Double Q-Learning

Yaroslav Ganin

February 10, 2017

Reinforcement Learning class (COMP 767)

Optimal action-value function and Q-Learning

We are interested in finding the solution to the **Bellman equation**:

$$Q^*(s,a) = \sum_{s'} P_{sa}^{s'} \left(R_{sa}^{s'} + \gamma \max_b Q^*(s',b) \right)$$

One possible option – **Q-learning**:

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha \left(r + \gamma \max_b Q_t(s',b) - Q_t(s,a) \right)$$

Major problem:

- Q-Learning **overestimates** Q^*
- Bias is accumulated at every update

Optimal action-value function and Q-Learning

We are interested in finding the solution to the **Bellman equation**:

$$Q^*(s,a) = \sum_{s'} P_{sa}^{s'} \left(R_{sa}^{s'} + \gamma \max_b Q^*(s',b) \right)$$

One possible option – **Q-learning**:

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha \left(r + \gamma \max_b Q_t(s',b) - Q_t(s,a) \right)$$

Major problem:

- Q-Learning **overestimates** Q^*
- Bias is accumulated at every update

Optimal action-value function and Q-Learning

We are interested in finding the solution to the **Bellman equation**:

$$Q^*(s,a) = \sum_{s'} P_{sa}^{s'} \left(R_{sa}^{s'} + \gamma \max_b Q^*(s',b) \right)$$

One possible option – **Q-learning**:

$$Q_{t+1}(s,a) \leftarrow Q_t(s,a) + \alpha \left(r + \gamma \max_b Q_t(s',b) - Q_t(s,a) \right)$$

Major problem:

- Q-Learning **overestimates** Q^*
- Bias is accumulated at every update

Why is that?

Q_t is a **noisy approximation** of Q^*.

In presence of noise we get:

$$\mathbb{E}\left[\max_b Q_t(s', b)\right] > \max_b \mathbb{E}\left[Q_t(s', b)\right]$$

That's because Q-Learning is based on a *so-called* **single estimator** for each variable (as opposed to the **double estimator** proposed in the paper).

The Single Estimator

Quite often (also in Q-Learning) we have a set of RVs $Z = \{Z_1, \dots, Z_M\}$ and we want to estimate:

$$\max_i \mathbb{E}[Z_i]$$

Say, we have a set of **unbiased** estimators $\{\mu_1, \dots, \mu_M\}$, s.t. $\mathbb{E}[\mu_i] = \mathbb{E}[Z_i]$. Then an obvious estimator is

$$\max_i \mu_i \approx \max_i \mathbb{E}[Z_i]$$

Turns out to be **positively biased**! From **Jensen's inequality**:

$$\mathbb{E}\left[\max_i \mu_i\right] \geq \max_i \mathbb{E}[\mu_i] = \max_i \mathbb{E}[Z_i]$$

So, often $\max_i \mu_i > \max_i \mathbb{E}[Z_i]$.

The Single Estimator

Quite often (also in Q-Learning) we have a set of RVs $Z = \{Z_1, \dots, Z_M\}$ and we want to estimate:

$$\max_i \mathbb{E}[Z_i]$$

Say, we have a set of **unbiased** estimators $\{\mu_1, \dots, \mu_M\}$, s.t. $\mathbb{E}[\mu_i] = \mathbb{E}[Z_i]$. Then an obvious estimator is

$$\max_i \mu_i \approx \max_i \mathbb{E}[Z_i]$$

Turns out to be **positively biased**! From **Jensen's inequality**:

$$\mathbb{E}\left[\max_i \mu_i\right] \geq \max_i \mathbb{E}[\mu_i] = \max_i \mathbb{E}[Z_i]$$

So, often $\max_i \mu_i > \max_i \mathbb{E}[Z_i]$.

