Batch reinforcement learning and fitted Q iteration

Akshay Krishnamurthy
akshaykr@microsoft.com

November 10, 2020

1 Batch reinforcement learning

Batch reinforcement learning broadly refers to reinforcement learning problems in which the learner does not get to interact with the environment. Instead, the learner is simply presented with a batch of experience collected by some decision-making policy, and the goal is to use this data to learn a near-optimal (or at the very least) better policy. This setting is quite important for high-stakes decision-making scenarios, like those that might arise in precision medicine or where safety is a serious concern. On the other hand, one significant challenge is that exploration is not controlled by the learner. Thus we will either (a) require some assumptions that ensure that the data-collection policy effectively covers the state-action space, or (b) not be able to find a global near-optimal policy.

In addition to this "coverage" issue, we will also face some representational issues when we turn to the function approximation setting. Due to our inability to collect data, the space of algorithms we may use is quite limited.

2 The formal setting

We consider the discounted MDP setting, where there is a state space $\mathcal{S}$, action space $\mathcal{A}$, transition operator P, reward function R, discount factor $\gamma \in [0, 1)$ and a starting state distribution d_0. We adopt the standard normalizations that $R(s, a) \in \Delta([0, 1])$ so that the value functions V^π and Q^π functions take values in $[0, V_{\max}]$ with $V_{\max} := 1/(1 - \gamma)$.

In batch RL, rather than interact with the environment to collect data, we will be presented with n tuples (x, a, r, x') where $(x, a) \sim \mu$, $r \sim R(x, a)$ and $x' \sim P(\cdot \mid x, a)$. Here μ is an approximation of the data collection policy, and it is only an approximation because we think of the tuples as iid. This is primarily to simplify the analysis, and it is possible to obtain results when we replace the iid dataset with one actually collected by a policy, which involves dealing with temporal correlations. See for example a paper by Antos, Munos, and Szepesvari for these kinds of arguments.

Given the dataset $\{(x_i, a_i, r_i, x'_i)\}_{i=1}^n \overset{iid}{\sim} \mu$ our goal is to output a near optimal policy for the MDP, that is we would like our algorithm to produce a policy $\hat{\pi}$ such that, with probability at least $1 - \delta$, $V(\hat{\pi}) \geq V^\star - \epsilon$, for some (ϵ, δ) pair. As usual, the number of samples n will depend on the accuracy parameters (ϵ, δ) and we would like n to scale favorably with these.

Example 1. *Suppose that $\gamma = 0$ and d_0 is supported on a single state x_0, so that we are only interested in optimizing the immediate reward from this state x_0. This is just a multi-armed bandit problem, but we do not get to control the data collection process. If $\mu = d_0 \circ \delta_a$ for some fixed action a, then while we can effectively learn the reward for this action, we have no hope of learning the reward for any other action. Therefore, if μ is not already the optimal policy, we have no hope of competing with the optimal policy.*

The issue in the above example is that μ is not providing sufficient coverage over the state-action space. And since we are not able to explore, this can completely prevent us from learning a near-optimal policy. To avoid this, it is common to impose some coverage assumptions.

2.1 Tabular setting, uniform approximation

Let's start out with perhaps the strongest one:

{assum:uni

Assumption 1 (Uniform concentrability)**.** *We assume that there exists some constant* $C_{\text{unif}} < \infty$ *such that*

$$(x,a) \in \mathcal{S} \times \mathcal{A} : \frac{1}{\mu(x,a)} \leq C_{\text{unif}}$$

Clearly this imposes a constraint on the number of states and actions in the system. Since in the best case, μ is the uniform distribution over $\mathcal{S} \times \mathcal{A}$, and there can be at most C_{unif} state-action pairs. So now we are effectively in the tabular setting and we can obtain some pretty strong guarantees.

