Reinforcement Learning and Optimal Control

IFT6760C, Fall 2021

Pierre-Luc Bacon

September 28, 2021

Optimality Equations over Q-factors

The version of the optimality equations that we used so far was:

$$v_\gamma^\star(s) = \max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v_\gamma^\star(j) \right\} \quad .$$

Now the same result can also derived over Q-factors/Q-values:

$$Q_\gamma^\star(s,a) = r(s,a) + \gamma \sum_{j \in \mathcal{A}(s)} p(j|s,a) \max_{a' \in \mathcal{A}(j)} Q_\gamma^\star(j,a') \quad .$$

where $v^\star(s)_\gamma = \max_{a \in \mathcal{A}(s)} Q_\gamma^\star(s,a) \quad .$

Advantage

This form will allow us to develop **model-free** algorithms. We saw earlier than an optimal policy $(d^\star)^\infty$ could be derived from a decision rule $d^\star$ such that:

$$d^\star(s) \in \arg\max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \gamma \sum_{j \in \mathcal{S}} p(j|s,a) v^\star_\gamma(j) \right\} \quad .$$

i.e. we need to have direct access to p and r to compute $d^\star$. In contrast, if we are given Q-factors, we can simply compute:

$$d^\star(s) \in \arg\max_{a \in \mathcal{A}(s)} Q^\star_\gamma(s,a) \quad ,$$

where p and r are encapsulated within $Q^\star_\gamma$.

Bellman operator for Q-factors

In the same way that we saw that $v_\gamma^\star$ is a fixed-point of L, $Q_\gamma^\star$ is the unique fixed-point of F defined as:

$$(FQ)(s,a) \triangleq r(s,a) + \gamma \sum_{j \in \mathcal{A}(s)} p(j|s,a) \min_{a' \in \mathcal{A}(j)} Q(j,a') \ ,$$

for $s \in \mathcal{S}, a \in \mathcal{A}(s)$ and $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

- We can also show that F is a γ-contraction under the sup-norm (infinity norm).
- Which also means that we can apply the method of successive approximation to find $Q_\gamma^\star$: aka value iteration.

Stochastic Value Iteration

While computing $d^\star$ can be done model-free given $Q^\star_\gamma$, $Q^\star_\gamma$ itself may not be unless we take a different approach.

Insight: applying F involves evaluating an expectation, which we can estimate by the sample mean:

$$(FQ)(s,a) \triangleq \mathbb{E}\left[r(S_t, A_t) + \gamma \max_{A_{t+1} \in \mathcal{A}(S_{t+1})} Q(S_{t+1}, A_{t+1}) \,\middle|\, S_t = s, A_t = a \right] \quad ,$$

where $S_{t+1} \sim p(\cdot | S_t, A_t)$. We define a random operator $\tilde{F}$ as:

$$(\tilde{F}_N Q)(s,a) = r(s,a) + \frac{\gamma}{N} \sum_{j=1}^{N} \max_{a' \in \mathcal{A}(s_j)} Q(s_j, a') \quad .$$

where $\{s_1, \ldots, s_N\}$ are drawn iid from the conditional $p(\cdot|s,a)$.

Approximate contraction

Note that the above process can be viewed as a noisy version of the deterministic Bellman operator for Q-factors.

Consider the following general iterative procedure:

$$x^{(k+1)} = Tx^{(k)} \quad ,$$

where T is a contractive operator. Now rather than observing the sequence $\{x^{(k)}\}$ directly, we get $\{y^{(k)}\}$. This is due to approximation error: noisy evaluation or discretization for example. We consider:

$$y^{(k+1)} = \tilde{T}_k y^{(k)} \quad ,$$

where $\tilde{T}_k$ is a contraction mapping.

What can we say about how $\{y^{(k)}\}$ is related to $\{x^{(k)}\}$?

Approximate contraction

Theorem (O&R 12.2.1) Let $T : \mathcal{V} \rightarrow \mathcal{V}$ be a γ-contraction and $x^\star$ its unique fixed point. Furthermore, let $\{y_k\} \in \mathcal{V}$ be any sequence, and define the local error for $k = 0, 1, \ldots$ as $\epsilon_k \triangleq \|Ty^{(k)} - y^{(k+1)}\|$. It then holds for $k = 0, 1, \ldots$ that:

$$\|y^{(k+1)} - x^\star\| \leq \frac{1}{1-\gamma}\left(\gamma\|y^{(k+1)} - y^{(k)}\| + \epsilon_k\right)$$

$$\|y^{(k+1)} - x^\star\| \leq \|x^{(k+1)} - x^\star\| + \sum_{j=0}^{k} \gamma^{k-j}\epsilon_j + \gamma^{(k+1)}\|x^{(0)} - y^{(0)}\|,$$

and $\lim_{k\rightarrow\infty} y^{(k)} = x^\star$ if and only if $\lim_{k\rightarrow\infty} \epsilon_k = 0$.

