17. Introduction

Batch learning is concerned with problems when a learning algorithm must work with data collected in some manner that is not under the control of the learning algorithm: on a batch of data. In batch RL the data is given in the form of a sequence of trajectories of varying length, where each trajectory is of the form $\tau = (S_0, A_0, R_0, S_1, A_1, R_1, \dots, S_t, A_t, R_t, S_{t+1})$, where A_i is chosen in a causal fashion (based on "past" data), $(R_t, S_{t+1}) \sim Q_{A_t}(S_t)$, where $Q = (Q_a(s))_{s,a}$ is a collection of probability distributions over pairs of reals and states, as usual (when we want to allow stochastic rewards).

Batch RL problems fall into two basic categories:

1. **Value prediction**: Predict the value μv^π of using a policy π from the initial distribution μ, where both μ and v^π are given in an explicit form.
2. **Policy optimization**: Find a good (ideally, near optimal) policy given the batch of data from an MDP.

These two problems are intimately related. On the one hand, a good value predictor can potentially be used to find good policies. On the other hand, a good policy optimizer can also be used to decide about whether the value of some policy is above or below some fixed threshold by appropriately manipulating the data fed to the policy optimizers. One can then put a binary search procedure around this decision routine to find out the value of some policy.

Value prediction problems have some common variations. In policy evaluation, rather than evaluating a policy for some fixed initial distribution, the goal is to estimate the entire value function of the policy. Of course, this is at least as hard as the simpler, initial value estimation problem. However, much of the hardness of the problem is already captured by the initial value estimation problem. In initial value prediction, oftentimes the goal is to predict an interval that contains the true unknown value with a prescribed probability, rather than just producing a "point estimate". In the case of policy evaluation, the analogue is to predict a set that contains the true unknown value function with a prescribed probability. Here, a simpler goal is to estimate confidence intervals for each potential input (state), which when "pasted together" can be visualized as forming a confidence band.

There is also the question of how to collect data. In statistics, the problem of designing a "good way" of collecting the data is called the experimental design problem. The best is of course, if data can be collected in an **active manner**: This is when the data collection strategy changes in response to what data has been collected so far.

The problem of designing good active data collection strategies belongs to the bigger group of designing **online learning** algorithms. These are defined exactly based on that the data is collected in a way that depends on what data has been previously collected. The last segment of the part will be solely devoted to these online learning strategies.

In many applications, active data collection is not an option. There can be many reasons for this: active data collection may be deemed to be risky, expensive, or just technically challenging. When data is collected in a passive fashion, it may simply miss key information that would allow for good solutions. Still, in this case, there may be better and worse ways collecting data. Optimizing **experimental designs** is the problem of choosing good passive data collection strategies that lead to good learning outcomes. This topic came up in the context planning algorithms as they also need to create value function estimates and for this the data collection is better to be planned so that learning can succeed.

Oftentimes though, there is no control over how data is collected. Even worse, the method that was used to collect data may be unknown. When this is the case, not much can be done, as the following example shows:

Consider a bandit problem with two actions, denoted by $0, 1$ and a Bernoulli reward. Assume that the reward distribution is Bernoulli with parameter 0.1 when $a = 1$ and Bernoulli with parameter 0.9 when $a = 0$. Let Z be a random variable, which is normally unavailable, but which, together with the action a taken completely determines the reward. For example, Z could have a Bernoulli distribution with parameter $p = 0.1$, and if action a is chosen, the reward $R(a)$ obtained is

$$R(a) = aZ + (1-a)(1-Z)\,.$$

This is indeed consistent with that $R(a)$ has Bernoulli 0.1 distribution when $a = 1$ and has Bernoulli 0.9 distribution when $a = 0$. Assume now that during data collection the actions are chosen based on Z: $A = \pi(Z)$ with some π. For concreteness, assume that during data collection $A = Z$. Then, the action is random, yet, if the data is composed of pairs that have the distribution shared by $(A, R(A))$, or $(Z, 1)$, clearly no method will be able to properly estimate the mean of $R(0)$ or $R(1)$, let alone choosing the action that leads to a higher reward. It is not hard to construct examples when the conditional mean of the observed data makes an optimal action look worse than a suboptimal action.

This is an example where the correct model cannot be estimated because of the way data is collected: The presence of the spurious correlation between a variable that controls outcomes but is not recorded can easily make the data collected useless, regardless of quantities. This is an instance when the model is unidentifiable even with an infinite amount of data.

When data collection is as arbitrary as in the above example, only a very careful study of the **domain** can tell us whether the model is identifiable or not from the data. Note that this is an activity that involves thinking about the structure of the problem at hand. The best is of course if data collection can be influenced to avoid building up spurious correlations. When data is collected in a causal way (following a policy, while recording both the decisions made and the data is used to make those decisions), spurious correlations are avoided and the remaining problem is to guarantee sufficient "coverage" to achieve statistical efficiency.

How good is the plug-in method?

The **plug-in method** estimates a model and uses the estimated model in place of the real one to solve the problem at hand. Let $M = (\mathcal{S}, \mathcal{A}, P, r)$ be a finite MDP, $\hat{M} = (\mathcal{S}, \mathcal{A}, \hat{P}, \hat{r})$ be an estimate. The estimate can be produced in a number of ways, but from the perspective of the result that comes, how the estimate is produced does not matter.

We consider the discounted case with a discount factor $0 \le \gamma < 1$. We will use $\hat{v}^\pi$ to denote the value function of a policy π in $\hat{M}$ (as opposed to v^π, which is the value function of policy in M), and similarly, we will use $\hat{v}^*$ to denote the optimal value function in $\hat{M}$. We analogously use $\hat{q}^\pi$ and $\hat{q}^*$. Every other quantity that is usually associated with an MDP but which now is associated with $\hat{M}$ receives a "hat". For example, we use $\hat{T}_\pi$ for the policy evaluation operator of memoryless policy π in $\hat{M}$ (either for the state values, or the action-values), while we use $\hat{T}$ to denote the Bellman optimality operator underlying $\hat{M}$ (again, both for the state and action-values).

We start with a generic result about contraction mappings:

Proposition (residual bound): Let $F : V \to V$ be a γ-contraction over a normed vector space V and let $x \in V$ be a fixed-point of F. Then for any $y \in V$,

$$\|x - y\| \le \frac{\|Fy - y\|}{1-\gamma} \,. \tag{1}$$

Proof: By the triangle inequality,

$$\|x-y\| \le \|Fx-Fy\| + \|Fy-y\| \le \gamma\|x-y\| + \|Fy-y\|\,.$$

Reordering and solving for $|x-y|$ gives the result. ■

An immediate implication is that good model estimates are guaranteed to give rise to (relatively) good value estimates.