Proposition 2. *In the batch RL setting, if Assumption 1 holds, then there is an algorithm such that*

$$V^\star - V(\hat{\pi}) \leq \tilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^2}\left(\sqrt{\frac{C_{\text{unif}}}{n}} + V_{\max}\gamma|\mathcal{S}|\sqrt{\frac{C_{\text{unif}}}{n}}\right)\right)$$

Proof sketch. If $C_{\text{unif}} < \infty$ then given n samples, we can expect that for any (x,a) pair, we will have n/C_{unif} samples from this (x,a) pair (and this will be concentrated). Thus, we can simply approximate the reward and transition operator using the empirical frequencies:

$$\hat{R}(x,a) = \frac{\sum_{i=1}^n \mathbf{1}\{(x_i,a_i) = (x,a)\} r_i}{\sum_{i=1}^n \mathbf{1}\{(x_i,a_i) = (x,a)\}}, \qquad \hat{P}(x' \mid x,a) = \frac{\sum_{i=1}^n \mathbf{1}\{(x_i,a_i,x_i') = (x,a,x')\}}{\sum_{i=1}^n \mathbf{1}\{(x_i,a_i) = (x,a)\}}.$$

A fairly standard concentration argument would reveal that

$$\forall x,a : \left|\hat{R}(x,a) - R(x,a)\right| \lesssim \sqrt{\frac{C_{\text{unif}}}{n}} =: \Delta_R \qquad \|\hat{P}(\cdot \mid x,a) - P(\cdot \mid x,a)\|_{\text{TV}} \lesssim |\mathcal{S}|\sqrt{\frac{C_{\text{unif}}}{n}} =: \Delta_P.$$

I am omitting constants and logarithmic factors here, and this analysis is not very sharp at all. In particular if you used Bernstein inequality for $\hat{P}$, I think you can shave off a factor of S.

This means that we have a model $\hat{M}$ that is uniformly accurate to the true MDP. If we plan in the model, we will obtain a policy $\hat{\pi}$ that turn out to be near-optimal. In particular

$$\begin{aligned}
V^\star - V(\hat{\pi}) &= \mathbb{E}_{d_0}\left[Q^\star(x_0,\pi^\star(x_0)) - Q^{\hat{\pi}}(x_0,\hat{\pi}(x_0))\right] \\
&\leq \mathbb{E}_{d_0}\left[Q^\star(x_0,\pi^\star(x_0)) - \hat{Q}(x_0,\pi^\star(x_0)) + \hat{Q}(x_0,\hat{\pi}(x_0)) - Q^{\hat{\pi}}(x_0,\hat{\pi}(x_0))\right] \\
&= \mathbb{E}_{d_0}\left[Q^\star(x_0,\pi^\star(x_0)) - \hat{Q}(x_0,\pi^\star(x_0)) + \hat{Q}(x_0,\hat{\pi}(x_0)) - Q^\star(x_0,\hat{\pi}(x_0)) + \gamma\mathbb{E}_{x_1 \sim P_{x_0,\hat{\pi}(x_0)}}(V^\star(x_1) - V^{\hat{\pi}}(x_1))\right]
\end{aligned}$$

(This is basically just the performance difference lemma.) This calculation telescopes, and we will get

$$V^\star - V(\hat{\pi}) \leq \sum_{h=0}^\infty \gamma^h \mathbb{E}_{d_{h,\hat{\pi}}}\left[\left|Q^\star(x,\pi^\star(x)) - \hat{Q}(x,\pi^\star(x))\right| + \left|\hat{Q}(x,\hat{\pi}(x)) - Q^\star(x,\hat{\pi}(x))\right|\right]$$