The Single Estimator

Quite often (also in Q-Learning) we have a set of RVs $Z = \{Z_1, \dots, Z_M\}$ and we want to estimate:

$$\max_i \mathbb{E}[Z_i]$$

Say, we have a set of **unbiased** estimators $\{\mu_1, \dots, \mu_M\}$, s.t. $\mathbb{E}[\mu_i] = \mathbb{E}[Z_i]$. Then an obvious estimator is

$$\max_i \mu_i \approx \max_i \mathbb{E}[Z_i]$$

Turns out to be **positively biased**! From **Jensen's inequality**:

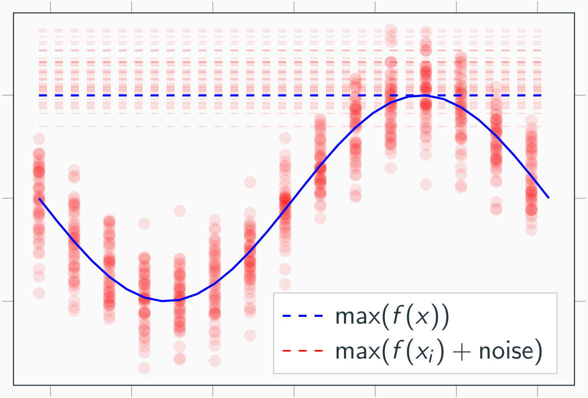
$$\mathbb{E}\left[\max_i \mu_i\right] \geq \max_i \mathbb{E}[\mu_i] = \max_i \mathbb{E}[Z_i]$$

So, often $\max_i \mu_i > \max_i \mathbb{E}[Z_i]$.

The Single Estimator (cont.)

Illustration:

- Let Z_i be deterministic $f(x_i)$ (blue curve)
- Let $\mu_i = f(x_i) + \text{noise}$ (an unbiased estimate of Z_i)



The Double Estimator

We tackle the overestimation by introducing **two sets of estimates**: $\mu^A = \{\mu_1^A, \ldots, \mu_M^A\}$ and $\mu^B = \{\mu_1^B, \ldots, \mu_M^B\}$:

- Obtained on two non-overlapping subsets of the samples $\Rightarrow$ independent
- Both are unbiased: $\mathbb{E}\left[\mu_i^A\right] = \mathbb{E}\left[\mu_i^B\right] = \mathbb{E}[Z_i]$

The **proposed estimator**:

- Select maximizing argument from one set: $a^* = \arg\max_i \mu_i^A$
- Plug it into the other set: $\mu_{a^*}^B \approx \max_i \mathbb{E}[Z_i]$

The Double Estimator underestimates

Lemma

Let $\mathcal{M} = \{ j \,|\, \mathbb{E}[Z_j] = \max_i \mathbb{E}[Z_i] \}$ *be the set of elements that maximize the expected values. Let* a^* *be an element that maximizes* μ^A*. Then*

$$\mathbb{E}\left[\mu^B_{a^*}\right] = \mathbb{E}\left[Z_{a^*}\right] \leq \max_i \mathbb{E}[Z_i] \,.$$

Furthermore, the inequality is strict **iff** $\mathbb{P}[a^* \notin \mathcal{M}] > 0$*.*

This lemma tells us that the Double Estimator has a **negative bias** (underestimates).

The Double Estimator underestimates (cont.)

Proof:

The maximizer a^* can be either in $\mathcal{M}$ or not in $\mathcal{M}$ (obviously):

Consider $a^* \in \mathcal{M}$, then

$$
\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*} \,|\, a^* \in \mathcal{M}\right] &= \mathbb{E}\left[\sum_i \mu^B_i \mathbb{1}\{a^* = i\} \,|\, a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i \,|\, a^* \in \mathcal{M}\right] \cdot \mathbb{P}\left[a^* = i \,|\, a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i\right] \cdot \mathbb{P}\left[a^* = i \,|\, a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[Z_i\right] \cdot \mathbb{P}\left[a^* = i \,|\, a^* \in \mathcal{M}\right] \\
&= \mathbb{E}\left[Z_{a^*} \,|\, a^* \in \mathcal{M}\right] = \max_i \mathbb{E}[Z_i]
\end{aligned}
$$

The Double Estimator underestimates (cont.)

Proof:

The maximizer a^* can be either in $\mathcal{M}$ or not in $\mathcal{M}$ (obviously):

Consider $a^* \in \mathcal{M}$, then

$$
\begin{aligned}
\mathbb{E}\left[\mu_{a^*}^B \mid a^* \in \mathcal{M}\right] &= \mathbb{E}\left[\sum_i \mu_i^B \mathbb{1}\{a^* = i\} \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu_i^B \mid a^* \in \mathcal{M}\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu_i^B\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[Z_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \mathbb{E}\left[Z_{a^*} \mid a^* \in \mathcal{M}\right] = \max_i \mathbb{E}[Z_i]
\end{aligned}
$$

The Double Estimator underestimates (cont.)