Proposition (value estimation error): Let $H_\gamma = 1/(1-\gamma)$ and assume that the rewards in M are in the $[0,1]$ interval. For any policy π, the following holds:

$$\begin{aligned}\|v^\pi - \hat v^\pi\|_\infty &\le H_\gamma \left(\|r_\pi - \hat r_\pi\|_\infty + \gamma\|(P_\pi - \hat P_\pi)v^\pi\|_\infty\right) &(2)\\ &\le H_\gamma \left(\|r - \hat r\|_\infty + \gamma H_\gamma\|P - \hat P\|_\infty\right). &(3)\end{aligned}$$

Also,

$$\begin{aligned}\|v^* - \hat v^*\|_\infty &\le H_\gamma \left(\|r - \hat r\|_\infty + \gamma\|(P - \hat P)v^*\|_\infty\right) &(4)\\ &\le H_\gamma \left(\|r - \hat r\|_\infty + \gamma H_\gamma\|P - \hat P\|_\infty\right). &(5)\end{aligned}$$

Similarly,

$$\begin{aligned}\|q^\pi - \hat q^\pi\|_\infty &\le H_\gamma \left(\|r - \hat r\|_\infty + \gamma\|(P - \hat P)v^\pi\|_\infty\right) &(6)\\ &\le H_\gamma \left(\|r - \hat r\|_\infty + \gamma H_\gamma\|P - \hat P\|_\infty\right). &(7)\end{aligned}$$

and

$$\begin{aligned}\|q^* - \hat q^*\|_\infty &\le H_\gamma \left(\|r - \hat r\|_\infty + \gamma\|(P - \hat P)v^*\|_\infty\right) &(8)\\ &\le H_\gamma \left(\|r - \hat r\|_\infty + \gamma H_\gamma\|P - \hat P\|_\infty\right). &(9)\end{aligned}$$

Note that in general the value estimates are more sensitive to errors in the transition probabilities then in the rewards. In particular, the transition errors can be magnified by a factor as large as H_γ, while the reward errors are magnified by at most H_γ. Also note that sometimes one can obtain tighter estimates with stopping earlier in the derivations of these bounds. We will see some examples of how this can help later.

Proof: To reduce clutter, we write $\|\cdot\|$ for $\|\cdot\|_\infty$. Let $F = \hat T_\pi$, where $\hat T_\pi$ is defined via $\hat T_\pi v = \hat r_\pi + \gamma \hat P_\pi v$. By the residual bound (1),

$$\|\hat v^\pi - v^\pi\| \le H_\gamma \|\hat T_\pi v^\pi - v^\pi\| = H_\gamma \|\hat T_\pi v^\pi - T_\pi v^\pi\| \le H_\gamma \left(\|r_\pi - \hat r_\pi\| + \gamma \|(P_\pi - \hat P_\pi) v^\pi\| \right).$$

The second inequality follows from separating v^π from the second term and bounding it using $\|v^\pi\| \le H_\gamma$ and also using that $r_\pi = M_\pi r$, $\hat r_\pi = M_\pi \hat r$, $P_\pi = M_\pi P$ and $\hat P_\pi = M_\pi \hat P$ and finally using that M_π is a nonexpansion. The remaining inequalities can be obtained in an entirely analogous manner and hence their proof is omitted. ■

The result just shown suffices to quantify the size of the value errors. For quantifying the **policy optimization error** that results from finding an optimal (or near optimal) policy for $\hat M$, recall the Policy Error Bound from Lecture 6:

Lemma (Policy error bound - I.): Let π be a memoryless policy and choose a function $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $\epsilon \ge 0$. Then, the following hold:

1. If π is ϵ-**optimizing** in the sense that $\sum_a \pi(a|s) q^*(s,a) \ge v^*(s) - \epsilon$ holds for every state $s \in \mathcal{S}$ then π is $\epsilon/(1-\gamma)$ suboptimal: $v^\pi \ge v^* - \frac{\epsilon}{1-\gamma} \mathbf{1}$.
2. If π is greedy with respect to q then π is 2ϵ-optimizing with $\epsilon = \|q - q^*\|_\infty$ and thus

$$v^\pi \ge v^* - \frac{2\|q - q^*\|_\infty}{1-\gamma} \mathbf{1} \,.$$

This leads to the following result:

Theorem (bound on policy optimization error): Assume that the rewards both in M and $\hat M$ belong to the $[0,1]$ interval. Take any $\varepsilon > 0$ and ε-optimal policy π in $\hat M$: $\hat v^\pi \ge \hat v^* - \varepsilon \mathbf{1}$. Then, π is δ-optimal in M with δ satisfying

$$\delta \le (1 + 2\gamma) H_\gamma \varepsilon + 2 H_\gamma^2 \left\{ \|r - \hat r\|_\infty + \gamma \|(P - \hat P) v^*\|_\infty \right\} .$$

Note that, up to a small constant factor, the optimization error is magnified by a factor of H_γ, the reward errors are magnified by a factor of H_γ^2, while the transition errors can get magnified by a factor of up to H_γ^3, depending on the magnitude of v^*.

Proof: Let π be a policy as in the theorem statement. Our goal now is to use the first part of the "Policy error bound", i.e., that π is ε'-optimizing with some $\varepsilon' > 0$.

On the one hand, we have

$$M_\pi \hat{q}^\pi = \hat{v}^\pi \ge \hat{v}^* - \varepsilon \boldsymbol{1} = M\hat{q}^* - \varepsilon \boldsymbol{1} \ge M\hat{q}^\pi - \varepsilon \boldsymbol{1}\,.$$

Let z be defined by $M_\pi \hat{q}^\pi = M\hat{q}^\pi + z$. From the previous inequality, we know that $\|z\|_\infty \le \varepsilon$. We also have

$$\begin{aligned}
M_\pi q^* &= M_\pi \hat{q}^\pi + M_\pi(q^* - \hat{q}^\pi) \\
&= M\hat{q}^\pi + M_\pi(q^* - \hat{q}^\pi) + z \\
&= Mq^* + M\hat{q}^\pi - Mq^* + M_\pi(q^* - \hat{q}^\pi) + z \\
&\ge Mq^* - (2\|\hat{q}^\pi - q^*\| + \varepsilon)\boldsymbol{1} \\
&= v^* - (2\|\hat{q}^\pi - q^*\| + \varepsilon)\boldsymbol{1}\,.
\end{aligned}$$

Hence, by Part 1. of the "Policy Error Bound I." lemma from above,

$$v^\pi \ge v^* - H_\gamma(2\|\hat{q}^\pi - q^*\| + \varepsilon)\boldsymbol{1}\,.$$

By the triangle inequality and the assumption on π,

$$\|\hat{q}^\pi - q^*\|_\infty \le \|\hat{q}^\pi - \hat{q}^*\|_\infty + \|\hat{q}^* - q^*\|_\infty \le \gamma\varepsilon + \|\hat{q}^* - q^*\|_\infty\,.$$

By Eq. (8),

$$\|q^* - \hat{q}^*\|_\infty \le H_\gamma \left(\|r - \hat{r}\|_\infty + \gamma \|(P - \hat{P})v^*\|_\infty \right)\,.$$

The result is obtained by chaining the inequalities:

$$\begin{aligned}
\|v^* - v^\pi\|_\infty &\le H_\gamma(2\|\hat{q}^\pi - q^*\| + \varepsilon) \\
&\le H_\gamma \left\{ 2\gamma\varepsilon + 2H_\gamma \left(\|r - \hat{r}\|_\infty + \gamma\|(P - \hat{P})v^*\|_\infty \right) + \varepsilon \right\}\,.
\end{aligned} \qquad \blacksquare$$

Model estimation error: Tabular case

As usual, it is worthwhile to clean up the foundations by considering the tabular case. In this case, the model can be estimared by using sample means. To allow for a unified presentation, let the data available be given in the form of triplets of the form $E_i = (S_i, A_i, R_i, S_{i+1})$ where $i = 1, \dots, n$ and $S_{i+1} \sim P_{A_i}(S_i)$ given $E_1, \dots, E_{i-1}, S_i, A_i$ and $\mathbb{E}[R_i | S_i, A_i, E_1, \dots, E_{i-1}] = r_{A_i}(S_i)$. Introducing the visit counts

$$N(s,a,s') = \sum_{i=1}^n \mathbb{I}(S_i = s, A_i = a, S_{i+1} = s')$$

and $N(s,a) = \sum_{s'} N(s,a,s')$, provided that the visit count for (s,a) is positive, for the transition probability estimates we have

$$\hat{P}_a(s,s') = \frac{N(s,a,s')}{N(s,a)}$$

and for the reward estimate we have

$$\hat{r}_a(s) = \frac{1}{N(s,a)} \sum_{i=1}^n \mathbb{I}(S_i = s, A_i = a) R_i \,.$$

For ensuring that these are always defined, let $\hat{P}_a(s)$ be the uniform distribution over the states and let $\hat{r}_a(s) = 0$ when $N(s,a) = 0$. From the perspective of the results to be presented, the particular values chosen here do not matter.

