Off-Policy Optimization and Tabular Q-Learning

Mohammad Sadegh Talebi
m.shahi@di.ku.dk
Department of Computer Science

Recap

Recap

Policy Evaluation (PE):

- Estimating V^π, in an unknown discounted MDP, using data collected according to a fixed π
- Data could be from dataset (offline) or via interaction (online)
- TD update:

$$V(s) \leftarrow \begin{cases} V(s) + \alpha_t \Big(r_t + \gamma V(s_{t+1}) - V(s) \Big) & s = s_t \\ V(s) & \text{else.} \end{cases}$$

- If (i) π is exploratory enough, and (ii) $(\alpha_t)_t$ satisfies Robbins-Monro conditions:

$$V \rightarrow_{t\to\infty} V^\pi \quad \text{almost surely}$$

OPE/OPO

Two related problems:

- **Off-Policy Evaluation (OPE):** Estimate V^π of a target policy π using data collected according to some behavior/logging policy π_b
- **Off-Policy Optimization (OPO):** Find an optimal policy $\pi^\star$ using data collected according to some behavior policy π_b

This lecture: Two algorithms for OPO

Off-Policy Optimization

Off-Policy Optimization

Given: Data $\mathcal{D}$ collected under some policy $\pi_{\rm b}$ (not necessarily fixed).

Mathematically, $\mathcal{D} = \Big\{(s_t, a_t, r_t), 1 \le t \le n\Big\}$ where

$$a_t \sim \pi_{\mathsf{b}}(\cdot|s_t), \quad r_t \sim R(s_t, a_t), \quad s_{t+1} \sim P(\cdot|s_t, a_t)$$

Goal: Find an optimal policy $\pi^\star$, or a near-optimal one.

Action-Value Function (Q-Function)

Action-Value Function

The action-value function of policy π (or simply, Q-value of π) is a mapping $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defined as (Under the bounded reward assumption)

$$Q^\pi(s,a) := \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \Big| s_0 = s, a_0 = a\right].$$

- Intuitively, $Q^\pi(s,a)$ measures the sum of future discounted rewards (in expectation) when the agent starts in s and takes action a in the first step (possibly $a \neq \pi(s)$), and then follows π afterwards.
- Again, recall that we assumed bounded rewards.
- We have

$$|Q^\pi(s,a)| \le \frac{R_{\max}}{1-\gamma}, \quad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$$

- For all $s \in \mathcal{S}$, $Q^\pi(s, \pi(s)) = V^\pi(s)$.

Bellman Optimality Equation

Recall

$$V^\star = \sup_{\pi \in \Pi^{\mathsf{HR}}} V^\pi = \max_{\pi \in \Pi^{\mathsf{SD}}} V^\pi$$

$Q^\star$ and $V^\star$ are related as

$$V^\star(s) = \max_{a \in \mathcal{A}} Q^\star(s,a)$$

Theorem

$V^\star$ and $Q^\star$ satisfy the optimal Bellman equation:

$$V^\star(s) = \max_{a \in \mathcal{A}} \Big(R(s,a) + \gamma \sum_{x \in \mathcal{S}} P(x|s,a) V^\star(x) \Big), \quad s \in \mathcal{S}$$

$$Q^\star(s,a) = R(s,a) + \gamma \sum_{x \in \mathcal{S}} P(x|s,a) \max_{b \in \mathcal{A}} Q^\star(x,b), \quad s \in \mathcal{S}, a \in \mathcal{A}$$

Optimality Theorems

A fundamental result in the theory of discounted MDPs:

Theorem

A stationary deterministic policy π is optimal if and only if it attains the maximum in the Bellman optimality equations: For all $s \in \mathcal{S}$,

$$\pi(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \Big(R(s,a) + \gamma \sum_{x \in \mathcal{S}} P(x|s,a) V^\star(x) \Big)$$

or equivalently,

$$\pi(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \Big(\underbrace{R(s,a) + \gamma \sum_{x \in \mathcal{S}} P(x|s,a) \max_{b \in \mathcal{A}} Q^\star(x,b)}_{=Q^\star(s,a)} \Big) .$$

In short, the optimal policy is the greedy policy w.r.t. $Q^\star$. Hence, enough to compute/learn $Q^\star$.