Let's focus on one of these pairs and one time step h, since the other terms will be analogous

$$\begin{aligned}
\mathbb{E}\left[\left|Q^\star(x,\pi^\star(x)) - \hat{Q}(x,\pi^\star(x))\right|\right] &\leq \mathbb{E}\left[\left|R(x,\pi^\star) - \hat{R}(x,\pi^\star(x))\right| + \gamma\left|\mathbb{E}_{x' \sim P_{x,\pi^\star(x)}} V^\star(x') - \gamma\mathbb{E}_{x' \sim \hat{P}_{x,\pi^\star(x)}}\hat{V}(x')\right|\right] \\
&\leq \Delta_R + \gamma\mathbb{E}\left[\left|\mathbb{E}_{x' \sim P}(V^\star(x') - \hat{V}(x')) + \left(\mathbb{E}_{x' \sim P} - \mathbb{E}_{x' \sim \hat{P}}\right)\hat{V}(x')\right|\right] \\
&\leq \Delta_R + V_{\max}\gamma\Delta_P + \gamma\mathbb{E}\mathbb{E}_{x' \sim P_{x,\pi^\star(x)}}\left|V^\star(x') - \hat{V}(x')\right|
\end{aligned}$$

Note that this calculation does not depend on the distribution over x or the policy $\pi^\star$. This last term is a recursive term, which can be seen as follows:

$$\mathbb{E}\left|V^\star(x') - \hat{V}(x')\right| = \mathbb{E}\left|\max_a Q^\star(x',a) - \max_a \hat{Q}(x',a)\right| \leq \mathbb{E}\left|Q^\star(x',\tilde{\pi}(x)) - \hat{Q}(x',\tilde{\pi}(x))\right|$$

where $\tilde{\pi}(x') = \text{argmax}_a \max\{Q^\star(x', a), \hat{Q}(x', a)\}$ chooses the larger of the two actions. Since we have uniform approximation of the MDP, this recursive term can be controlled similarly, so we get

$$\text{error} \leq \Delta_R + V_{\max}\gamma\Delta_P + \gamma \cdot \text{error} \Rightarrow \text{error} \leq \frac{1}{1-\gamma}\left(\Delta_R + V_{\max}\gamma\Delta_P\right).$$

We plug this in above to get

$$V^\star - V(\hat{\pi}) \leq \frac{2}{1-\gamma}\left(\Delta_R + V_{\max}\gamma\Delta_P\right) \cdot \sum_{h=0}^{\infty}\gamma^h = \frac{2}{(1-\gamma)^2}\left(\Delta_R + V_{\max}\gamma\Delta_P\right).$$

Plugging in the definitions of Δ_R and Δ_P, we'll obtain the result. □

This is kind of a simplistic setting, since it is tabular and we have great coverage of the environment, but it is I think the simplest setting where we can see some of the key analysis techniques for more complex settings.

3 Fitted Q iteration, its guarantee, and analysis

Let's now turn to a setting where there is function approximation. First we will relax the concentrability condition as follows. First recall that d_h^π is the state-action distribution induced by executing policy π for h steps.

{assum:con

Assumption 3 (Concentrability)**.** *There exists a constant C such that*

$$\forall \text{ nonstationary } \pi, h, x, a : \frac{d_h^\pi(x,a)}{\mu(x,a)} \leq C.$$

This assumption requires that the state visitation density ratio is bounded for any non-stationary policy π, which may not even be induced by the function class that we are going to use. We will see why we need this assumption in the proof. However, note that concentrability does not require that the state space is finite, but it does place some constraints on the system dynamics.

Since the state space could be large, we will have to resort to function approximation. A standard way to do this is to try to directly approximate the $Q^\star$ function using a class of functions $\mathcal{F} : \mathcal{S} \times \mathcal{A} \to [0, V_{\max}]$. With this setup, perhaps the most natural algorithm is *Fitted Q iteration*, which initializes $f_0 \in \mathcal{F}$ arbitrarily, and iterates:

$$f_k \leftarrow \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{n} (f(x_i, a_i) - r_i - \gamma \max_{a'} f_{k-1}(x_i', a'))^2$$

The analysis of this method typically relies on a representational condition for the function class $\mathcal{F}$. Recall the definition of the bellman optimality operator

$$\mathcal{T}f : (x,a) \mapsto \mathbb{E}\left[r + \gamma \max_{a'} f(x', a') \mid x, a\right].$$

We use $\mathcal{T}\mathcal{F} := \{\mathcal{T}f : f \in \mathcal{F}\}$ to denote the entire class of functions after applying the Bellman operator.