Consider now the simple case when the above triplets are so that for each state-action pair (s,a), $N(s,a) = n(s,a)$ for some deterministic counts $(n(s,a))_{s,a}$. Say, one has access to a generative model (simulator) and for each state-action pair the model is used to generate a fixed number of independent transitions. In this case, one can use Hoeffding's inequality.

In particular, defining

$$\beta(n,\zeta) = \sqrt{\frac{\log\left(\frac{\mathrm{SA}}{\zeta}\right)}{2n}}$$

provided that $R_i \in [0,1]$, Hoeffding's inequality gives that with probability $1-2\zeta$, for any s,a,

$$\begin{aligned} |\hat{r}_a(s) - r_a(s)| &\le \beta(n(s,a),\zeta)\,, \\ |\langle \hat{P}_a(s) - P_a(s), v^* \rangle| &\le H_\gamma \beta(n(s,a),\zeta)\,, \end{aligned}$$

from which it follows that with probability $1-2\zeta$,

$$\begin{aligned} \|\hat{r} - r\|_\infty &\le \beta(n_{\min},\zeta)\,, \\ \|(\hat{P} - P)v^*\|_\infty &\le H_\gamma \beta(n_{\min},\zeta)\,, \end{aligned}$$

where $n_{\min} = \min_{s,a} n(s,a)$. Plugging the obtained deviation bound into our policy suboptimality bound, we get that with probability $1-\zeta$,

$$\delta \le (1+2\gamma) H_\gamma \varepsilon + 2 H_\gamma^2 (1+\gamma H_\gamma) \beta(n_{\min},\zeta)\,.$$

One can alternatively write this in terms of the total number of observations, n. The best case is when $n(s,a) = n_{\min}$ for all (s,a) pairs, in which case $n = \mathrm{SA} n_{\min}$ and the above bound gives

$$\delta \le (1+2\gamma) H_\gamma \varepsilon + 2 H_\gamma^2 (1+\gamma H_\gamma) \sqrt{\mathrm{SA} \frac{\log\left(\frac{\mathrm{SA}}{\zeta}\right)}{2n}}\,.$$

It follows that for any target suboptimality δ_{trg}, as long as n, the number of observations satisfies

$$n \ge \frac{8 H_\gamma^6 SA \log\left(\frac{\mathrm{SA}}{\zeta}\right)}{\delta_{\text{trg}}^2}\,,$$

we are guaranteed that the optimal policy of the estimated model is at most δ_{trg} suboptimal. As we shall see soon, the optimal dependence on the horizon H_γ is cubic, unlike the dependence shown here.

Notes

Between batch and online learning

In applications it may happen that one can change the data collection strategy a limited number of times. This creates a scenario that is in between batch and online learning. This setting can be thought to be between batch and online learning. From the perspective of online learning, this is learning in the presence of constraints on the data collection strategy. One such widely studied constraint is the number of switches of the data collection strategy. As it happens, only very few switches are necessary to get the full power of online learning and this is not really specific to reinforcement learning but follows because the empirical distribution converges are a slow rate to the true distribution. For parametric problems, the rate is $O(1/\sqrt{n})$ where n is the number of observations. Thus, to change "accuracy" of the estimates of any quantity in a significant fashion, the sample size should increase by much, which means, few changes to the data collection are sufficient. In other words, there is no reason to change the data collection strategy before one obtains sufficient new evidence that can help with deciding in what way the data collection strategy should be changed. This usually means that with only logarithmically many changes in the total sample size, one gets the full power of online methods.

Batch RL with no access to state information

For simplicity, we stated the batch learning problem in a way that assumes that the states in the transitions are observed. This may be seen as problematic. One "escape" is to treat the whole history as the state: Indeed, in a causal, controlled stochastic process, the history can always be used as a Markov state. Because of this, the assumption that the state is observed is not restrictive, though the state space becomes exponential in the length of the trajectories. This reduces to the problem to learning in large state-space MDPs. Of course, even lower bounds for planning tell us that in lack of extra structure, all algorithms need a sample size proportional to the size of the state-action space, hence, one needs to add extra structure to deal with this case, such as function approximation. It also holds that if one uses, say, linear function approximation, then only the features of the states (or state-action pairs) need to be recorded in the data.

Causal reasoning and batch RL

Whether a causal effect can be learned from a batch of data (to be more precise, from data drawn from a specific distribution) is the topic of **causal reasoning**. In batch RL, the "effect" is the value of a policy, which, in the language of causal reasoning, would be called a multistage treatment. As the example in the text shows, in batch RL, just because of our assumptions on how the data is collected, the identifiability problem is just "assumed away". When the assumption on how the data is generated/collected is not met, the tools of causal reasoning can potentially be still used. It is important to emphasize though that there is no causality without assuming causality. The statements that causal reasoning can make are conditional on the data sampling assumptions met. Even "causal discovery" is contingent on these assumptions. However, with care, oftentimes it is possible to argue for that some suitable assumptions are met (e.g., arguing based on what information is available at what time in a process), in which case, the nontrivial tools of causal reasoning may be very useful.

Nevertheless, especially in engineered systems, our standard data collection assumptions are reasonable and can be arranged for, though in large engineered systems, mistakes, such as not logging critical quantities may happen. One example of this is an action to be taken is overriden by some part of a system, which will, say, later be turned off. Clearly, if no one logs the actual actions taken, the effects of actions become unidentifiable. As we shall see later, batch RL and the causality literature share some of their vocabulary, such as "instrumental variables", "propensity scores", etc.

Plug-in or certainty equivalence

Plug-in generally means that a model is estimated and then is used as if it was the "true" model. In control, when a controller (policy) is derived with this approach, this is known as the "certainty equivalence" controller. The "certainty equivalence principle" states that the "random" errors can be neglected. The principle originates from the observation that in various scenarios, the optimal controller (optimal policy) has a special form that confirms this principle. In particular, this was first observed in the control of linear quadratic Gaussian control, where the optimal controller can be obtained by solving for the optimal control under perfect state information then substituting optimal state prediction for the the perfect state information. This strict optimality result is quite brittle. As we shall see soon, from the perspective of minimax optimality, certainty equivalent policies are not a bad choice.

Bibliographic remarks

In the early RL literature, online learning was dominant. When people tried to apply RL to various "industrial"/"applied" settings, they were forced to think about how to learn from data collected before learning starts. One of the first papers to push this agenda is the following one:

- Tree-Based Batch Mode Reinforcement Learning Damien Ernst, Pierre Geurts, Louis Wehenkel; 6(18):503-556, 2005.