Optimal Bellman Operator

The optimal Bellman operator is a mapping $\mathcal{T} : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$, such that for any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\mathcal{T} f(s,a) := R(s,a) + \gamma \sum_{x\in\mathcal{S}} P(x|s,a) \max_{b\in\mathcal{A}} f(x,b), \quad s \in \mathcal{S}, a \in \mathcal{A}$$

$\mathcal{T}$ applies to (or *operates on*) a function defined on $\mathcal{S}$ and returns another function defined on $\mathcal{S}$.

- $Q^\star$ satisfies $\mathcal{T}Q^\star = Q^\star$.
- In words, $Q^\star$ is the *unique* fixed-point of the operator $\mathcal{T}^\star$.

`CE-OPO`: A Model-Based Method based on Certainty Equivalence

Known Model

When the model (MDP) is known, we simply solve the Bellman optimality equations using, e.g., `VI` or `QVI`:

`QVI` is quite similar to `VI`. It starts from Q_0 and iterates for $n \geq 1$:

$$Q_{n+1} = \mathcal{T} Q_n$$

I.e.,

$$Q_{n+1}(s,a) = R(s,a) + \gamma \sum_{x \in \mathcal{S}} P(x|s,a) \max_{b \in \mathcal{A}} Q_n(x,b)$$

$\Longrightarrow$ Q_n converges to $Q^\star$ since $\mathcal{T}$ is contractive.

CE-OPO: Certainty Equivalence OPO

Model (MDP) is unknown, so one cannot solve the Bellman optimality equations.

Idea: Estimate the MDP using data and apply the certainty equivalence principle.

- **Step 1:** Compute estimate $\widehat{P}$ (of P) and $\widehat{R}$ (of R)
- **Step 2:** Solve the Bellman optimality equations using $\widehat{P}$ and $\widehat{R}$

CE-OPO

Idea: Estimate the MDP using data and apply the certainty equivalence principle.

- **Step 1:** Compute estimate $\widehat{P}$ (of P) and $\widehat{R}$ (of R)
- **Step 2:** Solve the Bellman optimality equations using $\widehat{P}$ and $\widehat{R}$

We introduce visit counts for various triplets (s, a, s').

Given a dataset $\mathcal{D} = \{(s_t, a_t, r_t), 1 \le t \le n\}$, define for any (s, a, s'),

$$N(s,a,s') = \sum_{t=1}^{n-1} \mathbb{I}\{s_t = s, a_t = a, s_{t+1} = s'\}$$

$$N(s,a) = \sum_{s' \in \mathcal{S}} N(s,a,s')$$

A better choice in practice is

$$N(s,a) = \max\left\{1, \sum_{s' \in \mathcal{S}} N(s,a,s')\right\}$$

CE-OPO

Idea: Estimate the MDP using data and apply the certainty equivalence principle.

- **Step 1:** Compute estimate $\widehat{P}$ (of P) and $\widehat{R}$ (of R)
- **Step 2:** Solve the Bellman optimality equations using $\widehat{P}$ and $\widehat{R}$

Smoothed Estimator for P:

$$\widehat{P}(s'|s,a) = \frac{N(s,a,s') + \alpha}{N(s,a) + \alpha S}$$

- S denotes the number of states.
- $\alpha \geq 0$ is an arbitrary choice controlling the level of smoothing.
- $\alpha = 0$ corresponds to Maximum Likelihood Estimator (unbiased).
- $\alpha = 1/S$ corresponds to Laplace Smoothed Estimator (biased, but the bias vanishes as $N(s,a)$ increases).
- If $\alpha = 0$, for $N(s,a) = 0$, define $\widehat{P}(s'|s,a) = 1/S$.

CE-OPO

Idea: Estimate the MDP using data and apply the certainty equivalence principle.

