Notes on Tabular Methods

Nan Jiang

September 27, 2022

1 Overview of the methods

1.1 Tabular certainty-equivalence

Certainty-equivalence is a **model-based** RL algorithm, that is, it first estimates an MDP model from data, and then performs policy evaluation or optimization in the estimated model as if it were true. To specify the algorithm it suffices to specify the model estimation step.

Given a dataset D of trajectories, $D = \{(s_1, a_1, r_1, s_2, \dots, s_{H+1})\}$, we first convert it into a bag of $\{(s, a, r, s')\}$ tuples, where each trajectory is broken into H tuples: (s_1, a_1, r_1, s_2), (s_2, a_2, r_2, s_3), …, (s_H, a_H, r_H, s_{H+1}). For every $s \in \mathcal{S}, a \in \mathcal{A}$, define $D_{s,a}$ as the subset of tuples where the first element of the tuple is s and the second is a, and we write $(r, s') \in D_{s,a}$ since all tuples in $D_{s,a}$ share the same state-action pair. The tabular certainty-equivalence model uses the following estimation of the transition function $\widehat{P}$: let $\mathbf{e}_{s'}$ be the unit vector whose s'-th entry is 1 and all other entries are 0,

$$\widehat{P}(s, a) = \frac{1}{|D_{s,a}|} \sum_{(r,s') \in D_{s,a}} \mathbf{e}_{s'}. \tag{1}$$

Here $\mathbb{I}(\cdot)$ is the indicator function. In words, $\widehat{P}(s'|s, a)$ is simply the empirical frequency of observing s' after taking a in state s. Similarly when reward function also needs to be learned, the estimate is

$$\widehat{R}(s, a) = \frac{1}{|D_{s,a}|} \sum_{(r,s') \in D_{s,a}} r. \tag{2}$$

$\widehat{P}$ and $\widehat{R}$ are the maximum likelihood estimates of the transition and the reward functions, respectively. Note that for the transition function to be well-defined we need $n(s, a) > 0$ for every $s, a \in \mathcal{S}$.

1.2 Value-based tabular methods

Certainty-equivalence explicitly stores an estimated MDP model, which has $O(|\mathcal{S}|^2|\mathcal{A}|)$ space complexity, and the algorithm has a *batch* nature, i.e., it is invoked after all the data are collected. In contrast, there is another popular family of RL algorithms that (1) only model the Q-value functions hence has $O(|\mathcal{S}||\mathcal{A}|)$ sample complexity, (2) can be applied in an online manner, i.e., the algorithm runs as more and more data are collected. Well-known examples include Q-learning [1] and Sarsa [2].

Another very appealing property of these methods is that it is relatively easy to incorporate sophisticated generalization schemes, such as deep neural networks, which has recently led to many

empirical successes [3]. On the other hand, such methods are typically less sample-efficient than model-based methods and will not be discussed in more details in this note.[1]

2 Analysis of certainty-equivalence RL

Here we analyze the method introduced in Section 1.1. For simplicity we further assume that data are generated by sampling each (s, a) a fixed number of times. We are interested in deriving high-probability guarantees for the optimal policy of $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{R}, \gamma)$ as a function of $n \equiv |D_{s,a}|$.

We provide three different analyses for the algorithm, and we should see some interesting trade-off between state space and horizon.

2.1 Naive analysis

The basic idea is, when n is sufficiently large, we expect $\widehat{R} \approx R$ and $\widehat{P} \approx P$. In particular, by Hoeffding's inequality and union bound, the following inequalities hold with probability at least $1 - \delta$:

$$\max_{s,a} |\widehat{R}(s, a) - R(s, a)| \le R_{\max} \sqrt{\frac{1}{2n} \ln \frac{4|\mathcal{S} \times \mathcal{A}|}{\delta}} \tag{3}$$

and

$$\max_{s,a,s'} |\widehat{P}(s'|s, a) - P(s'|s, a)| \le \sqrt{\frac{1}{2n} \ln \frac{4|\mathcal{S} \times \mathcal{A} \times \mathcal{S}|}{\delta}}. \tag{4}$$

Note that we first split the failure probability δ evenly between the reward estimation events and the transition estimation events. Then for reward, we split $\delta/2$ evenly among all (s, a); for transition, we split $\delta/2$ evenly among all (s, a, s'). From Eq.(4) we further have

$$\max_{s,a} \|\widehat{P}(s, a) - P(s, a)\|_1 \le \max_{s,a} |\mathcal{S}| \cdot \|\widehat{P}(s, a) - P(s, a)\|_\infty \le |\mathcal{S}| \cdot \sqrt{\frac{1}{2n} \ln \frac{4|\mathcal{S} \times \mathcal{A} \times \mathcal{S}|}{\delta}}. \tag{5}$$

To bound the suboptimality of $\pi^\star_{\widehat{M}}$, we first introduce the *simulation lemma* [6].