{assum:com

Assumption 4 (Completeness)**.** *We assume that $\mathcal{T}\mathcal{F} \subset \mathcal{F}$.*

Note that this implies that $Q^\star \in \mathcal{F}$, which is the weaker assumption we would hope is sufficient. This follows since $Q^\star$ is the fixed point of the Bellman operator, and if we repeatedly apply $\mathcal{T}$, we remain in $\mathcal{F}$.

Theorem 5 (FQI guarantee)**.** *The k^{th} iterate of Fitted Q Iteration guarantees that with probability $1-\delta$*

$$V^\star - V^{\pi_k} \leq \mathcal{O}\left(\frac{V_{\max}}{(1-\gamma)^2}\sqrt{\frac{C\log(|\mathcal{F}|/\delta)}{n}}\right) + \frac{2\gamma^k V_{\max}}{1-\gamma}$$

The first term is the estimation error term, which goes to 0 as we get more data. The second term is "optimization error" term that goes to 0 as we do more iterations. This term can always be made arbitrarily small at the expense of more computation.

Proof. We'll use the "norm" notation to make things more concise. Here $\|f\|_{2,\nu}^2 := \mathbb{E}_{x,a\sim\nu} f^2(x,a)$ where ν is a distribution over the state-action pairs.

The first step is a uniform convergence argument. This would yield

$$\forall k : \|f_k - \mathcal{T} f_{k-1}\|_{2,\mu}^2 \le \mathcal{O}\left(\frac{V_{\max}^2 \log |\mathcal{F}|/\delta}{n}\right) =: \Delta_{\text{sq}}$$

We can go through this calculation if you all think it would be useful, but the main point is that we are doing least squares regression, and due to Assumption 4, we are actually realizable. In particular, the minimizer of the risk function $f \mapsto \mathbb{E}_\mu(f(x,a) - r - \gamma \max_{a'} f_{k-1}(x',a'))^2$ is the "bellman backup" function $\mathcal{T} f_{k-1}$, which happens to belong to $\mathcal{F}$. When we write-out the excess risk, we'll see that this equals $\|f - \mathcal{T} f_{k-1}\|_{2,\mu}^2$, which is what we can bound using the generalization argument.

Next, we will use a similar argument to the one above. By performance difference lemma (note that d_{h,π_k} here is a distribution over states *only*!)

$$\begin{aligned}
V^\star - V(\pi_k) &= \sum_{h=0}^{\infty} \gamma^h \mathbb{E}_{d_{h,\pi_k}} \left[Q^\star(x, \pi^\star(x)) - Q^\star(x, \pi_k(x)\right] \\
&\le \sum_{h=0}^{\infty} \gamma^h \left(\|Q^\star - f_k\|_{1, d_{h,\pi_k} \circ \pi^\star} + \|Q^\star - f_k\|_{1, d_{h,\pi_k} \circ \pi_k} \right) \\
&\le \sum_{h=0}^{\infty} \gamma^h \left(\|Q^\star - f_k\|_{2, d_{h,\pi_k} \circ \pi^\star} + \|Q^\star - f_k\|_{2, d_{h,\pi_k} \circ \pi_k} \right).
\end{aligned}$$