Earlier mentions of "batch-mode RL" include

- Efficient Value Function Approximation Using Regression Trees (1999) by Xin Wang , Thomas G. Dietterich, Proceedings of the IJCAI Workshop on Statistical Machine Learning for Large-Scale Optimization. pdf

Even in online learning, efficient learning may force one to save all the data to be used for learning. The so-called LSTD algorithm, and later the LSPI algorithm, were explicitly proposed to address this challenge:

- J. A. Boyan. Technical update: least-squares temporal difference learning. Machine Learning, 49 (2-3):233–246, 2002.
- M. G. Lagoudakis and R. Parr. Least-squares policy iteration. Journal of Machine Learning Research, 4:1107–1149, 2003a.

Off-policy learning refers to the case when an algorithm needs to produce value function (or action-value function) estimates for some policy and the data available is **not** generated by the policy to be evaluated. In all the above examples, we are thus in the setting of off-policy learning. The policy evaluation problem, accordingly, is often called the **off-policy policy evaluation** (OPPE) problem, while the problem of finding a good policy is called the **off-policy policy optimization** (OPPO) problem.

For a review of the literature of around 2012, consult the following paper:

- S. Lange, T. Gabel, M. Riedmiller (2012) Batch Reinforcement Learning. In: M. Wiering, M. van Otterlo (eds) Reinforcement Learning. Adaptation, Learning, and Optimization, vol 12. Springer, Berlin, Heidelberg pdf

Batch RL / 18. Sample complexity in finite MDPs

18. Sample complexity in finite MDPs

Let $Z = \mathcal{S} \times \mathcal{A}$ be the set of state-action pairs. A Z**-design** assigns a count to every member of Z, that is, to every state-action pair. In the last lecture we saw that

$$n = \tilde{O}\left(\frac{H^6 \mathrm{SA}}{\delta_{\text{trg}}}\right)$$

samples are sufficient to obtain a δ_{trg}-suboptimal policy with high probability provided that data is generated from a Z-design that assigns the same count to each state-action pair and to get a policy one uses the straightforward **plug-in approach** that estimates the rewards and transitions using empirical estimates and uses the policy that is optimal with respect to the estimated model. Above, the dependence on the number of state-action pairs is optimal, but the dependence on the horizon $H = \frac{1}{1-\gamma}$ is suboptimal. In the first half of this lecture, I sketch how the analysis presented in the previous lecture can be improved to get the optimal cubic dependence, together with a sketch that shows that the cubic dependence is indeed optimal.

In the second half of the lecture, we consider policy-based data collection, or experimental designs, where the goal is to find a near optimal policy from an initial state, where the data consists of trajectories obtained by rolling out the data-collection policy from the said initial state. Here, we will show a lower bound that shows that the sample complexity in this case is at least as large $\Omega(\mathrm{A}^{\min(\mathrm{S},H)})$, which shows that there exist an exponential separation between both Z-designs and policy-based designs, and also between passive and active learning. To see the latter, note that in the presence of a simulator, with only a reset to an initial state, one can use approximate policy iteration with rollouts, or Politex with rollouts, to get a policy that is near-optimal when started from the initial state that one can reset to but with polynomially many samples in S, A and H (cf. Lecture 8 and Lecture 14).

Improved analysis of the plug-in method: First attempt

The improvement in the analysis of the plug-in method comes from two sources:

1. Using a version of the value-difference identity and avoiding the use of the policy error bound
2. Using Bernstein's inequality in place of Hoeffding's inequality

In this section, we focus on the first aspect. The second aspect will be considered in the next section.

We continue to use the same notation as in the previous lecture. In particular, M denotes the "true" MDP, $\hat{M}$ denotes the estimated MDP and we put $\hat{\cdot}$ on quantities related to this second MDP. We further let π^* be one of the memoryless optimal policies of M. For simplicity, we will assume that **the reward function in $\hat{M}$ is the same as in M**: As we have seen, the higher order term in our error bound came from errors in the transition probability; the simplifying assumption allows us to focus on reducing this term while minimizing clutter. The arguments are easy to extend to the case when $\hat{r} \neq r$.

Let $\hat{\pi}$ be a policy whose suboptimality in M we want to bound. The idea is to bound the suboptimality of $\hat{\pi}$ by its suboptimality in $\hat{M}$ and also by how much value functions for **fixed policies** differ when we switch from P to $\hat{P}$. In particular, we have

$$\begin{aligned} v^* - v^{\hat{\pi}} &= v^* - \hat{v}^* \,+\, \hat{v}^* - v^{\hat{\pi}} \\ &\le v^{\pi^*} - \hat{v}^{\pi^*} \,+\, \underbrace{\hat{v}^* - \hat{v}^{\hat{\pi}}}_{\text{opt. error}} \,+\, \hat{v}^{\hat{\pi}} - v^{\hat{\pi}}\,, \end{aligned} \tag{1}$$

where $\hat{\pi}^*$ denotes an optimal policy in $\hat{M}$ and the inequality holds because $\hat{v}^* = \hat{v}^{\hat{\pi}^*} \ge \hat{v}^{\pi^*}$. The term marked as "opt. error" is the optimization error that arises when $\hat{\pi}$ is not (quite) optimal in $\hat{M}$. This term is controlled by the choice of $\hat{\pi}$. For simplicity, assume for now that $\hat{\pi}$ is an optimal policy in $\hat{M}$, so that we can drop this term. We further assume that $\hat{\pi}$ is a deterministic optimal policy of $\hat{M}$.

It remains to bound the first and last terms. Both of these terms have the form $v^\pi - \hat{v}^\pi$, i.e., the difference between the value functions of the same policy π in the two MDPs (here, π is either π^* or $\hat{\pi}$). This difference, similar to the value difference identity, can be expressed as a function of the difference $P - \hat{P}$, as shown in the next result:

Lemma (value difference from transition differences): Let M and $\hat{M}$ be two MDPs sharing the same state-action space, rewards, but differing in their transition probabilities. Let π be a memoryless policy over the shared state-action space of the two MDPs. Then, the following identities holds:

$$v^\pi - \hat v^\pi = \gamma\underbrace{(I-\gamma P_\pi)^{-1} M_\pi (P-\hat P)\hat v^\pi}_{\delta(\hat v^\pi)}\,, \qquad (2)$$

$$\hat v^\pi - v^\pi = \gamma\underbrace{(I-\gamma \hat P_\pi)^{-1} M_\pi (\hat P-P) v^\pi}_{\hat\delta(v^\pi)}\,. \qquad (3)$$

Proof: We only need to prove (2) since (3) follows from this identity by symmetry. Concerning the proof of (2), we start with the closed form expression for value functions. From this we get

$$v^\pi - \hat v^\pi = (I-\gamma P_\pi)^{-1} r_\pi - (I-\gamma \hat P_\pi)^{-1} r_\pi\,.$$

Inspired by the elementary identity that states that $\frac{1}{1-x} - \frac{1}{1-y} = \frac{x-y}{(1-x)(1-y)}$, we calculate

$$\begin{aligned} v^\pi - \hat v^\pi &= (I-\gamma P_\pi)^{-1}\Big[(I-\gamma\hat P_\pi)-(I-\gamma P_\pi)\Big](I-\gamma\hat P_\pi)^{-1} r_\pi \\ &= \gamma (I-\gamma P_\pi)^{-1}\Big[P_\pi - \hat P_\pi\Big](I-\gamma\hat P_\pi)^{-1} r_\pi \\ &= \gamma (I-\gamma P_\pi)^{-1} M_\pi\Big[P-\hat P\Big]\hat v^\pi\,, \end{aligned}$$

finishing the proof. ■

Note that in (3), the empirical transition kernel $\hat P$ appears through its inverse by left-multiplying $M_\pi(\hat P - P)$, while in (2), through $\hat v^\pi$, it appears by right-multiplying the same deviation term. In the remainder of this section we use (3), but in the next section we will use (2).