Lemma 1 (Simulation Lemma). *If* $\max_{s,a} |\widehat{R}(s, a) - R(s, a)| \le \epsilon_R$ *and* $\max_{s,a} \|\widehat{P}(s, a) - P(s, a)\|_1 \le \epsilon_P$*, then for any policy* $\pi : \mathcal{S} \to \mathcal{A}$*, we have*

$$\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \le \frac{\epsilon_R}{1 - \gamma} + \frac{\gamma\, \epsilon_P\, V_{\max}}{2(1 - \gamma)},$$

where $V_{\max} := R_{\max}/(1 - \gamma)$.

[1] Techniques such as experience replay can be used the improve the sample efficiency of many online algorithms [4], but it also blurs the boundary between value-based and model-based methods [5].

Proof. For any $s \in \mathcal{S}$,

$$
\begin{aligned}
&|V^\pi_{\widehat{M}}(s) - V^\pi_M(s)| \\
&= |\widehat{R}(s,\pi) + \gamma\langle \widehat{P}(s,\pi), V^\pi_{\widehat{M}}\rangle - R(s,\pi) - \gamma\langle P(s,\pi), V^\pi_M\rangle| \\
&\le \epsilon_R + \gamma|\langle \widehat{P}(s,\pi), V^\pi_{\widehat{M}}\rangle - \langle P(s,\pi), V^\pi_{\widehat{M}}\rangle + \langle P(s,\pi), V^\pi_{\widehat{M}}\rangle - \langle P(s,\pi), V^\pi_M\rangle\rangle| \\
&\le \epsilon_R + \gamma|\langle \widehat{P}(s,\pi) - P(s,\pi), V^\pi_{\widehat{M}}\rangle| + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \\
&= \epsilon_R + \gamma|\langle \widehat{P}(s,\pi) - P(s,\pi), V^\pi_{\widehat{M}} - \tfrac{V_{\max}}{2}\cdot\mathbf{1}\rangle| + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \\
&\le \epsilon_R + \gamma\|\widehat{P}(s,\pi) - P(s,\pi)\|_1\|V^\pi_{\widehat{M}} - \tfrac{V_{\max}}{2}\|_\infty + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \\
&\le \epsilon_R + \tfrac{\gamma\epsilon_P V_{\max}}{2} + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty.
\end{aligned}
$$

Since this holds for all $s \in \mathcal{S}$, we can also take infinite-norm on the LHS, which yields the desired result. Note that we subtract $\frac{V_{\max}}{2}\cdot\mathbf{1}$ ($\mathbf{1}$ is the all-one vector) to center the range of $V^\pi_{\widehat{M}}$ around the origin, which exploits the fact that both $\widehat{P}(s,\pi)$ and $P(s,\pi)$ are valid probability distributions and sum up to 1. □

Alternative proof of Simulation Lemma Here we sketch an alternative and more "modern" proof to the simulation lemma. The proof relies on the following identity: $\forall f \in \mathbb{R}^{\mathcal{S}}$, $s_0 \in \mathcal{S}$, [2]

$$
f(s_0) - V^\pi_M(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi, a\sim\pi(\cdot|s), r\sim R(s,a), s'\sim P(\cdot|s,a)}[f(s) - r - \gamma f(s')]. \tag{6}
$$

To see why, first note that $V^\pi_M(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi,s_0}, a\sim\pi(\cdot|s), r\sim R(s,a)}[r]$, so the corresponding terms can be dropped on the two sides of the equation. For the remaining terms, the RHS is

$$
\begin{aligned}
&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi,s_0}, a\sim\pi(\cdot|s), s'\sim P(\cdot|s,a)}[f(s) - \gamma f(s')] \\
=&\sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{E}_{s\sim d^{\pi,s_0}_t, a\sim\pi(\cdot|s), s'\sim P(\cdot|s,a)}[f(s) - \gamma f(s')].
\end{aligned}
$$