Here we are using the norm notation, but this is precisely the inequality we had previously. As before, we can handle both of these terms in a similar way. For an distribution ν induced by a nonstationary policy, we have

$$\begin{aligned}
\|Q^\star - f_k\|_{2,\nu} &\le \|Q^\star - \mathcal{T} f_{k-1}\|_{2,\nu} + \|f_k - \mathcal{T} f_{k-1}\|_{2,\nu} \\
&\le \gamma \sqrt{\mathbb{E}_{x,a\sim\nu} \left[\left| \mathbb{E}_{x'\sim P_{xa}} \max_a Q^\star(x',a) - \max_a f_{k-1}(x',a) \right|^2 \right]} + \|f_k - \mathcal{T} f_{k-1}\|_{2,\nu} \\
&\le \gamma \|Q^\star - f_{k-1}\|_{2, \nu \circ \tilde{\pi}} + \sqrt{C} \|f_k - \mathcal{T} f_{k-1}\|_{2,\mu} \\
&\le \ldots \le \sum_{t=0}^{k-1} \sqrt{C} \gamma^t \|f_{k-t} - \mathcal{T} f_{k-t-1}\|_{2,\mu} + \sup_{\tilde{\nu}} \gamma^k \|Q^\star - f_0\|_{2,\tilde{\nu}} \\
&\le \frac{1}{1-\gamma} \sqrt{C \cdot \Delta_{\text{sq}}} + \gamma^k V_{\max}.
\end{aligned}$$

Here $\tilde{\nu}$ is some distribution induced by a nonstationary policy. Plugging this in to the performance difference calculation yields the theorem. □

4 Some remarks on concentrability

The concentrability assumption is worth thinking about for a bit. It's clear that uniform concentrability imposes an upper bound on the number of state-action pairs in the system, but what about Assumption 3?

This assumption does place some strong constraints on the system and the exploratory policy, but it does not preclude infinitely large state spaces. One nice example is the block MDP model that some of us have been studying for a couple of years.

Concentrability can also be modified if we restrict ourselves to linear functions. If all functions are linear in some feature space $\phi(x,a)$, then the norm

$$\|f - f'\|_{2,\nu}^2 = \mathbb{E}_\nu(\langle f - f', \phi(x,a)\rangle)^2 = (f-f')^\top \Sigma_\nu (f - f')$$

Where $\Sigma_\nu = \mathbb{E}_\nu \phi(x,a)\phi(x,a)^\top$. Exploiting this special structure, it suffices to measure coverage in terms of these second moment matrices. In particular, we can instead assume

$$\lambda_{\max}\left(\Sigma_\mu^{-1/2}\Sigma_{d_h^\pi}\Sigma_\mu^{-1/2}\right) \le C_{\text{linear}}$$

And we can see why as follows:

$$\begin{aligned}\|f\|_{2,\nu}^2 = f^\top \Sigma_\nu f = f^\top \Sigma_\mu^{1/2}\Sigma_\mu^{-1/2}\Sigma_\nu\Sigma_\mu^{-1/2}\Sigma_\mu^{1/2} f = \operatorname{tr}(\Sigma_\mu^{-1/2}\Sigma_\nu\Sigma_\mu^{-1/2}\cdot \Sigma_\mu^{1/2} f f^\top \Sigma_\mu^{1/2}) \\ \le \lambda_{\max}\left(\Sigma_\mu^{-1/2}\Sigma_{d_h^\pi}\Sigma_\mu^{-1/2}\right)\cdot \|f\|_{2,\mu}^2\end{aligned}$$

So for linear functions, this allows us to shift distributions. Note that this is exactly how we used concentrability in the FQI proof, and this is the only place we used it. So for linear functions, you could use this definition and just replace the dependence on C with C_{linear}. As a final remark, this notion clearly allows us to scale to large state spaces, since if μ is supported on the standard basis and the features are bounded in norm, we would always have C_{linear} bounded.

5 Some remarks on completeness

The other assumption we should scrutinize is the completeness assumption. This one is quite a strong representational constraint on the function class. In particular, one thing that is quite disappointing is that the assumption is "non-monotonic" in the sense that a small function class may satisfy this, while a larger class that contains the small class may not.