Proof Omitted, see theorem 12.2.1 p. 395 of O&R .

ℹ If $x^{(k)} = y^{(k)}$ then the above result coincides with the error estimate that we have seen in the contraction mapping theorem, namely: $\|x^{(k+1)} - x^\star\| \leq \frac{\gamma}{1-\gamma}\|x^{(k+1)} - x^{(k)}\|$.

Approximate nonstationary contraction

In the above, we assumed that the underlying approximate operator $\tilde{T}_k$ was kept fixed throughout. We can extend the above result to the nonstationary case:

Theorem (12.2.2 in O&R) Let $T : \mathcal{V} \to \mathcal{V}$ be a γ-contraction and $x^\star$ its unique fixed point. Furthermore, let $\{\tilde{T}_k\}, k = 0, 1, \ldots$ be a set of mappings $\tilde{T}_k : \mathcal{V} \to \mathcal{V}$ such that:

$$\lim_{k\to\infty} \|\tilde{T}_k x - Tx\| = 0, \quad \text{uniformly for } x \in \mathcal{V} \quad , .$$

Then the sequence defined by $y^{(k+1)} = \tilde{T}_k y^{(k)}$ converges to the unique fixed point $x^\star$ of T.

Application to stochastic value iteration

12.2.2 provides almost all we need to show convergence of SVI. The argument may look like:

- Given $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$, $F_k Q(s, a)$ converges to $FQ(s, a)$ by the law of large numbers, hence SVI must converge. But…

⚠ 12.2.2 requires $\lim_{k\to\infty} \|\tilde{T}_k x - Tx\| = \sup_{s\in\mathcal{S}} |(\tilde{T}_k x)(s) - (Tx)(s)| = 0$. Therefore, we'd want to not only draw from only **one** conditional $p(\cdot|s, a)$ but for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$ infinitely often.

Online algorithm

Rather than keeping the set of realizations from the conditional $p(\cdot|s,a)$ for each $s \in \mathcal{S}, a \in \mathcal{A}(s)$, we can maintain an online average:

$$Q^{(k+1)}(s,a) = (1-\eta_k)Q^{(k)}(s,a) + \eta_k(F_k Q^{(k)})(s,a), \quad s \in \mathcal{S}, a \in \mathcal{A}(s)$$

$$(F_k Q^{(k)})(s,a) = \begin{cases} r(s_t,a_t) + \gamma \max_{a' \in \mathcal{A}(s_t)} Q^{(k)}(s_t,a') & \text{if } (s,a) = (s_t,a_t) \\ Q^{(k)}(s,a) & \text{otherwise} \end{cases}$$

and $\forall k = 0, 1, \ldots$:

$$\eta_k > 0, \quad \sum_{k=0}^{\infty} \eta_k = \infty, \quad \sum_{k=0}^{\infty} \eta_k^2 < \infty \quad .$$

Function approximation

⚠ What we have done so far is to get rid of the assumption that the transition probability and reward functions are known precisely. All we require is to be able to obtain samples of the next state, infinitely often for every state-action pairs. We haven’t dealt yet with the problem of having **a lot of states**. This is where function approximation comes in. But we’re not there yet…

Stochastic approximation

The online algorithm in the last slide is a stochastic approximation (SA) algorithm.

Remember, there are two things in life:

1. Fixed point problems
2. Root-finding problems

SA is a derivative-free counterpart to Newton's method for finding the zeros of a noisy function. More precisely, given $f : \mathbb{R}^n \to \mathbb{R}^m$ we want to find a $x^\star \in \mathbb{R}^n$ such that:

$$f(x^\star) = 0 \quad .$$

but only based on noisy evaluations of f.

Noisy evaluations

Rather than observing $f(x)$ directly, assume that we have access to noisy measurements. That is:

$$y^{(k)} = f(x) + \epsilon^{(k)}, \quad k = 0, 1, \ldots$$

where ϵ_k is a noise term (unobserved) and $x \in \mathbb{R}^n$. The root-finding SA algorithm (Robbins-Monro, 1951) is of the form:

$$\tilde{x}^{(k+1)} = \tilde{x}^{(k)} - \eta_k y^{(k)} \quad .$$

Example: sample mean

Let $\mathbb{E}[X] = \mu$, we define the same mean estimator as:

$$\bar{X}_{k+1} \triangleq \frac{1}{k+1} \sum_{i=1}^{k+1} X_i \quad ,$$

which can also be written recursively as:

$$= \bar{X}_k - \frac{1}{k+1} \underbrace{\left(\bar{X}_k - X_{k+1}\right)}_{\text{noisy observation}} \quad .$$

with $\bar{X}_0 = 0$.

The root-finding problem is that corresponding to $f(x) \triangleq x - \mu = 0$.