Proof:

The maximizer a^* can be either in $\mathcal{M}$ or not in $\mathcal{M}$ (obviously):

Consider $a^* \in \mathcal{M}$, then

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] &= \mathbb{E}\left[\sum_i \mu^B_i \mathbb{1}\{a^* = i\} \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i \mid a^* \in \mathcal{M}\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[Z_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \mathbb{E}\left[Z_{a^*} \mid a^* \in \mathcal{M}\right] = \max_i \mathbb{E}[Z_i]
\end{aligned}$$

The Double Estimator underestimates (cont.)

Proof:

The maximizer a^* can be either in $\mathcal{M}$ or not in $\mathcal{M}$ (obviously):

Consider $a^* \in \mathcal{M}$, then

$$
\begin{aligned}
\mathbb{E}\left[\mu_{a^*}^B \mid a^* \in \mathcal{M}\right] &= \mathbb{E}\left[\sum_i \mu_i^B \, \mathbb{1}\{a^* = i\} \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu_i^B \mid a^* \in \mathcal{M}\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu_i^B\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[Z_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \mathbb{E}\left[Z_{a^*} \mid a^* \in \mathcal{M}\right] = \max_i \mathbb{E}[Z_i]
\end{aligned}
$$

The Double Estimator underestimates (cont.)

Proof:

The maximizer a^* can be either in $\mathcal{M}$ or not in $\mathcal{M}$ (obviously):

Consider $a^* \in \mathcal{M}$, then

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] &= \mathbb{E}\left[\sum_i \mu^B_i \mathbb{1}\{a^* = i\} \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i \mid a^* \in \mathcal{M}\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[Z_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \mathbb{E}\left[Z_{a^*} \mid a^* \in \mathcal{M}\right] = \max_i \mathbb{E}[Z_i]
\end{aligned}$$

The Double Estimator underestimates (cont.)

Proof:

The maximizer a^* can be either in $\mathcal{M}$ or not in $\mathcal{M}$ (obviously):

Consider $a^* \in \mathcal{M}$, then

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] &= \mathbb{E}\left[\sum_i \mu^B_i \mathbb{1}\{a^* = i\} \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i \mid a^* \in \mathcal{M}\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[Z_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \mathbb{E}\left[Z_{a^*} \mid a^* \in \mathcal{M}\right] = \max_i \mathbb{E}[Z_i]
\end{aligned}$$

The Double Estimator underestimates (cont.)

Proof:

The maximizer a^* can be either in $\mathcal{M}$ or not in $\mathcal{M}$ (obviously):

Consider $a^* \in \mathcal{M}$, then

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] &= \mathbb{E}\left[\sum_i \mu^B_i \, \mathbb{1}\{a^* = i\} \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i \mid a^* \in \mathcal{M}\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[\mu^B_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \sum_i \mathbb{E}\left[Z_i\right] \cdot \mathbb{P}\left[a^* = i \mid a^* \in \mathcal{M}\right] \\
&= \mathbb{E}\left[Z_{a^*} \mid a^* \in \mathcal{M}\right] = \max_i \mathbb{E}[Z_i]
\end{aligned}$$

The Double Estimator underestimates (cont.)

Now consider $a^* \notin \mathcal{M}$ and choose $j \in \mathcal{M}$, then

$$\mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] = \mathbb{E}\left[Z_{a^*} \mid a^* \notin \mathcal{M}\right] < \mathbb{E}\left[Z_j\right] = \max_i \mathbb{E}[Z_i]$$

Let's combine the expectations above:

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*}\right] &= \mathbb{P}\left[a^* \in \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] + \\
&\quad \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&= \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&\leq \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \max_i \mathbb{E}[Z_i] \\
&= \max_i \mathbb{E}[Z_i]
\end{aligned}$$

This inequality is strict **iff** $\mathbb{P}[a^* \notin \mathcal{M}] > 0$.

The Double Estimator underestimates (cont.)