- **Step 1:** Compute estimate $\widehat{P}$ (of P) and $\widehat{R}$ (of R)
- **Step 2:** Solve the Bellman optimality equations using $\widehat{P}$ and $\widehat{R}$

Smoothed Estimator for R:

$$\widehat{R}(s,a) = \frac{\alpha + \sum_{t=1}^{n-1} r_t \mathbb{I}\{s_t = s, a_t = a\}}{\alpha + N(s,a)}$$

- $\alpha \geq 0$ is an arbitrary choice controlling the level of smoothing.
- $\alpha = 0$ corresponds to Maximum Likelihood Estimator (unbiased).

CE-OPO

Idea: Estimate the MDP using data and apply the certainty equivalence principle.

- **Step 1:** Compute estimate $\widehat{P}$ (of P) and $\widehat{R}$ (of R)
- **Step 2:** Solve the Bellman optimality equations using $\widehat{P}$ and $\widehat{R}$

Using $\widehat{P}$ and $\widehat{R}$, we can solve empirical Bellman optimality equations:

$$\widehat{Q}^\star(s,a) = \widehat{R}(s,a) + \gamma \sum_{x\in\mathcal{S}} \widehat{P}(x|s,a) \max_{b\in\mathcal{A}} \widehat{Q}^\star(x,b)$$

or

$$\widehat{Q}^\star = \widehat{\mathcal{T}}\widehat{Q}^\star$$

$\widehat{\mathcal{T}}$ is the empirical Bellman operator.

$\widehat{Q}^\star = \widehat{\mathcal{T}}\widehat{Q}^\star$ can be solved using `QVI`.

CE-OPO: Certainty Equivalence OPO

CE-OPO: Certainty Equivalence OPO

- **input:** $\mathcal{D} = \{(s_t, a_t, r_t)\}_{1\le t\le n}$, α (optional)
- Compute estimates $\widehat{P}(s'|s,a)$ and $\widehat{R}(s,a)$ for all (s,a,s')
- Find $\widehat{\pi}^\star$, the optimal policy in the empirical MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{R})$.
- **output:** $\widehat{\pi}^\star$

$\widehat{M}$ could be solved using `VI`, `PI`, or `QVI`.

CE-OPO: Asymptotic Convergence

$$\widehat{P}(s'|s,a) \longrightarrow_{N(s,a)\to\infty} P(s'|s,a) \quad \text{almost surely}$$
$$\widehat{R}(s,a) \longrightarrow_{N(s,a)\to\infty} R(s,a) \quad \text{almost surely}$$

If π_{b} is exploratory enough in the sense that $N(s,a) \to_{n\to\infty} \infty$ for all (s,a), then $\widehat{P}$ and $\widehat{R}$ converge to P and R as $n \to \infty$. Thus, we can show

$$\widehat{\mathcal{T}} \longrightarrow_{n\to\infty} \mathcal{T} \qquad \widehat{Q}^\star \longrightarrow_{n\to\infty} Q^\star \quad \text{almost surely}$$

which guarantees

$$\widehat{\pi}^\star \longrightarrow_{n\to\infty} \pi^\star \quad \text{almost surely}$$

Theorem

If all state-action pairs are visited infinitely often under π_{b}, then $\widehat{Q}^\star$ converges to $Q^\star$ almost surely:

$$\mathbb{P}\Big(\lim_{n\to\infty} \widehat{Q}^\star = Q^\star\Big) = 1$$

- Strong guarantee, but only asymptotically (unfortunately).

CE-OPO: Pros and Cons

Disadvantages of the model-based solution:

- Often leads to large variance in the estimation of $Q^\star$
- Computational complexity is $O(S^3)$, and space complexity is $O(S^2)$.
- May not be easily converted to an incremental procedure.