We would hope that we can get away with $Q^\star \in \mathcal{F}$ only, which is a monotonic assumption. This was shown to be possible in a recent paper by Tengyang Xie and Nan Jiang, but it requires very strong concentrability conditions, stronger than Assumption 3, and it has a non-parametric flavor that loses the $1/\epsilon^2$ convergence rate. The assumption is that $\mu(a \mid s) \ge 1/C_{\mathcal{A}}$ and $P(s' \mid s,a)/\mu(s') \le C_{\mathcal{S}}$. I think the basic idea here is that we can run a tournament, where we always compare two functions, discretized to a certain resolution. In doing so, the bellman back up can be written in terms of a "coarse" state-abstraction of the space, and this allows us to achieve the bayes optimal error using functions that are piecewise constant over the state abstraction.

On the other hand, with weaker concentrability assumptions, this kind of guarantee is not possible. This was proved in a very recent paper by Ruosong Wang, Dean Foster, and Sham. They show that if you have the Linear concentrability condition and $Q^\star$ is a linear function in the feature map, then you cannot avoid $\exp(d)$ sample complexity for batch RL.

6 Other results

A couple of other remarks

1. A popular approach these days is this "minimax" approach where you optimize

$$\operatorname*{argmin}_{Q\in\mathcal{Q}} \max_{w\in\mathcal{W}} |\mathcal{L}(Q,w)|, \qquad \mathcal{L}(Q,w) := \mathbb{E}_\mu\left[w(x,a)\left(r+\gamma\max_{a'} Q(x',a') - Q(x,a)\right)\right].$$

   Intuitively this is saying that you want to have low "average" bellman error on all data-distributions (induced by importance weighting). These algorithms have some nice properties, such as they provide valid confidence intervals under pretty weak conditions.

2. It is also possible to avoid the $1/(1-\gamma)^2$ dependence using various minimax algorithms, as shown in a paper by Tengyang and Nan recently. Algorithmically, one approach is to use "bellman residual minimization"

$$\hat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \mathbb{E}_\mu[(f(x,a) - r - \gamma\max_{a'} f(x',a'))^2] - \mathbb{E}_\mu[(g(x,a) - r - \gamma\max_{a'} f(x',a'))^2]$$

   The intuition is that we'd like to optimize just the first part, but this is susceptible to the so called "double-sampling" issue, since the conditional expectation over x' is outside the square. The second term intuitively

subtracts off the bayes error for the square loss problem to account for this. In the proof, the key lemma is a different performance difference decomposition

$$V(\pi) - V(\pi_f) \leq \frac{1}{1-\gamma}\left(\mathbb{E}_{d_\pi}[\mathcal{T}f - f] + \mathbb{E}_{d_{\pi_f}}[f - \mathcal{T}f]\right)$$

The terms in this decomposition look quite similar to the optimization problem we solve. This is similar to what we did in the FQI proof, except we had terms depending on $Q^\star - f$. To relate these terms to the FQI optimization problem, we incurred another factor if $1/(1-\gamma)$.

3. Lastly, in the absence of coverage conditions, we clearly cannot compete with the optimal policy, but maybe we can get something? This has been studied in a couple of recent papers (by Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill, and another by Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims). These papers introduce some "penalty" to discourage the learned policy from drifting off of the data distribution. With such approaches it is possible to compete with the best policy that is well-supported by the data distribution. However it is much harder to run these algorithms, because the penalty functions typically require some kind of density estimation.

References

1. Tengyang Xie and Nan Jiang. $Q^\star$ approximation schemes for batch reinforcement learning: A theoretical comparison. 2020.

2. Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. 2020.

3. Ruosong Wang, Dean P. Foster, and Sham M. Kakade. What are the statistical limits of offline RL with linear function approximation?

4. Andras Antos, Csaba Szepesvari, and Remi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. 2007.

5. Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. 2020.

6. Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. 2020.