log \frac{\pi^{(t+1)}(a|s)\widehat{Z}_t(s)}{\pi^{(t)}(a|s)} + \frac{\varepsilon}{1-\gamma} \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim d^\star} \left(KL(\pi^\star||\pi^{(t)}) - KL(\pi^\star||\pi^{(t+1)}) + \log \widehat{Z}_t(s) \right) + \frac{\varepsilon}{1-\gamma}
\end{aligned}
$$

where $d^\star = d_\rho^{\pi^\star}$.

Proof of global convergence continues

Using the previous lemma:

$$\frac{1}{\eta}\mathbb{E}_{s\sim d^\star} \log \widehat{Z}_t(s) \leq \frac{1}{1-\gamma}\left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star)\right) + \frac{\varepsilon}{(1-\gamma)^2}$$

So

$$\min_{t\leq T}\left\{V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right\} \leq \frac{1}{T}\sum_{t=0}^{T-1}\left(V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right)$$

$$\leq \frac{1}{\eta T}\sum_{t=0}^{T-1}\mathbb{E}_{s\sim d^\star}\left(KL(\pi^\star||\pi^{(t)}) - KL(\pi^\star||\pi^{(t+1)}) + \log \widehat{Z}_t(s)\right)$$

$$+ \frac{\varepsilon}{1-\gamma}$$

Proof of global convergence ends

$$\leq \frac{\mathbb{E}_{s\sim d^\star} KL(\pi^\star || \pi^{(0)})}{\eta T} + \frac{1}{(1-\gamma)T} \sum_{t=0}^{T-1} \left(V^{(t+1)}(d^\star) - V^{(t)}(d^\star) \right)$$

$$+ \frac{\varepsilon}{(1-\gamma)^2} + \frac{\varepsilon}{1-\gamma}$$

$$= \frac{\mathbb{E}_{s\sim d^\star} KL(\pi^\star || \pi^{(0)})}{\eta T} + \frac{V^{(T)}(d^\star) - V^{(0)}(d^\star)}{(1-\gamma)T} + \frac{2-\gamma}{(1-\gamma)^2}\varepsilon$$

$$\leq \frac{\log |\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T} + \frac{2-\gamma}{(1-\gamma)^2}\varepsilon$$

Global convergence with estimated gradients

In conclusion we have the following.

Theorem

Suppose we run the NPG update with $\theta^{(0)} = 0$. Fix $\eta > 0$. For all $T > 0$, given the sampler $\widehat{A}^{\pi}$ previously defined, with probability at least $1 - exp\left(-\frac{3N\varepsilon^2(1-\gamma)}{6+4\varepsilon}\right)$, we have:

$$\min_{t \leq T}\left\{V^{\pi^\star}(\rho) - V^{(t)}(\rho)\right\} \leq \frac{\log|\mathcal{A}|}{\eta T} + \frac{1}{(1-\gamma)^2 T} + \frac{2-\gamma}{(1-\gamma)^2}\varepsilon$$

Generalization

What if the $\mathcal{S}$ and A are too big? Then we consider a restricted policy class

$$\left\{\pi_\theta | \theta \in \mathbb{R}^d\right\}$$

with $d \ll |\mathcal{S}||\mathcal{A}|$. The parameter update becomes a minimization problem:

$$\left(F_\rho^{\theta^{(t)}}\right)^\dagger \nabla_\theta V^{(t)}(\rho) = \frac{1}{1-\gamma} w^\star$$

where

$$w^\star \in \operatorname*{argmin}_w \mathbb{E}_{s \sim d_\rho^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} \left[\left(w^\top \nabla_\theta \pi_\theta(\cdot|s) - A^{\pi_\theta}(s,a) \right)^2 \right]$$