Combining (3) with our previous inequality, we immediately get that

$$v^* - v^{\hat\pi} \le \frac{\gamma}{1-\gamma}\left[\|(P-\hat P)v^{\pi^*}\|_\infty + \|(P-\hat P)v^{\hat\pi}\|_\infty\right]. \qquad (4)$$

Assume that $\hat P$ is obtained by sampling m next states at each state-action pair. By Hoeffding's inequality and a union bound over the state-action pairs, for any fixed $v \in [0,H]^{SA}$ and $0 \le \zeta < 1$, with probability $1-\zeta$, we have

$$\|(P-\hat{P})v\|_\infty = H\sqrt{\frac{\log(SA/\zeta)}{2m}} \tag{5}$$

and in particular with $v = v^{\pi^*}$, we have

$$\|(P-\hat{P})v^{\pi^*}\|_\infty = \tilde{O}\left(H/\sqrt{m}\right).$$

Controlling the second term in (4) requires more care as $\hat{\pi}$ is random and depends on the same data that is used to generate $\hat{P}$. To deal with this term, we use another union bound. Let $\tilde{V} = \{v^\pi \,:\, \pi : \mathcal{S} \to \mathcal{A}\}$ be the set of all possible value functions that we can obtain by considering deterministic policies. Since by construction $\hat{\pi}$ is also a deterministic policy, $\hat{v}^{\hat{\pi}} \in \tilde{V}$. Hence,

$$\|(P-\hat{P})\hat{v}^{\hat{\pi}}\|_\infty \le \sup_{v \in \tilde{V}} \|(P-\hat{P})v\|_\infty .$$

and thus by a union bound over the $|\tilde{V}| \le A^S$ functions v in $\tilde{V}$, we get that with probability $1-\zeta$,

$$\|(P-\hat{P})\hat{v}^{\hat{\pi}}\|_\infty \le H\sqrt{\frac{\log(SA|\tilde{V}|/\zeta)}{2m}} = H\sqrt{\frac{\log(SA/\zeta)+S\log(A)}{2m}} = \tilde{O}\left(H\sqrt{S/m}\right).$$

Putting things together, we see that

$$v^* - v^{\hat{\pi}} = \tilde{O}\left(H^2\sqrt{S/m}\right),$$

which reduces the dependence on H of the sample size bound from H^6 to H^4. As we shall see soon, this is not the best possible dependence on H. This method also falls short of giving the best possible dependence on the number of states. In particular, inverting the above bound, we see that with this method we can only guarantee a δ-optimal policy if the total number of samples, $n = SAm$ is at least

$$\tilde{O}(S^2AH^4/\delta^2)$$

while below we will see that the optimal bound is $\tilde{O}(SAH^3/\delta^2)$.

Improved analysis of the plug-in method: Second attempt

There are two further ideas that help one achieve the sample complexity which will be seen to be optimal. One is to use what is known as Bernstein's inequality in place of Hoeffding's inequality, together with a clever observation on the "total variance" and the second is to improve the covering argument. The first idea helps with improving the horizon dependence, the second helps with improving the dependence on the number of states. In this lecture, we will only cover the first idea and sketch the second.

Bernstein's inequality is a classic result in probability theory:

Theorem (Bernstein's inequality): Let $b > 0$ and let $X_1, \dots, X_m \in [0, b]$ be an i.i.d. sequence and define $\bar{X}_m$ as the sample mean of this sequence: $\bar{X}_m = \frac{1}{m}(X_1 + \dots + X_m)$. Then, for any $\zeta \in (0, 1)$, with probability at least $1 - \zeta$,

$$|\bar{X}_m - \mathbb{E}[X_1]| \le \sigma \sqrt{\frac{2\log(2/\zeta)}{m}} + \frac{2}{3}\,\frac{b\log(2/\zeta)}{m}\,,$$

where $\sigma^2 = \mathrm{Var}(X_1)$.

To set expectations, it will be useful to compare this bound to Hoeffding's inequality. In particular, in the setting of the lemma Hoeffding's inequality also applies and gives

$$|\bar{X}_m - \mathbb{E}[X_1]| \le b\sqrt{\frac{\log(2/\zeta)}{2m}}\,.$$

Since in our case $b = H$ (the value functions take values in the $[0, H]$ interval), using $m = H^3/\delta^2$ (which would give rise to the optimal sample size), Hoeffding's inequality gives a bound of size $HH^{-3/2}\delta = H^{-1/2}\delta$ (cf. (5)). This is a problem: Ideally, we would like to see $H^{-1}\delta$ here, because inequality \cref{eq:vdeb} introduces an additional H factor.

We immediately see that for Bernstein's inequality to make a difference, just focusing on the first term in Bernstein's inequality, we need $\sigma = O(H^{1/2})$. In fact, since $b/m = H^{-2}\delta^2 = o(H^{-1}\delta)$, we see that this is also sufficient to take off the H factor from the sample complexity bound. It thus remains to be seen whether the variance could indeed be this small.

To find this out, fix a state-action pair (s, a) and let $S'_1, \dots, S'_m \sim P_a(s)$ be an i.i.d. sequence of next states at (s, a). Then, $((\hat{P} - P)v^\pi)(s, a) = (\hat{P}_a(s) - P_a(s))v^\pi$ has the same distribution as

$$\Delta(s,a) = \frac{1}{m}\sum_{i=1}^m v^\pi(S_i') - P_a(s)v^\pi\,.$$

Defining $X_i = v^\pi(S_i')$ and $\sigma_\pi^2(s,a) = \mathrm{Var}(X_1)$, we see that it is $\sigma_\pi(s,a)$ that appears when Bernstein's inequality is used to bound $((\hat{P}-P)v^\pi)(s,a)$. It remains to be seen how large values $\sigma_\pi(s,a)$ can take on. Sadly, one can quickly discover that the range of $\sigma_\pi(s,a)$ is sometimes also as large as H. Is Bernstein's inequality a dead-end then?

Of course, it is not, otherwise we would not have introduced it. In particular, a better bound is possible by directly bounding the maximum-norm of

$$\delta(v^\pi) = (I-\gamma P_\pi)^{-1} M_\pi (P-\hat{P})v^\pi\,,$$

which is close to the actual term that we need to bound. Indeed, by \cref{(2)} from the value difference lemma, $v^\pi - \hat{v}^\pi = \gamma\delta(\hat{v}^\pi)$ and thus

$$v^\pi - \hat{v}^\pi = \gamma\delta(v^\pi) + \gamma(\delta(\hat{v}^\pi) - \delta(v^\pi))\,.$$

The second term on the right-hand side is of order $1/m$ (since $(P-\hat{P})(\hat{v}^\pi - v^\pi)$ appears there and both $P-\hat{P}$ and $\hat{v}^\pi - v^\pi$ have been seen to be of order $1/\sqrt{m}$). As we expect $\delta(v^\pi)$ to be of order $1/\sqrt{m}$, we will focus on this term.