Model-Free Method for OPO: Tabular Q-Learning (and Friends)

Bellman Optimality Equations

Bellman optimality equations (using Q):

$$Q^\star(s,a) = R(s,a) + \gamma \sum_{x \in \mathcal{S}} P(x|s,a) \max_b Q^\star(x,b)$$
$$\mathcal{T}Q^\star = Q^\star$$

Equivalently, $Q^\star$ is the root of functional $F(Q) = \mathcal{T}Q - Q$, namely the solution to the nonlinear system:

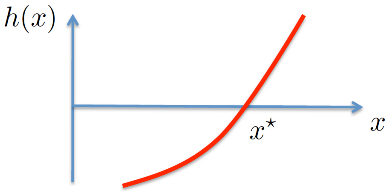
$$F(Q) = \mathcal{T}Q - Q = 0, \qquad \text{where} \;\; Q \in \mathbb{R}^{S \times A}$$

- **Known model**: Find the root of F using VI (or Q-iteration).
- **Unknown model**: We have only samples from R and P (as in `TD`).

We need a root finding method from noisy measurements.

Stochastic Approximation

Stochastic Approximation (SA) is method to find the root of an increasing function from noisy measurements.



The setting:

- At the n-th iteration, you select x_n
- You get a noisy measurement $y_n = h(x_n) + \xi_n$
- ξ_n is a noise with zero-mean but may depend on the selected point x_n
- $\mathbb{E}[\xi_n | \xi_1, \dots, \xi_{n-1}] = 0$

Robbins-Monro Algorithm (1951)

SA proposed by Robbins & Monro in (1951)

$$x_{n+1} = x_n - \alpha_n y_n = x_n - \alpha_n(h(x_n) + \xi_n), \quad n \geq 1$$

with $(\alpha_n)_n$ satisfying the Robbins-Monro conditions:

$$\alpha_n > 0, \quad \sum_{n=1}^{\infty} \alpha_n = \infty, \quad \text{and} \sum_{n=1}^{\infty} \alpha_n^2 < \infty$$

Theorem

Under the following assumptions

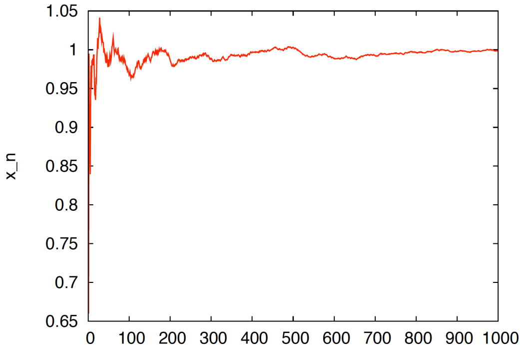
1. $\mathbb{E}[\xi_n | \xi_1, \ldots, \xi_{n-1}] = 0$
2. $\mathbb{E}[\|\xi_n\|^2 | \xi_1, \ldots, \xi_{n-1}] \leq K(1 + \|x_n\|^2)$, *almost surely for some* K
3. h *is Lipschitz*
4. $\sup_n \|x_n\| < \infty$, *almost surely*

$$\lim_{n\to\infty} x_n = x^\star \qquad \textit{almost surely.}$$

Example

Solving $h(x) = x^2 - 1 = 0$ through noisy samples from h using SA

$$h(x) = x^2 - 1, \quad a_n = 1/n, x_0 = 0$$



SA for $F(Q) = \mathcal{T}Q - Q$

We apply SA to $F(Q) = \mathcal{T}Q - Q$.

- Consider a sample (s_t, a_t, r_t, s_{t+1}).
- We show that $Y_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$, conditioned on (s_t, a_t) and Q is an unbiased sample from $F(Q)(s_t, a_t)$.

$$\begin{aligned}
\mathbb{E}[Y_t|Q, s_t, a_t] &= \mathbb{E}\Big[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\Big|Q, s_t, a_t\Big] \\
&= \underbrace{\mathbb{E}\Big[r_t\Big|Q, s_t, a_t\Big]}_{=R(s_t,a_t)} + \gamma\mathbb{E}\Big[\max_{a'} Q(s_{t+1}, a')\Big|Q, s_t, a_t\Big] - Q(s_t, a_t) \\
&= R(s_t, a_t) + \gamma \sum_{x\in\mathcal{S}} P(x|s_t, a_t) \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \\
&= \mathcal{T}Q(s_t, a_t) - Q(s_t, a_t) = F(Q)(s_t, a_t)
\end{aligned}$$

Hence, $\mathbb{E}[Y_t|\mathcal{H}_{t-1}] = F(Q)(s_t, a_t)$.

