Importance sampling in off-policy learning

Pierre-Luc Bacon

June 17, 2015

Importance Sampling

IS considers the case where samples come from a different sampling distribution $q(s)$. We use the following simple identity:

$$
\begin{aligned}
J &= \mathbb{E}_{s\sim p}\{f(s)\} \\
&= \sum_{s\in\mathcal{S}} p(s)f(s) \\
&= \sum_{s\in\mathcal{S}} q(s)\frac{p(s)}{q(s)}f(s) \\
&= \mathbb{E}_{s\sim q}\left\{\frac{p(s)}{q(s)}f(s)\right\} \\
&\approx \frac{1}{N}\sum_{i=1}^{N} f(s_i)\frac{p(s_i)}{q(s_i)}
\end{aligned}
$$

In the continuous world, we would talk about a *change of measure*.

$\rho(s) = \frac{p(s)}{q(s)}$ is called the **likelihood ratio**, $q(s)$ is the **importance sampling distribution** and $p(s)$ is the **nominal distribution**.

Normalized Importance Sampling

Also called *self-normalized IS* or *weighted IS*.

Note:

$$\rho(s) = \mathbb{E}_{s\sim q}\left\{\rho(s)\right\} = \sum_{s\in\mathcal{S}} q(s)\frac{p(s)}{q(s)} = 1$$

Then:

$$\begin{aligned} J &= \frac{\mathbb{E}_{s\sim q}\left\{\rho(s)f(s)\right\}}{\mathbb{E}_{s\sim q}\left\{\rho(s)\right\}} \\ &\approx \frac{\sum_{i=1}^{N} f(s_i)\frac{p(s_i)}{q(s_i)}}{\sum_{i=1}^{N}\rho_i} \end{aligned}$$

Useful if only know $\tilde{p}(s) = cp(s)$ and $\tilde{q}(s) = bq(s)$.

Properties

- Normalized IS converges with pr. 1 to the true value, in the limit.
- It is however unbiased: the expected difference with the true value might not be zero.

Optimal sampling density

Let's pretend that we don't have any of the usual RL constraints. What if we wanted to find the IS sampling distribution yielding the minimum variance ?

$$q^*(s) = \arg\min_q \mathrm{Var}(f(s)\rho(s)) = \frac{|f(s)|p(s)}{\sum_{s\in\mathcal{S}} |f(s)|p(s)}$$

In the normalized case, we get $q^*(s) \propto |f(s) - J|p(x)$.

Cross entropy method

Using the previous result, we could also try to *push* $q(s)$ closer to $f(s)p(s)$ (assuming that f(s) is non-negative). We want to minimize:

$$
\begin{aligned}
&D_{\mathrm{KL}}(f(s)p(s); q_\theta(s)) \\
&= \sum_s f(s)p(s) \log \frac{f(s)p(s)}{q_\theta(s)} \\
&= \sum_s f(s)p(s) \left(\log f(s)p(s) - \log q_\theta(s)\right) \\
&= \left[\sum_s f(s)p(s) \log f(s)p(s)\right] - \left[\sum_s f(s)p(s) \log q_\theta(s)\right]
\end{aligned}
$$

The first term does not depend on on θ, so we can just focus on maximizing the second term:

$$
\begin{aligned}
&\sum_s p(s) f(s) \log q_\theta(s) \\
&= \sum_s q_\theta(s) \frac{p(s)}{q_\theta(s)} f(s) \log q_\theta(s) \\
&= \mathbb{E}_{s \sim q} \left\{ \rho(s) f(s) \log q_\theta(s) \right\} \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \rho_i f(s_i) \log q_\theta(s_i)
\end{aligned}
$$

Mixture Importance Sampling

We can try to avoid *light tails* by ensuring a proper coverage through a mixture distribution $q(s) = \sum_{i=1}^{k} \alpha_i q_i(s)$

$$J \approx \frac{1}{N} \sum_{i=1}^{N} f(s_i) \frac{p(s_i)}{\sum_{j=1}^{k} \alpha_j q_j(s_i)}$$

When we mix the nominal distribution, we get a **defensive importance sampling** scheme. The variance in a two-components defensive mixture them becomes:

$$\mathrm{Var}(J) \leq \frac{1}{N\alpha_1} \left(\sigma^2 + \alpha_2 \mu^2\right)$$

IS for policy gradient methods

Off-policy actor-critic (Degris 2012) approximates the gradient as:

$$
\begin{aligned}
\nabla_\theta J(\pi) &= \sum_s d^b(s) \sum_a \nabla_\theta \left[\pi(s,a) Q^\pi(s,a)\right] \\
&= \sum_s d^b(s) \sum_a \nabla_\theta \pi(s,a) Q^\pi(s,a) + \pi(s,a) Q^\pi(s,a) \\
&\approx \sum_s d^b(s) \sum_a \nabla_\theta \pi(s,a) Q^\pi(s,a) \\
&= \sum_s d^b(s) \sum_a \pi(s,a) \frac{1}{\pi(s,a)} \nabla_\theta \pi(s,a) Q^\pi(s,a) \\
&= \sum_s d^b(s) \sum_a \pi(s,a) \nabla_\theta \log \pi(s,a) Q^\pi(s,a) \\
&= \mathbb{E}_{s\sim d^b, a\sim\pi} \left\{ \nabla_\theta \log \pi(s,a) Q^\pi(s,a) \right\}
\end{aligned}
$$

*Note that in the continuous world, interchange of differentiation and integration must generally be justified by the bounded convergence theorem.

Off-policy actor critic

Now the case where samples all come from a different behavior policy:

$$
\begin{aligned}
\nabla_\theta J(\pi) &\approx \mathbb{E}_{s\sim\pi}\left\{\nabla_\theta \log \pi(s,a) Q^\pi(s,a)\right\} \\
&= \sum_s d^b(s) \sum_a b(s,a) \frac{\pi(s,a)}{b(s,a)} \frac{1}{\pi(s,a)} \nabla_\theta \pi(s,a) Q^\pi(s,a) \\
&= \sum_s d^b(s) \sum_a b(s,a) \frac{\pi(s,a)}{b(s,a)} \frac{1}{\pi(s,a)} \nabla_\theta \pi(s,a) Q^\pi(s,a) \\
&= \mathbb{E}_{s\sim d^b, a\sim b}\left\{\rho(s,a) \nabla_\theta \log \pi_\theta(s,a) Q^\pi(s,a)\right\}
\end{aligned}
$$

A glimpse at off-policy evaluation

The $Q^\pi(s, a)$ term above will have to be evaluated seperately in an actor-critic manner. Q-learning might be suitable for this (although it is known to diverge with function approximation). Note that we can write:

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \{R(\tau)\}$$

where τ is a r.v. denoting length k trajectories, and $R(\cdot)$ is the cumulative return for a given trajectory. By the Markov property, the importance ratios will be of the form:

$$\rho(\tau) = \frac{\prod_{i=1}^{k} \pi(s, a)}{\prod_{i=1}^{k} b(s, a)}$$

and ratios over trajectories will somehow be more difficult to work with...

References

- Thomas Degris, Martha White, Richard S. Sutton: **Linear Off-Policy Actor-Critic**. ICML 2012

Available online, chapter 9 of:

- Art B. Owen: **Monte Carlo theory, methods and examples**

About the CE method:

- Reuven Y. Rubinstein. **Simulation and the Monte Carlo Method**, New York, Wiley

Available in our lab:

- Christian P. Robert, George Casella. **Monte Carlo Statistical Methods**. Springer