Recall that d^{π,s_0}_t is the distribution of s_t under π starting from s_0. In this summation, the $\gamma f(s')$ term for t cancels out exactly with the $f(s)$ term for $t+1$, because both s and s' are distributed according to d^{π,s_0}_{t+1} and the difference in discount factor (γ^{t-1} vs. γ^t) accounts for the γ in $\gamma f(s')$. As a result, only the first term $f(s_0)$ is left, which is the same as the remaining term on the LHS. In fact, this term is effectively the Bellman flow equation for d^π, written in a form where f serves as a "test function" or discriminator to the Bellman flow equation.

To prove simulation lemma, we simply let $f = V^\pi_{\widehat{M}}$. On the RHS, we marginalize out r and s', and obtain

$$
\begin{aligned}
&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi, a\sim\pi(\cdot|s)}[V^\pi_{\widehat{M}}(s) - R(s,a) - \gamma\langle P(\cdot|s,a), V^\pi_{\widehat{M}}\rangle] \\
=&\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi, a\sim\pi(\cdot|s)}[\widehat{R}(s,a) + \gamma\langle \widehat{P}(\cdot|s,a), V^\pi_{\widehat{M}}\rangle - R(s,a) - \gamma\langle P(\cdot|s,a), V^\pi_{\widehat{M}}\rangle].
\end{aligned}
$$

The rest of the proof follows similarly as the original proof.

[2] The RHS can also be conveniently written as $\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^\pi}[f - \mathcal{T}^\pi f]$.

Turning back to the analysis of certainty-equivalence, the following lemma translates the policy evaluation error to the suboptimality of $\pi^\star_{\widehat{M}}$:

Lemma 2 (Evaluation error to decision loss)**.** $\forall s \in \mathcal{S}$, $V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) \le 2 \sup_{\pi:\mathcal{S}\to\mathcal{A}} \|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty$.

Proof. For any $s \in \mathcal{S}$,

$$\begin{aligned} V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) &= V^{\pi^\star_M}_M(s) - V^{\pi^\star_M}_{\widehat{M}}(s) + V^{\pi^\star_M}_{\widehat{M}}(s) - V^{\pi^\star_{\widehat{M}}}_M(s) \\ &\le V^{\pi^\star_M}_M(s) - V^{\pi^\star_M}_{\widehat{M}}(s) + V^{\pi^\star_{\widehat{M}}}_{\widehat{M}}(s) - V^{\pi^\star_{\widehat{M}}}_M(s) && (\pi^\star_{\widehat{M}} \text{ maximizes } v_{\widehat{M}}) \\ &\le \|V^{\pi^\star_M}_M - V^{\pi^\star_M}_{\widehat{M}}\|_\infty + \|V^{\pi^\star_{\widehat{M}}}_{\widehat{M}} - V^{\pi^\star_{\widehat{M}}}_M\|_\infty. \end{aligned}$$ □

Putting Lemmas 1 and 2 together with the concentration inequalities, we can see that the suboptimality we incur is

$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) = \tilde{O}\left(\frac{|\mathcal{S}|V_{\max}}{\sqrt{n}(1-\gamma)}\right), \ \forall s \in \mathcal{S}.$$

Here $\tilde{O}(\cdot)$ supresses poly-logarithmic dependences on $|\mathcal{S}|$ and $|\mathcal{A}|$; in this note we also omit the dependence on $R_{\max}$ and $1/\delta$, and only highlight the dependence on $|\mathcal{S}|$, n, and $1/(1-\gamma)$.

2.2 Improving $|\mathcal{S}|$ to $\sqrt{|\mathcal{S}|}$

The previous analysis proves concentration for each individual $P(s'|s,a)$ and adds up the errors to give an ℓ_1 error bound, which is loose. We can obtain a tighter analysis by proving an ℓ_1 concentration bound for multinomial distribution directly.