- The other technical conditions of SA can be verified. (Technical and tedious, so omitted here.)

SA for $F(Q) = \mathcal{T}Q - Q$

Application of SA to $F(Q) = \mathcal{T}Q - Q$:

The Q-Learning (`QL`) update rule:

$$\underbrace{Q(s_t, a_t)}_{\text{new value}} \leftarrow \underbrace{Q(s_t, a_t)}_{\text{new value}} + \underbrace{\alpha_t \Big(r_t + \gamma \max_{b \in \mathcal{A}} Q(s_{t+1}, b) - Q(s_t, a_t) \Big)}_{\text{correction}}$$

And $Q(s, a)$ unchanged if $(s, a) \neq (s_t, a_t)$.

QL: Learning Rate

To guarantee convergence, learning rates $(\alpha_t)_{t\geq 1}$ must satisfy the *Robbins-Monro conditions*:

$$\alpha_t > 0, \qquad \sum_{t=1}^{\infty} \alpha_t = \infty, \qquad \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

(I.e., a positive sequence that is *square-summable-but-not-summable.*)

Examples:

- $\alpha_t = \frac{1}{t+1}$,
- $\alpha_t = \frac{c}{t^a}$ for $a \in (\frac{1}{2}, 1]$
- $\alpha_t = \alpha_t(s,a) = \frac{1}{N_t(s,a)+1}$, where $N_t(s,a)$ is the number of times (s,a) is sampled in the first $t-1$ rounds —i.e., learning rate can be personalized to (s,a), assuming that Robbins-Monro conditions could be met.

QL

input: $\mathcal{D} = \{(s_t, a_t, r_t)\}_{1\le t\le n}$, $(\alpha_t)_{t\ge 1}$

initialization: Select Q_1 arbitrarily

for $t = 1, \dots, n-1$:

- $\delta_t = r_t + \gamma \max_{b\in\mathcal{A}} Q_t(s_{t+1}, b) - Q_t(s_t, a_t)$
- Update:

$$Q_{t+1}(s,a) = \begin{cases} Q_t(s,a) + \alpha_t \delta_t & (s,a) = (s_t, a_t) \\ Q_t(s,a) & \text{else.} \end{cases}$$

output: Greedy policy w.r.t. Q_n

- Q_n is an estimate of $Q^\star$, giving an estimate $\widehat{\pi^\star}$ of the optimal policy:

$$\widehat{\pi^\star}(s) \in \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, Q_n(s,a)$$

QL: Asymptotic Convergence

Theorem

If all state-action pairs are visited infinitely often in $\mathcal{D}$ and $(\alpha_t)_{t\geq 1}$ satisfies the Robbins-Monro conditions, then Q_t converges to the true value function $Q^\star$ almost surely:

$$\mathbb{P}\left(\forall s \in \mathcal{S}, a \in \mathcal{A},\ \lim_{t\to\infty} Q_t(s,a) = Q^\star(s,a)\right) = 1$$

In other words, if π_{b} (used to collect $\mathcal{D}$) is exploratory enough, Q_t converges to $Q^\star$, in the following sense:

$$\mathbb{P}\left(\exists \mathcal{D}, \exists (s,a) : \lim_{t\to\infty} Q_t(s,a;\mathcal{D}) \neq Q^\star(s,a)\right) = 0$$

I.e., datasets for which $Q_\infty \neq Q^\star$ will occur with probability 0.