For simplicity, take now the case when π is a fixed, nonrandom policy (we need to bounded $\delta(v^\pi)$ for $\pi=\pi^*$ and also for $\pi=\hat{\pi}$, the second of which is random). In this case, by a union bound and Bernstein's inequality, with probability $1-\zeta$,

$$|(P-\hat{P})v^\pi| \le \sqrt{\frac{2\log(2SA/\zeta)}{m}}\sigma_\pi + \frac{2H}{3}\frac{\log(2/\zeta)}{m}\mathbf{1}\,.$$

Multiplying both sides by $(I-\gamma P_\pi)^{-1}M_\pi$, using a triangle inequality and the special properties of $(I-\gamma P_\pi)^{-1}M_\pi$, we get

$$\begin{aligned}|\delta(v^\pi)| &\le (I-\gamma P_\pi)^{-1}M_\pi|(P-\hat{P})v^\pi| \\ &\le \sqrt{\frac{2\log(2SA/\zeta)}{m}}(I-\gamma P_\pi)^{-1}M_\pi\sigma_\pi + \frac{2H^2}{3}\frac{\log(2SA/\zeta)}{m}\mathbf{1}\,. \end{aligned} \qquad (6)$$

The following beautiful result, whose proof is omitted, gives an $O(H^{3/2})$ bound on the first term appearing on the right-hand side of the above display:

Lemma (total discounted variance bound): For any discounted MDP M and policy π in M, $\|(I-\gamma P_\pi)^{-1} M_\pi \sigma_\pi\|_\infty \le \sqrt{\frac{2}{(1-\gamma)^3}}$.

Since the bound that we get from here is $H^{3/2}$ and not H^2, “we are saved”. Indeed, plugging this into (6) gives

$$\|\delta(v^\pi)\|_\infty \le 2\sqrt{\frac{H^3 \log(2SA/\zeta)}{m}} + \frac{2H^2}{3}\frac{\log(2SA/\zeta)}{m}\,,$$

which holds with probability $1-\zeta$. Choosing $m = H^3/\delta^2$, we see that both terms are $O(\delta)$. It remains to show that a similar result holds for $\pi = \hat{\pi}$. If we use the union bound that we used before, we introduce an extra S factor. Avoiding this extra S factor requires new ideas, but with these we get the following result:

Theorem (upper bound for Z-designs): Let $\hat{\pi}$ be an optimal policy in the MDP whose transition kernel is $\hat{P}$, a kernel estimated based on a sample of m next states from each state-action pair. Letting $0 \le \zeta < 1$ and $0 \le \delta \le \sqrt{H}$, if

$$m \ge \frac{c\gamma H^3 \log(SAH/\delta)}{\delta^2}$$

then with probability $1-\zeta$, $\hat{\pi}$ is δ-optimal, where c is a universal constant. In short, for any $0 \le \delta \le \sqrt{H}$ there exist an algorithm that produces a δ-optimal policy from a total number of

$$\tilde{O}\left(\frac{\gamma S A H^3}{\delta^2}\right)$$

samples under a uniform Z-design.

It remains to be seen whether the same sample complexity holds for larger values of δ, e.g., for $\delta = H/2$.

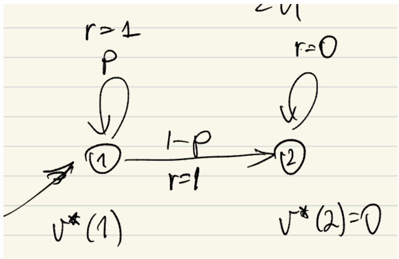
Lower bound for Z-designs

A natural question is whether we can improve on the H^3SA/δ^2 upper bound, or whether this can be matched by a lower bound. For this, we have the following result:

Theorem (lower bound for Z-designs): Any algorithm that uses Z-designs and is guaranteed to produce a δ-optimal policy needs at least $\Omega(H^3SA/\delta^2)$ samples.

Proof (sketch): As we have seen in the proof of the upper bound, the key to achieve the cubic dependence was that the sample mean $\bar{X}_m$ of m i.i.d. bounded random variables is within a distance of $\sigma\sqrt{1/m}$ to the true mean. In a way, the converse of this is also true: It is "quite likely" that the distance between the sample and true mean is this large. This is not too hard to see for specific distributions, such as when the X_i are normally distributed, or when X_i are Bernoulli distributed (in a way, this is the essence of the central-limit theorem, though the central-limit theorem is restricted for $m \to \infty$).

So how can we use this to establish the lower bound? In an MDP the randomness comes either from the rewards or the transitions. But in the upper bound above, the rewards were given, so the only source of randomness is transitions. Also, the cubic dependence must hold even if the number of states is a constant. What all this implies is that somehow learning the transition structure with a few states is what makes the sample complexity large as $\gamma \to 1$ (or $H \to \infty$). Clearly, this can only happen if the (small) MDP has self-loops. The smallest example of an MDP with a self-loop is if one has an action and state such that taking that action from that state leads to same action with some positive probability, while with the complementary probability the next state is some other state. This leads to the structure shown on the figure on the right.



As can be seen, there are two states. The transition at the first state, call it state 1, is stochastic and leads to itself with probability p, while it leads to state 2 with probability $1-p$. The reward associated with both transitions is 1. The second state, call it state 2, has a self-loop. The reward associated with this transition is zero.

There are no actions (alternatively, there is only one action at both states). However, if we can show that in the lack of knowledge of p, estimating the value of state 1 up to a precision of δ takes $\Omega(H^3/\delta^2)$ samples, the sample complexity result will follow. In particular, if we repeat the transition structure A times (sharing the same two states), one can make the value of p for one of this actions ever slightly so different from the others so that its value differs by (say) 2δ from the others. Then, by construction, a learner who uses fewer than $\Omega(AH^3/\delta^2)$ total samples at state 1 will not be able to reliably tell the difference between the value of the special action and the other actions, hence, will not be to choose the right action and will thus be unable to produce a δ-optimal policy. To also add the state dependence, the structure can then be repeated S times.

So it remains to be seen whether the said sample complexity result holds for estimating the value of state 1. Rather than giving a formal proof, we give a quick heuristic argument, hoping that readers will find this more intuitive.

The starting point for this heuristic argument is the general observation that **sample complexity questions concerning estimation problems are essentially questions about the sensitivity of the quantity to be estimated to the unknown parameters**. Here, sensitivity means how much the quantity changes if we change the underlying parameter. This sensitivity for small deviations and a single parameter is exactly the derivative of the quantity of interest with respect to the parameter.