Note that for any vector $v \in \mathbb{R}^{|\mathcal{S}|}$,

$$\|v\|_1 = \sup_{u\in\{-1,1\}^{|\mathcal{S}|}} u^\top v.$$

Each $u \in \{-1,1\}^{|\mathcal{S}|}$ projects the vector v to some scalar value. If v can be written as the sum of zero-mean i.i.d. vectors, we can prove concentration for $u^\top v$ first, and then union bound over all u to obtain the ℓ_1 error bound. Concretely, for any fixed (s,a) pair and any fixed $u \in \{-1,1\}^{|\mathcal{S}|}$, with probability at least $1 - \delta/(2|\mathcal{S}\times\mathcal{A}|\cdot 2^{|\mathcal{S}|})$, we have

$$u^\top(\widehat{P}(s,a) - P(s,a)) \le 2\sqrt{\frac{1}{2n}\ln\frac{2|\mathcal{S}\times\mathcal{A}|\cdot 2^{|\mathcal{S}|}}{\delta}}, \tag{7}$$

because $u^\top\widehat{P}(s,a)$ is the average of i.i.d. random variables $u^\top \mathbf{e}_{s'}$ with bounded range $[-1,1]$. [3] This leads to the following improvement over Eq.(5): w.p. at least $1-\delta/2$,

$$\max_{s,a}\|\widehat{P}(s,a) - P(s,a)\|_1 = \max_{s,a}\max_{u\in\{-1,1\}^{|\mathcal{S}|}} u^\top(\widehat{P}(s,a) - P(s,a)) \le 2\sqrt{\frac{1}{2n}\ln\frac{2|\mathcal{S}\times\mathcal{A}|\cdot 2^{|\mathcal{S}|}}{\delta}}. \tag{8}$$

Rougly speaking, the $\tilde{O}(|\mathcal{S}|\sqrt{\frac{1}{n}})$ bound in Eq.5 is improved to $\tilde{O}(\sqrt{\frac{|\mathcal{S}|}{n}})$, and propagating the improvement through the remainder of the analysis yields

$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) = \tilde{O}\left(\frac{\sqrt{|\mathcal{S}|}V_{\max}}{\sqrt{n}(1-\gamma)}\right), \ \forall s \in \mathcal{S}.$$

[3] Also note that we only bound the deviation from one side, so we save a factor of 2 in ln compared to bounding the absolute deviation. Another tiny improvement: for $u \in \{-1,1\}^{|\mathcal{S}|}$, one can ingore $u = \pm\mathbf{1}$ as $\pm\mathbf{1}(\widehat{P}(s,a) - P(s,a)) \equiv 0$.

2.3 No dependence on $|\mathcal{S}|$

The last analysis removes the dependence of n on $|\mathcal{S}|$, at the cost of an additional dependence on $\frac{1}{1-\gamma}$. Note that the total number of samples still scales with $|\mathcal{S}|$ as we require n samples per (s,a).

The core idea is to show $Q^\star_{\widehat{M}} \approx Q^\star_M$, and then upper bound loss by Lemma 4 from Note 1. First, by contraction we have,

$$\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty \le \frac{1}{1-\gamma}\|Q^\star_M - \mathcal{T}_{\widehat{M}} Q^\star_M\|_\infty. \tag{9}$$

This is because

$$\begin{aligned}\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty &= \|\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}} - \mathcal{T}_{\widehat{M}} Q^\star_M + \mathcal{T}_{\widehat{M}} Q^\star_M - Q^\star_M\|_\infty \\ &\le \gamma\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty + \|\mathcal{T}_{\widehat{M}} Q^\star_M - Q^\star_M\|_\infty. && (\mathcal{T}_{\widehat{M}} \text{ is a } \gamma\text{-contraction})\end{aligned}$$

Then we bound the RHS in the following lemma.

Lemma 3. *For any fixed $s \in \mathcal{S}, a \in \mathcal{A}$, with probability at least $1-\delta$,*

$$\left|Q^\star_M(s,a) - \left(\widehat{R}(s,a) + \gamma\langle \widehat{P}(s,a), V^\star_M\rangle\right)\right| \le \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

Proof. The bound follows directly from Hoeffding's inequality upon the following observation:

$$\widehat{R}(s,a) + \gamma\langle \widehat{P}(s,a), V^\star_M\rangle = \frac{1}{n}\sum_{(r,s')\in D_{s,a}} \left(r + \gamma V^\star_M(s')\right).$$

Note that the RHS is the average of i.i.d. random variables $(r + \gamma V^\star_M(s'))$ in the interval of $[0, \frac{R_{\max}}{1-\gamma}]$, whose expectation is exactly $Q^\star_M(s,a)$. Therefore, the LHS of the lemma statement is the deviation of average of i.i.d. variables from the expectation, where Hoeffding's inequality applies. □