On Behavior Policy

- (Asymptotic) convergence requires that state-action pairs are visited infinitely often.
- The behavior policy π_{b} could change during the learning, as long as it is kept exploratory enough.
  - E.g., ε-greedy policy (for some $\varepsilon > 0$)

    $$\pi_{\varepsilon\text{-greedy}}(s) = \begin{cases} \mathrm{argmax}_a\, Q_t(s,a) & \text{w.p. } 1-\varepsilon \\ \text{sample uniformly at random from } \mathcal{A} & \text{w.p. } \varepsilon \end{cases}$$

    Note that $\pi_{\varepsilon\text{-greedy}}(s)$ is non-stationary.
  - E.g., Boltzmann's policy (a.k.a. softmax):

    $$\text{at state } s\text{, select action } a \in \mathcal{A} \text{ w.p.} \quad \frac{e^{\eta Q_t(s,a)}}{\sum_{b\in\mathcal{A}} e^{\eta Q_t(s,b)}}$$

    where $\eta > 0$ is a parameter controlling exploration.
  - Incremental `QL` (cf. the very last slides)

`QL`: Advantages

- `QL` is model-free: It does not require to estimate a model of the MDP, and only relies on collected experience.
- `QL` can be incremental (unlike the model-based methods).
- Space complexity is $O(SA)$ and computational complexity, per round, is $O(A)$. Much cheaper than the model-based method.

QL: Non-Asymptotic Convergence

- Asymptotic convergence results often do not tell us much information about the speeds of convergence.
- We are interested in knowing what happens with small datasets. So we study the non-asymptotic convergence.

Sample complexity for OPO

Given $\delta \in (0, 1)$ and $\varepsilon > 0$, define the PAC off-policy sample complexity as the number $\mathsf{SC}(\varepsilon, \delta)$ of samples from the MDP such that for all $n \geq \mathsf{SC}(\varepsilon, \delta)$,

$$\|Q^\star - Q_n\|_\infty \leq \varepsilon, \qquad \text{with probability } \geq 1 - \delta$$

Two Definitions

Two notions arising in sample complexity of OPO:

Cover Time t_{cover}. Given $t_1 > 0$, let $t_2 > t_1$ denote the first time step such that all (s, a) pairs are visited at least once with probability at least $\frac{1}{2}$. Then, $t_{\text{cover}} = t_2 - t_1$ defines the cover time of M.

- $t_{\text{cover}} \geq SA$.
- A quantity related to π_{b}.

Effective Horizon. Given $\varepsilon > 0$, the effective horizon is

$$H_{\text{eff}} := \frac{-1}{1-\gamma} \log(\varepsilon(1-\gamma))$$

- Truncating ∞-horizon to H_{eff} would bring at most ε error to $V^\star$.

QL: Non-Asymptotic Convergence

Theorem (Even-dar & Mansour (2003))

Let $\delta \in (0,1)$ *and* $\varepsilon \in (0, \frac{1}{1-\gamma}]$, *and assume that* n *satisfies:*

$$n \geq c \cdot \frac{\left[t_{\mathsf{cover}}\right]^{H_{\mathsf{eff}}}}{\varepsilon^2(1-\gamma)^4} \log\left(\frac{SAn}{\delta}\right) \log\left(\frac{SA}{\varepsilon\delta(1-\gamma)^2}\right)$$

where c *is a universal constant. Then,* QL *with* $\alpha_t(s,a) = \frac{1}{N_t(s,a)+1}$ *satisfies:*

$$\|Q^\star - Q_n\|_\infty \leq \varepsilon, \qquad \textit{with probability} \geq 1-\delta.$$

Essentially, it establishes a sample complexity for QL proportional to

$$\widetilde{O}\left(\frac{\left[t_{\mathsf{cover}}\right]^{H_{\mathsf{eff}}}}{\varepsilon^2(1-\gamma)^4}\right)$$

where $\widetilde{O}(\cdot)$ hides poly-log terms.

QL: Non-Asymptotic Convergence

Theorem (Li et al. (2020))

Let $\delta \in (0,1)$ and $\varepsilon \in (0, \frac{1}{1-\gamma}]$, under QL one has:

$$\|Q^\star - Q_n\|_\infty \leq \varepsilon, \qquad \textit{with probability} \geq 1-\delta.$$

provided that

$$n \geq c \cdot \frac{t_{\mathsf{cover}}}{\varepsilon^2 (1-\gamma)^5} \log^2\left(\frac{SAn}{\delta}\right) \log\left(\frac{1}{\varepsilon(1-\gamma)^2}\right)$$

$$\alpha_t = \frac{c'}{\log(SAn/\delta)} \min\left(\frac{(1-\gamma)^4 \varepsilon^2}{\gamma^2}, 1\right)$$

where c, c' are universal constants.