In our special case, the value of state 1, call it $v_p(1)$ (also showing the dependence on p) is the quantity to be estimated. Since the value of state 2 is zero, $v_p(1)$ must satisfy $v_p(1) = p(1+\gamma v_p(1)) + (1-p)1$. Solving this we get

$$v_p(1) = \frac{1}{1-p\gamma}\,.$$

The derivative of this with respect to p is

$$\frac{d}{dp} v_p(1) = \frac{\gamma}{(1-\gamma p)^2}\,.$$

To get a δ-accurate estimate of $v_{p_0}(1)$, we need

$$\begin{aligned}\delta \geq |v_{p_0}(1) - v_{\bar X_m}(1)| \approx \frac{d}{dp} v_p(1)|_{p=p_0} |p_0 - \bar X_m| = \frac{\gamma}{(1-\gamma p_0)^2} |p_0 - \bar X_m| \\ \approx \frac{\gamma}{(1-\gamma p_0)^2} \sqrt{\frac{p_0(1-p_0)}{m}}\,.\end{aligned}$$

Inverting for m, we get that

$$m \gtrsim \frac{\gamma^2 p_0(1-p_0)}{(1-\gamma p_0)^4 \delta^2}\,.$$

It remains to choose p_0 as a function of γ to show that the above can be lower bounded by $1/(1-\gamma)^3$. If we choose $p_0 = \gamma$, we have $1-\gamma p_0 = 1-\gamma^2 = (1-\gamma)(1+\gamma) \le 2(1-\gamma)$ and hence

$$\frac{\gamma^2 p_0(1-p_0)}{(1-\gamma p_0)^4 \delta^2} \ge \frac{\gamma^2 \gamma(1-\gamma)}{2^4(1-\gamma)^4\delta^2} = \frac{\gamma^3}{2^4(1-\gamma)^3\delta^2}\,.$$

Putting things together finishes the proof sketch. ■

A homework problem is included which explains how to fill in the gaps in the last section of the proof, while pointers to the literature are given that one can use to figure out how to fill the remaining gaps.

Policy-based designs

When the data is generated by following some policy, we talk about policy based designs. Here, the design decision is what policy to use to generate the data. The **sample complexity of learning with policy based designs** is the number of observations necessary and sufficient for some algorithm to figure out a policy of a fixed target suboptimality, from a fixed initial state, based on data generated by following a policy where the MDP where the policy is followed can be any of the MDPs within the class.

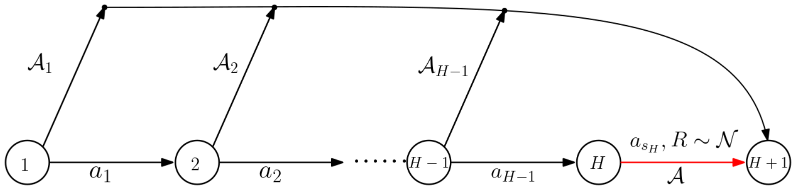
Three questions arise then. *(i)* The first question (the **design question**) is what policy to follow during data collection. If the policy can use the full history, the problem is not much different than online learning, which we will consider later. From this perspective the interesting (and perhaps more realistic) case is when **the data-collection policy is memoryless and is fixed before the data collection begins**. Hence, in what follows, we will restrict our attention to this case. *(ii)* The second question is what algorithm to use to compute the policy given the data generated. *(iii)* The final, third question is how large is the sample complexity of learning with policy induced data for some fixed MDP class.

Learning and estimating a good policy from policy induced data is much closer to reality than the same problem from Z-designs. Practical problems, such as problems in health care, robotics, etc., are so that we can obtain data generated by following some fixed policy, while it is usually not possible to demand obtaining sample transitions from **arbitrary** state-action pairs.

For simplicity, let us still consider the case of finite state-action MDPs, but to further simplify matters, let us now consider (homogenous kernel) **finite horizon problems** with a horizon H. As it turns out, the **plug-in algorithm** of the previous section is still a good algorithm in the sense that it achieves the optimal (minimax) sample complexity. However, the minimax sample complexity is much higher than it is for Z-designs:

Theorem (sample complexity lower bound with policy induced data): For any S, A, H, $0 \le \delta$, any (memoryless) data collection policy π over the state-action spaces $[S]$ and $[A]$, for any $n \le cA^{\min(S-1,H)}/\delta^2$ and any algorithm $\mathcal{L}$ that maps data that has n transitions to a policy there exist an MDP M with state space $[S]$ and action space $[A]$ such that with constant probability, the policy $\hat{\pi}$ produced by $\mathcal{L}$ is not δ-optimal with respect to the H-horizon total reward criterion when the algorithm is fed with data by following π in M.

Proof (sketch): Without loss of generality assume that $S = H + 1$; if there are more states, just ignore them, while if there are fewer states then just decrease H. Consider an MDP where states $1, \dots, H$ are organized in a chain under the effect of some actions, and state $H + 1$ is an absorbing state with zero associated reward. For $1 \le i \le H$, let action a_i be the one that gets to be chosen with the smallest probability in state i under the data generating policy π: $a_i = \arg\min_{a \in [A]} \pi(a|i)$. We choose action a_i as the action that moves the state from i to $i + 1$, deterministically. Any other action leads to state $H + 1$, deterministically. All rewards are zero, except when transitioning from state H to state $H + 1$ under action a_H, where the reward is stochastic with a normal distribution with mean μ either -2δ or $+2\delta$ and a variance of one. The structure of the MDP is shown on the figure in the left-hand side.



Now, because of the choice of a_i, $\pi(a_i|i) \le 1/A$. Hence, the probability that starting from state 1, following policy π for H steps will generate the sequence of states $1, 2, \dots, H, H + 1$, including the critical transition from state H to state $H + 1$, is at most $(1/A)^H$. This transition is critical in the sense is that only data from this transition decides whether in state 1 it is worth taking action a_1 or not. In particular, if $\mu = -2\delta$, taking a_1 is a poor choice, while if $\mu = 2\delta$, taking a_1 is the optimal choice. The expected number of times this critical transition is seen is at most $m = n(1/A)^H$. With m observations, the value of μ will be estimated up to an accuracy of $O(\sqrt{1/m})$. When this is smaller than 2δ, with constant probability, the sign of μ cannot be decided and thus with constant probability, any algorithm will fail to identify whether a_1 should be taken in state 1 or not (with a probability, of say, at least $1/2$). Plugging in the expected value of m, we get that the condition on n is that $\sqrt{cA^H/n} \le 2\delta$ where $c > 0$ is some universal constant. Equivalently, the condition is that $n \ge cA^H/(4\delta^2)$, which is the statement to be proven. ■

The lower bound construction suggests that the best policy to be used in the lack of extra information about the MDPs is the uniform policy. Note that a similar statement holds for the discounted setting. The contrast between this lower bound and the polynomial upper bound of the previous section are in strike contrast: Data obtained from following policies can be very poor. One may wonder whether the situation can be improved assuming that the data is obtained from a good policy (say, 2δ optimal policy), but the proof of the previous result in fact shows that this is not the case.

While the exponential lower bound on the sample complexity of learning from policy induced data is already bad enough, one may worry that the situation could be even worse. Could it happen that even the best algorithm needs double exponential number of samples? Or even infinite? A moment of thought shows that the latter is the case is switch to the average reward setting: This is because in the average reward setting the value of an action can depend on the value of a state whose hitting probability within an arbitrary fixed number of transitions is positive, just arbitrarily low. Can something similar happen perhaps in the finite-horizon setting, or the discounted setting? As it turns out, the answer is no. The previous lower bound gives the correct order of the sample complexity of finding a near-optimal policy using policy induced data:

Theorem (sample complexity upper bound with policy induced data): With $m = \Omega(S^3 H^4 A^{\min(H,S-1)+2}/\delta^2)$ episodes of length H collected with the uniform policy from a fixed initial distribution μ, with a constant probability, the plug-in algorithm produces a policy that is δ-optimal when started from μ.