Note that the LHS of the lemma statement is simply the (s,a)-th entry of $(Q^\star_M - \mathcal{T}_{\widehat{M}} Q^\star_M)$. The final result we can get is

$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) = \tilde{O}\left(\frac{V_{\max}}{\sqrt{n}(1-\gamma)^2}\right), \ \forall s \in \mathcal{S}.$$

The cubic dependence on horizon comes from 3 different sources: (1) the range of value, (2) translating Bellman error to the difference in optimal Q-value functions, and (3) error accumulation over time when taking actions greedily wrt $\widehat{Q}$. The previous analyses only paid quadratic dependence on horizon because (3) was not present.

Some notes on Eq.(9) One can also obtain the following inequality by swapping the roles of M and $\widehat{M}$ in Eq.(9):

$$\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty \le \frac{1}{1-\gamma}\|Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}\|_\infty.$$

In fact, the RHS of the above inequality is the more standard notion of *Bellman errors* (or *Bellman residuals*): it measures how much an *approximate* Q-value function (here $Q^\star_{\widehat{M}}$) deviates from itself when

updated using the *true* Bellman update operator. In fact we can attempt to complete the analysis based on this inequality instead of Eq.(9), by noticing that the RHS is (ignoring $\frac{1}{1-\gamma}$ and the max-norm)

$$\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}.$$

This way we also introduce $\mathcal{T}_{\widehat{M}}$ into the expression and compare it with $\mathcal{T}_M$, which should allow us to use concentration inequalities to bound the difference between $\mathcal{T}_{\widehat{M}}$ and $\mathcal{T}_M$.

Now the (s,a)-th entry of the above expression is

$$\left(\widehat{R}(s,a) + \gamma \langle \widehat{P}(s,a), V^\star_{\widehat{M}} \rangle\right) - \left(R(s,a) + \gamma \langle P(s,a), V^\star_{\widehat{M}} \rangle\right)$$

It is attempting to use the techniques in the proof of Lemma 3, by claiming that $(r + V^\star_{\widehat{M}}(s'))$ are i.i.d. random variables for $(r,s') \in D_{s,a}$, with expected value $R(s,a) + \gamma \langle P(s,a), V^\star_{\widehat{M}} \rangle$. This is not true in general, because the function $V^\star_{\widehat{M}}(s')$ itself is random and depends on the data in $D_{s,a}$! Hence Hoeffding does not apply. One workaround is to consider a deterministic function class that always contains $V^\star_{\widehat{M}}$ and do a union bound over that class; in fact, if we choose all tabular functions in the range of $[0, V_{\max}]$, the analysis is basically identical to Section 2.2.

Now you should see why we use $Q^\star_M$ and $\mathcal{T}_{\widehat{M}}$ in Eq.(9), as this way we compare $\mathcal{T}_M$ and $\mathcal{T}_{\widehat{M}}$ against $V^\star_M$, which is a *deterministic* function.

In cases where M's state space forms a directed acyclic graph (DAG), the argument with $V^\star_{\widehat{M}}$ can still work as $V^\star_{\widehat{M}}(s')$ only depends on the datasets for later state-action pairs, which do not include the current (s,a) under consideration. This argument is straightforward here because we have a very simple and clean data collection procedure. One has to be extremely careful when using this argument in more realistic settings: for example, in the exploration setting, even if $V^\star_{\widehat{M}}(s')$ is estimated from datasets not including $D_{s,a}$, the outcomes in $D_{s,a}$ might have determined which later states we have sufficient samples and which not, which introduces very subtle interdependence with $V^\star_{\widehat{M}}$.

Connection to MCTS Interestingly, the independence of n on $|\mathcal{S}|$ in the last analysis is the core idea that leads to Sparse Sampling [7], which is a prototype algorithm for the family of Monte-Carlo tree search algorithm that played a crucial role in the success of AlphaGo.

One way to view Sparse Sampling is the following: conceptually we run the tabular method with n set according to the last analysis (no dependence on $|\mathcal{S}|$). Of course, when $|\mathcal{S}|$ is large this is impractical, but if we only need to know $\pi^\star(s_0)$ for some particular state s_0 (which is the setting of online planning with MCTS), we can perform "lazy evaluation": only generate the datasets for state-action pairs that contribute to the calculation of $V^\star_{\widehat{M}}(s_0)$ and truncate at the effective horizon. Roughly speaking, this requires a total of $(n|\mathcal{A}|)^{O(\frac{1}{1-\gamma})}$ samples to compute $\pi^\star(s_0)$, where has no dependence on $|\mathcal{S}|$.

