ARTIFICIAL INTELLIGENCE

Why does the "reward to go" trick in policy gradient methods work?

Asked 6 years, 3 months ago Modified 1 year, 3 months ago Viewed 7k times

9

In the policy gradient method, there's a trick to reduce the variance of policy gradient. We use causality, and remove part of the sum over rewards so that only actions happened after the reward are taken into account (See here http://rail.eecs.berkeley.edu/deeprlcourse/static/slides/lec-5.pdf, slide 18).

Why does it work? I understand the intuitive explanation, but what's the rigorous proof of it? Can you point me to some papers?

reinforcement-learning math policy-gradients rewards reward-to-go

Share Improve this question Follow

edited Oct 10, 2020 at 15:52
nbro
42.3k 12 119 213

asked Dec 20, 2018 at 1:00
Konstantin Solomatov
298 2 10

3 Answers

Sorted by: Highest score (default)

12

An important thing we're going to need is what is called the "Expected Grad-Log-Prob Lemma here" (proof included on that page), which says that (for any t):

$$\mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \right] = 0.$$

Taking the analytical expression of the gradient (from, for example, slide 9) as a starting point:

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right)\left(\sum_{t=1}^{T} r(s_t,a_t)\right)\right] \\ &= \sum_{t=1}^{T}\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=1}^{T} r(s_{t'},a_{t'})\right] \\ &= \sum_{t=1}^{T}\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=1}^{t-1} r(s_{t'},a_{t'}) + \nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=t}^{T} r(s_{t'},\right. \\ &= \sum_{t=1}^{T}\left(\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=1}^{t-1} r(s_{t'},a_{t'})\right]\right. \\ &\quad \left.+\,\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=t}^{T} r(s_{t'},a_{t'})\right]\right)\end{aligned}$$

At the t^{th} "iteration" of the outer sum, the random variables $\sum_{t'=1}^{t-1} r(s_{t'},a_{t'})$ and $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$ are independent (we assume, by definition, the action only depends on the most recent state), which means we are allowed to split the expectation:

$$\begin{aligned}\nabla_\theta J(\theta) = \sum_{t=1}^{T}\Bigg(&\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\sum_{t'=1}^{t-1} r(s_{t'},a_{t'})\right]\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right] \\ &+\,\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=t}^{T} r(s_{t'},a_{t'})\right]\Bigg)\end{aligned}$$

The first expectation can now be replaced by 0 due to the lemma mentioned at the top of the post:

$$\begin{aligned}\nabla_\theta J(\theta) &= \sum_{t=1}^{T}\mathbb{E}_{\tau\sim\pi_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\sum_{t'=t}^{T} r(s_{t'},a_{t'})\right] \\ &= \mathbb{E}_{\tau\sim\pi_\theta(\tau)}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\left(\sum_{t'=t}^{T} r(s_{t'},a_{t'})\right).\end{aligned}$$

The expression on slide 18 of the linked slides is an unbiased, sample-based estimator of this gradient:

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t} \mid s_{i,t})\left(\sum_{t'=t}^{T} r(s_{i,t'},a_{i,t'})\right)$$

For a more formal treatment of the claim that we can pull $\sum_{t'=1}^{t-1} r(s_{t'},a_{t'})$ out of an expectation due to the Markov property, see this page:

https://spinningup.openai.com/en/latest/spinningup/extra_pg_proof1.html

Share Improve this answer Follow

edited Nov 25, 2020 at 20:14
Mike Land
103 3

answered Feb 2, 2019 at 19:16
Dennis Soemers ♦
10.5k 2 29 70

1 Why is the form of "Expected Grad-Log-Prob Lemma" mentioned at the top of the post legit? It is different than the form of the proof shown in the link (the proof in the link has no problem). Your lemma is an expectation over **trajectories**. – starriet 차주녕 Jan 4, 2022 at 3:09

@starriet For the individual choice of a single action at one point in time within a trajectory, the trajectory as a whole has no influence (our policy does not look at the history or the future, just at the current state), so $\mathbb{E}_{\tau \sim \pi_\theta(\tau)} [\nabla_\theta \log \pi_\theta(a_t \mid s_t)] = \mathbb{E}_{a_t \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a_t \mid s_t)]$, where the right-hand side more closely matches that notation from the link. – Dennis Soemers ♦ Jan 4, 2022 at 9:43

@starriet However, we use the trajectory-based notation because in the larger equation, we have these sums over all time steps, and that's the part where the full trajectory actually becomes relevant. – Dennis Soemers ♦ Jan 4, 2022 at 9:45

Thanks, I think I got your point. If I understood correctly, the right-hand side is at a *specific state*, and the left-hand side is at the various states and actions of *all possible trajectories*, right? Since the RHS is 0 and it's about one state, LHS is also 0 because LHS is an expectation over all states (this is how I understood. please let me know if I'm wrong). – starriet 차주녕 Jan 4, 2022 at 12:41

But still, It's difficult to understand why the form of the LHS is legit. It's an expectation over trajectories, but the term inside of the square bracket is a function of action and state. Given that a trajectory is composed of many states and actions, how the expectation over trajectories is possible if the term inside of the brackets is smaller than a trajectory? – starriet 차주녕 Jan 4, 2022 at 12:51

1

A correct proof is given at https://spinningup.openai.com/en/latest/spinningup/extra_pg_proof1.html. It uses, among other techniques, the probability chain rule.