Proof (sketch): For simplicity assume that the reward function is known. Let $\pi_{\log}$ be the logging policy, which is uniform. Again, assume that the plug-in algorithm produces a deterministic policy.

The proof is based on the decomposition of the suboptimality gap of the policy $\hat\pi$ produced that was used before. In particular, by (1),

$$v^*(\mu) - v^{\hat\pi}(\mu) \le v^{\pi^*}(\mu) - \hat v^{\pi^*}(\mu) \,+\, \hat v^{\hat\pi}(\mu) - v^{\hat\pi}(\mu)\,, \tag{7}$$

where as before, $\hat v^\pi$ denotes the value function of policy π in the empirically estimated MDP. Further, we also used $v(\mu)$ as a shorthand for $\sum_s \mu(s)v(s)(=\langle \mu, v\rangle)$, where $v:[S]\to\mathbb{R}$,

One then needs a counterpart of the value difference lemma. In this case, the following version is convenient: For any policy π,

$$q^\pi_H - \hat q^\pi_H = \sum_{h=0}^{H-1} (P_\pi)^h (P-\hat P)\hat v^\pi_{H-h-1}\,,$$

and

$$\hat q^\pi_H - q^\pi_H = \sum_{h=0}^{H-1} (\hat P_\pi)^h (\hat P-P) v^\pi_{H-h-1}\,,$$

where P_π and $\hat P_\pi$ are $SA\times SA$ matrices. These can be proved by using the Bellman equation for the action-value functions and a simple recursion and noting that $q_0 = r = \hat q_0$.

Next, we can observe that $v^\pi(\mu) = \langle \mu^\pi, q^\pi\rangle$ where μ^π is a distribution over $[S]\times[A]$ which assigns probability $\mu(s)\pi(a|s)$ to $(s,a)\in[S]\times[A]$. This, combined with the value difference identity makes $\nu^\pi_h := \mu^\pi(P_\pi)^h$ appear in the bounds. This is the probability distribution over the state-action space after using π for h steps when the initial distribution is μ^π. Now, as this is multiplied by $P-\hat P$, and for a given state-action pair (s,a), $\|P(s,a)-\hat P(s,a)\|_1 \lesssim 1/\sqrt{N(s,a)} \le 1/\sqrt{N_h(s,a)} \approx 1/\sqrt{m\nu_h^{\pi_{\log}}}$, using that $\nu^\pi_h(s,a) \le \sqrt{\nu^\pi_h(s,a)}$ which holds because $0\le \nu^\pi_h \le 1$, we see that it suffices if the ratios $\rho^\pi_h(s,a) := \nu^\pi_h(s,a)/\nu_h^{\pi_{\log}}(s,a)$ (or their square root) are controlled. Above, $N(s,a)$ is the number of times (s,a) is seen in the data, and $N_h(s,a)$ is the number of times (s,a) is seen in the data in the hth transition. Here we should also mention that we only control these terms for state-action pairs (s,a) that satisfy $\nu^\pi_h(s,a)\gtrsim 1/m$ as the total contribution of the other state-action pairs is $O(1/m)$, i.e., small. For these state action pairs, $\nu_h^{\pi_{\log}}(s,a)$ is also positive and with high probability, the counts are also positive.

Next, one can show that

$$\rho_h^\pi(s,a) \le A^{\min(h+1,S)}\,.$$

This is done in two steps. First, show that $\nu_h^\pi(s,a) \le A^{h+1}\nu_h^{\pi_{\log}}(s,a)$. This follows from the law of total probability: Write $\nu_h^\pi(s,a)$ as the sum of probabilities of all trajectories that end with (s,a) after h transitions. Next, for a given trajectory, replace each occurrence of π with $\pi_{\log}$ at the expense of introducing a factor of A^{h+1} (this comes from $\pi(a'|s') \le 1 \le A\pi_{\log}(a'|s')$). The next step is to show that $\nu_h^\pi(s,a) \le A^S\nu_h^{\pi_{\log}}(s,a)$ also holds. This inequality follows by observing that the uniform policy and the uniform mixture of all deterministic (memoryless) policies induce the same distribution over the trajectories. Then by letting DET denote the set of all deterministic policies, using that $\pi \in \mathrm{DET}$, we have $\nu_h^\pi(s,a) \le \sum_{\pi' \in \mathrm{DET}} \nu_h^{\pi'}(s,a) = A^S\nu_h^{\pi_{\log}}(s,a)$, where we used that $|\mathrm{DET}| = A^S$.

Putting things together, applying a union bound when it comes to argue for $\hat\pi$ and collecting terms gives the result. ■

Bibliographic remarks

Finding a good policy from a sample drawn from a Z-design and finding a good policy from a sample given a **generative model**, or **random access simulator** of the MDP (which we extensively studied in previous lectures on planning) are almost the same. The random access model however allows the learner to determine which state-action pair the next transition data should be generated at in reaction to the sample collected in a sequential fashion. Thus, computing a good policy with a random access simulator gives more power to the "learner" (or planner). The lower bound presented for Z-design can in fact be shown to hold for the generative setting, as well (the proof in the paper cited below goes through in this case with no changes). This shows that in the tabular case, adaptive random access to the simulator provides no benefits to the planner over non-adaptive random access.

The result of the $O(H^3SA/\delta^2)$ sample complexity bound to find a δ-optimal policy with uniform Z-design using the plug-in method is from the following paper:

- Agarwal, Alekh, Sham Kakade, and Lin F. Yang. 2020. "Model-Based Reinforcement Learning with a Generative Model Is Minimax Optimal." COLT, 67–83. arXiv link

This paper also contains a number of pointers to the literature. Interestingly, earlier approaches often used more complicated approaches which directly worked with value functions rather than the more natural plug-in approach. The problem of whether the plug-in method is minimax optimal in Z design for finite-horizon problem is open.

The result which was included in this lecture limits the range of δ to $\sqrt{H}$. Equivalently, the result is not applicable for a small number of observations m per state-action pair. This limitation has been removed in a follow-up to this work:

- Li, Gen, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. 2020. "Breaking the Sample Size Barrier in Model-Based Reinforcement Learning with a Generative Model." NeurIPS

This paper still uses the plug-in method, but adds random noise to the observed rewards to help with tie-breaking.

The variance bound, which is the key to achieving the cubic dependence on the horizon is from the following paper:

- Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. Machine learning, 91(3):325–349, 2013.

This paper also has the essential ideas for the matching lower bound. The 2020 paper is notable for some novel proof techniques, which were developed to bound the error terms whose control is not included in this lecture.

The results for learning with policy-induced data are from

- Xiao, Chenjun, Ilbin Lee, Bo Dai, Dale Schuurmans, and Csaba Szepesvari. 2021. "On the Sample Complexity of Batch Reinforcement Learning with Policy-Induced Data." arXiv

which also has the details that were omitted in these notes. This paper also gives a modern proof for the Z-design sample complexity lower bound.

One may ask whether the results for Z-design that show cubic dependence on the horizon H extend to the case of large MDPs when value function approximation is used. In a special case, this has been positively resolved in the following paper:

- Yang, Lin F., and Mengdi Wang. 2019. "Sample-Optimal Parametric Q-Learning Using Linearly Additive Features." ICML arXiv version

which uses an approach similar to Politex in a more restricted setting, but achieves an optimal dependence on H.

19. Scaling with value function approximation

Under construction.