Now consider $a^* \notin \mathcal{M}$ and choose $j \in \mathcal{M}$, then

$$\mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] = \mathbb{E}\left[Z_{a^*} \mid a^* \notin \mathcal{M}\right] < \mathbb{E}\left[Z_j\right] = \max_i \mathbb{E}[Z_i]$$

Let's combine the expectations above:

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*}\right] = {} & \mathbb{P}\left[a^* \in \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] + \\
& \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
= {} & \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
\leq {} & \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \max_i \mathbb{E}[Z_i] \\
= {} & \max_i \mathbb{E}[Z_i]
\end{aligned}$$

This inequality is strict **iff** $\mathbb{P}[a^* \notin \mathcal{M}] > 0$.

The Double Estimator underestimates (cont.)

Now consider $a^* \notin \mathcal{M}$ and choose $j \in \mathcal{M}$, then

$$\mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] = \mathbb{E}\left[Z_{a^*} \mid a^* \notin \mathcal{M}\right] < \mathbb{E}\left[Z_j\right] = \max_i \mathbb{E}[Z_i]$$

Let's combine the expectations above:

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*}\right] &= \mathbb{P}\left[a^* \in \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] + \\
&\qquad \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&= \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&\leq \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \max_i \mathbb{E}[Z_i] \\
&= \max_i \mathbb{E}[Z_i]
\end{aligned}$$

This inequality is strict **iff** $\mathbb{P}[a^* \notin \mathcal{M}] > 0$.

The Double Estimator underestimates (cont.)

Now consider $a^* \notin \mathcal{M}$ and choose $j \in \mathcal{M}$, then

$$\mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] = \mathbb{E}\left[Z_{a^*} \mid a^* \notin \mathcal{M}\right] < \mathbb{E}\left[Z_j\right] = \max_i \mathbb{E}[Z_i]$$

Let's combine the expectations above:

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*}\right] &= \mathbb{P}\left[a^* \in \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] + \\
&\quad \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&= \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&\le \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \max_i \mathbb{E}[Z_i] \\
&= \max_i \mathbb{E}[Z_i]
\end{aligned}$$

This inequality is strict **iff** $\mathbb{P}[a^* \notin \mathcal{M}] > 0$.

The Double Estimator underestimates (cont.)

Now consider $a^* \notin \mathcal{M}$ and choose $j \in \mathcal{M}$, then

$$\mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] = \mathbb{E}\left[Z_{a^*} \mid a^* \notin \mathcal{M}\right] < \mathbb{E}\left[Z_j\right] = \max_i \mathbb{E}[Z_i]$$

Let's combine the expectations above:

$$\begin{aligned}
\mathbb{E}\left[\mu^B_{a^*}\right] &= \mathbb{P}\left[a^* \in \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \in \mathcal{M}\right] + \\
&\quad \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&= \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \mathbb{E}\left[\mu^B_{a^*} \mid a^* \notin \mathcal{M}\right] \\
&\leq \mathbb{P}\left[a^* \in \mathcal{M}\right] \max_i \mathbb{E}[Z_i] + \mathbb{P}\left[a^* \notin \mathcal{M}\right] \max_i \mathbb{E}[Z_i] \\
&= \max_i \mathbb{E}[Z_i]
\end{aligned}$$

This inequality is strict **iff** $\mathbb{P}[a^* \notin \mathcal{M}] > 0$.

Double Q-Learning

Algorithm Double Q-Learning

Initialize $Q^A(s,a)$ and $Q^B(s,a)$, $\forall s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
for each episode **do**
  Initialize s
  for each step of the episode **do**
    Choose a from s using policy derived from Q_A and Q_B
    Take action a, observe r, s'
    Toss a fair coin
    if heads **then**
      $Q^A(s,a) \leftarrow Q^A(s,a) + \alpha \left(r + \gamma Q^B \left(s', \arg\max_a Q^A(S', a) \right) - Q^A(s,a) \right)$
    else
      $Q^B(s,a) \leftarrow Q^A(s,a) + \alpha \left(r + \gamma Q^A \left(s', \arg\max_a Q^B(S', a) \right) - Q^B(s,a) \right)$
    end if
    $s \leftarrow s'$
  end for
end for