On the contrary, an attempt to prove "reward to go" by applying $X, Y \; independent \implies E[XY] = E[X]E[Y]$ does not work.

The random variables $\sum_{t'=1}^{t-1} r(s_{t'}, a_{t'})$ and $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$ are generally not independent. They would be independent, if $P[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \mid \sum_{t'=1}^{t-1} r(s_{t'}, a_{t'})] = P[\nabla_\theta \log \pi_\theta(a_t \mid s_t)]$ but this does not follow from $\pi_\theta(a_t \mid s_t) = \pi_\theta(a_t \mid s_t, s_{t-1}, \ldots, s_1, a_{t-1}, \ldots, a_1)$

Consider the following setup as an example for the dependency between the two random variables in question.

times = {1, 2, 3, ..., T}

states = {1, 2, 3, ..., T}

actions = {0, 1}

$\pi(a_t = 0 \mid s_t) = \frac{1}{s_t}$

$r(s_t, a_t) = 1$ i.e. the event reward at time t equals 1 and the total reward from time 1 to time t-1 equals t-1

$P(s_1 = 1) = 1$ i.e. all trajectories start in state 1

$P(s_{t+1} = s_t + 1 \mid s_t, a_t) = 1$ i.e. the next state, given the current state and the current action, is always the current state + 1

As a result, the higher time t, the higher the state s_t, the higher the probability for action $a_t = 1$ and the higher the total reward from time 0 to time $t - 1$, i.e. the values of $\pi(a_t \mid s_t)$ and thus the values of $\nabla_\theta \log \pi(a_t \mid s_t)$ are highly correlated with the total rewards from time 0 to time $t - 1$, hence respective random variables are not independent of each other.

Share Improve this answer Follow

answered Sep 27, 2022 at 14:36

blubb3456
11 1

0

By linearity of expection and using the distributive property:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t) \right) \left(\sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \right) \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_\theta \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) \sum_{t'=1}^{T} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) \right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_\theta \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) \sum_{t'=1}^{t-1} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) + \nabla_\theta \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) \sum_{t'=t}^{T} r(\mathbf{s}_{t'} \right. \\
&= \sum_{t=1}^{T} \sum_{t'=1}^{t-1} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t)\, r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) \right] + \\
&\quad \sum_{t=1}^{T} \mathbb{E}_{\tau \sim p_{\boldsymbol{\theta}}(\tau)} \left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\mathbf{a}_t \mid \mathbf{s}_t) \sum_{t'=t}^{T} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) \right]
\end{aligned}$$

and the first term can be further expanded as follows:

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{t'=1}^{t-1}\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,r(\mathbf{s}_{t'},\mathbf{a}_{t'})\right]\\
=&\sum_{t=1}^{T}\sum_{t'=1}^{t-1}\mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t,\mathbf{s}_{t'},\mathbf{a}_{t'})\sim p_{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t,\mathbf{s}_{t'},\mathbf{a}_{t'})}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,r(\mathbf{s}_{t'},\mathbf{a}_{t'})\right]\\
=&\sum_{t=1}^{T}\sum_{t'=1}^{t-1}\mathbb{E}_{(\mathbf{s}_{t'},\mathbf{a}_{t'})\sim p_{\boldsymbol{\theta}}(\mathbf{s}_{t'},\mathbf{a}_{t'})}\left[\mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim p_{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t\mid\mathbf{s}_{t'},\mathbf{a}_{t'})}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,r(\mathbf{s}_{t'},\mathbf{a}_{t'})\right]\right]\\
=&\sum_{t=1}^{T}\sum_{t'=1}^{t-1}\mathbb{E}_{(\mathbf{s}_{t'},\mathbf{a}_{t'})\sim p_{\boldsymbol{\theta}}(\mathbf{s}_{t'},\mathbf{a}_{t'})}\left[r(\mathbf{s}_{t'},\mathbf{a}_{t'})\,\mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim p_{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t\mid\mathbf{s}_{t'},\mathbf{a}_{t'})}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\right]\right]
\end{aligned}
$$

The innermost expectation can be broken down further into

$$
\begin{aligned}
&\mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim p_{\boldsymbol{\theta}}(\mathbf{s}_t,\mathbf{a}_t\mid\mathbf{s}_{t'},\mathbf{a}_{t'})}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\right]\\
=&\int\int\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\;p(\mathbf{s}_t\mid\mathbf{s}_{t'},\mathbf{a}_{t'})\,d\mathbf{a}_t\;d\mathbf{s}_t\\
=&\int p(\mathbf{s}_t\mid\mathbf{s}_{t'},\mathbf{a}_{t'})\int\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,d\mathbf{a}_t\;d\mathbf{s}_t
\end{aligned}
$$

where $\int\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\,d\mathbf{a}_t=\mathbb{E}_{\mathbf{a}_t\sim\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\right]=0$ due to the EGLP lemma

As a result, we have

$$
\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})=\sum_{t=1}^{T}\mathbb{E}_{\tau\sim p_{\boldsymbol{\theta}}(\tau)}\left[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\mathbf{a}_t\mid\mathbf{s}_t)\sum_{t'=t}^{T}r(\mathbf{s}_{t'},\mathbf{a}_{t'})\right]
$$

Share Improve this answer Follow edited Jan 5, 2024 at 15:52 answered Jan 5, 2024 at 15:45

Jus Forever
1 1