Essentially, it establishes a sample complexity for QL proportional to

$$\widetilde{O}\left(\frac{t_{\mathsf{cover}}}{\varepsilon^2 (1-\gamma)^5}\right)$$

QL: Overestimation Bias

QL could exhibit weak empirical performance due overestimation bias.

- Overestimation bias stems from the term

$$\max_{b \in \mathcal{A}} Q_t(s_{t+1}, b)$$

  to approximate $\max_{b \in \mathcal{A}} Q^\star(s_{t+1}, b)$ in the update equation of QL.

- It is one major reason behind slow convergence of QL in practice.

Could we update Q_t in a wiser way?

Idea: $\max_{b \in \mathcal{A}} Q^\star(s_{t+1}, b)$ is related to the classical problem of Estimating the Maximum Expected Value. So let's use a wiser such estimate.

Estimating the Maximum Expected Value

Consider r.v.'s $X_1, \dots, X_m$ with $\mathbb{E}[X_i] = \mu_i$.

- We wish to estimate $\mu_\star = \max_i \mathbb{E}[X_i]$.
- Distributions of $X_i, \dots, X_m$ unknown.
- We have a set S_i of i.i.d. samples from each X_i.

Maximum Estimator (ME): We construct $\widehat{\mu}_i := \widehat{\mu}_i(S_i) = \frac{1}{|S_i|}\sum_{x \in S_i} x$, and set

$$\widehat{\mu}_\star^{\mathsf{ME}} := \max_i \widehat{\mu}_i.$$

$\widehat{\mu}_\star^{\mathsf{ME}}$ is positively biased since: $\mathbb{E}[\widehat{\mu}_\star^{\mathsf{ME}}] = \mathbb{E}[\max_i \widehat{\mu}_i] \geq \max_i \mathbb{E}[\widehat{\mu}_i] = \max_i \mu_i = \mu_\star$

Double Estimator (DE): Randomly partition each sample set as $S_i = S_i^A \cup S_i^B$.

$$\bar{i} \in \operatorname*{argmax}_i \widehat{\mu}_i(S_i^A) \qquad \text{Then} \qquad \widehat{\mu}_\star^{\mathsf{DE}} := \widehat{\mu}_{\bar{i}}(S_{\bar{i}}^B).$$

It can be shown that $\widehat{\mu}_\star^{\mathsf{DE}}$ is negatively biased.

Combining Double Estimator with QL

The Double Estimator could be incorporated into QL:

- Let's maintain two estimates of Q-values Q^A and Q^B, each updates using half of the samples from $\mathcal{D}$:
- Update for Q^A

$$Q^A_{t+1}(s,a) = \begin{cases} Q^A_t(s,a) + \alpha_t\Big(r_t + \gamma Q^B_t(s_{t+1}, \overline{\boldsymbol{a}}) - Q^A_t(s,a)\Big) & (s,a) = (s_t, a_t) \\ Q^A_t(s,a) & \text{else.} \end{cases}$$

  with $\overline{\boldsymbol{a}} = \mathrm{argmax}_b\, Q^A_t(s_{t+1}, b)$.

- A similar update will be made for Q^B

The corresponding algorithm is called `Double QL` (van Hasselt, 2010).

Double QL

input: $\mathcal{D} = \{(s_t, a_t, r_t)\}_{1\le t\le n}$, $(\alpha_t)_{t\ge 1}$

initialization: Select Q_1^A, Q_1^B arbitrarily

for $t = 1, \dots, n-1$:

- Set `update-A` $=$ True w.p. 0.5
- **if** `update-A`:
  - $\overline{a} = \mathrm{argmax}_a\, Q_t^A(s_{t+1}, a)$
  - $\delta_t = r_t + \gamma Q_t^B(s_{t+1}, \overline{a}) - Q_t^A(s_t, a_t)$
  - Update: $Q_{t+1}^A(s,a) = \begin{cases} Q_t^A(s,a) + \alpha_t\delta_t & (s,a) = (s_t, a_t) \\ Q_t^A(s,a) & \text{else.} \end{cases}$
- **else:**
  - $\overline{a} = \mathrm{argmax}_a\, Q_t^B(s_{t+1}, a)$
  - $\delta_t = r_t + \gamma Q_t^A(s_{t+1}, \overline{a}) - Q_t^B(s_t, a_t)$
  - Update: $Q_{t+1}^B(s,a) = \begin{cases} Q_t^B(s,a) + \alpha_t\delta_t & (s,a) = (s_t, a_t) \\ Q_t^B(s,a) & \text{else.} \end{cases}$

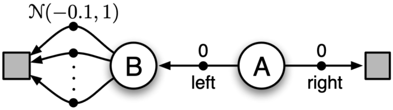
output: Policy greedy w.r.t. $Q_n^A + Q_n^B$

- Q_n is an estimate of $Q^\star$, giving an estimate $\widehat{\pi^\star}$ of the optimal policy:

$$\widehat{\pi^\star}(s) \in \underset{a\in\mathcal{A}}{\mathrm{argmax}}\, Q_n^A(s,a) + Q_n^B(s,a)$$

Double QL vs. QL

Double QL vs. QL in a simple MDP(Source: Sutton & Barto):

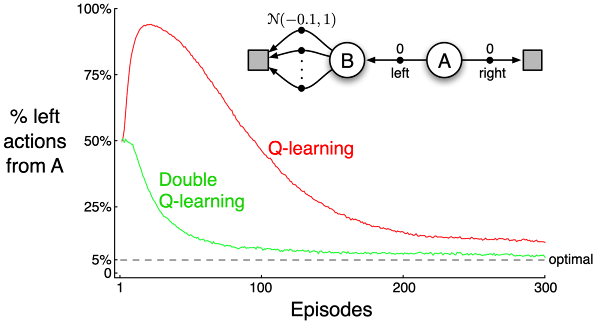


- 4 states: A, B and two terminal states denoted by $\square$
- At A: Two actions ('left' and 'right'), each with $r = 0$
- At B: Multiple actions, each with $r \sim \mathcal{N}(-0.1, 1)$
- $\implies \pi^\star(A) =$ right

However, OPO methods may choose 'left' since maximization bias making B appear to have a positive value.

Double QL vs. QL

Double QL vs. QL in a simple MDP (Source: Sutton & Barto):
Averaged over 10000 runs. π_b, is ε-greedy with $\varepsilon = 0.1$.



- QL initially learns 'left' much more often than 'right'
- In contrast, Double QL is less affected by maximization bias.

Off-Policy vs. Offline

Off-Policy Learning/Optimization $\neq$ Offline RL

- In offline RL, the goal is to learn an optimal policy (or a near-optimal one) from a dataset – *we're offline; no further exploration*.
- Offline RL $\subset$ OPO
- Note that OPO could take place in an online fashion *(but behavior must be generated off-the-target-policy)*

Historical Account

- Christopher Watkins presented QL in 1989 in his PhD thesis.
- In 1994, Tsitsiklis established the almost sure convergence of QL by showing its relation to SA. See the paper for a detailed proof of asymptotic convergence guarantee and verification of SA conditions (Tsitsiklis, 1994).
- Non-asymptotic convergence of QL was done in (Even-dar & Mansour, 2003). State-of-the-art is (Li et al., 2020).
- Double QL is presented in (van Hasselt, 2010).
- Research on improved sample complexity of QL as well as improved variants is ongoing.

References

1. C. Watkins. Learning from delayed rewards. *PhD thesis at King's College*, 1989.
2. H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.
3. J. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 1994.
4. E. Even-Dar and Y. Mansour. Learning rates for Q-learning. *Journal of machine learning Research*, 2003.
5. G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Advances in Neural Information Processing Systems*, 2020.
6. H. van Hasselt. Double Q-learning. *Advances in Neural Information Processing Systems*, 2010.