2.4 "Best of both worlds" in the large-sample regime

We have seen two different analyses, where one pays an extra factor in $|\mathcal{S}|$ and the other pays an extra factor in horizon (i.e., $1/(1-\gamma)$). Can we get the best of both worlds and pay neither factors?

The answer turns out to be positive, as long as we allow a slow "burn-in" term. To start, recall that in the comments below Eq.(9), we show that $\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty$ can also be bounded if we swap the

role between M and $\widehat{M}$ on the RHS, i.e., the RHS becomes $\frac{1}{1-\gamma}\|Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}\|_\infty$. The trouble is that we can no longer directly apply Hoeffding's inequality due to the data-dependence on $Q^\star_{\widehat{M}}$.

However, if we manage to control this term, it will save us a horizon factor (which is a good reason to consider this idea more carefully)! This is shown in the following lemma: (the lemma only involves the true MDP M, which is omitted in the subscripts of value functions and occupancies)

Lemma 4 ([8]). *For any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, any initial state s_0, and any policy π,*

$$V^\pi(s_0) - V^{\pi_f}(s_0) \le \frac{1}{1-\gamma}(\mathbb{E}_{d^\pi,s_0}[\mathcal{T}f - f] + \mathbb{E}_{d^{\pi_f},s_0}[f - \mathcal{T}f]),$$

where terms in the form of $\mathbb{E}_\mu[f]$ are the shorthand for $\mathbb{E}_{(s,a)\sim\mu}[f(s,a)]$.

Proof. We use the following identity, which is the Q-function variant of Eq.(6) (what we used to give the alternative proof of simulation lemma) and can be proved in very similar ways: for any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$,

$$f(s_0,\pi) - V^\pi(s_0) = \frac{1}{1-\gamma}\mathbb{E}_{d^\pi,s_0}[f - \mathcal{T}^\pi f]. \tag{10}$$

Then,

$$V^\pi(s_0) - V^{\pi_f}(s_0) \le V^\pi(s_0) - f(s_0,\pi) + f(s_0,\pi_f) - V^{\pi_f}(s_0).$$

For the two pairs of differences on the RHS, we invoke Eq.(10) twice: one with π, and the other with π rebound to π_f. This gives:

$$\begin{aligned}&V^\pi(s_0) - f(s_0,\pi) + f(s_0,\pi_f) - V^{\pi_f}(s_0)\\ =\ &\frac{1}{1-\gamma}\mathbb{E}_{d^\pi,s_0}[\mathcal{T}^\pi f - f] + \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_f},s_0}[f - \mathcal{T}^{\pi_f} f].\end{aligned}$$

The proof is completed by recognizing that $\mathcal{T}^\pi f \le \mathcal{T}f$, and $\mathcal{T}^{\pi_f} f = \mathcal{T}f$. □

Using Lemma 4, we immediately have that

$$\|V^\star_M - V^{\pi^\star_{\widehat{M}}}_M\|_\infty \le \frac{2}{1-\gamma}\|Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}\|_\infty,$$

which contrasts the $2/(1-\gamma)^2$ factor in Section 2.3. However, this brings back the earlier problem: how to handle the data dependence of $Q^\star_{\widehat{M}}$ in concentration?

Bounding $\|Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}\|_\infty$ We now show how to bound this term. $\forall(s,a)$,

$$\begin{aligned}&|Q^\star_{\widehat{M}}(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)| = |(\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}})(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)|\\ =\ &|(\mathcal{T}_{\widehat{M}})Q^\star_M(s,a) - (\mathcal{T}_M Q^\star_M)(s,a) + (\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}})(s,a) - (\mathcal{T}_{\widehat{M}} Q^\star_M)(s,a) + (\mathcal{T}_M Q^\star_M)(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)|.\end{aligned}$$

So the idea is to replace $Q^\star_{\widehat{M}}$ with $Q^\star_M$, then the difference is the same as in Section 2.3 which we know how to handle (without paying extra $|\mathcal{S}|$). The consequence of doing so is that we will get some extra terms (the last 4 terms above).

To handle those 4 terms, we plug in the definition of $\mathcal{T}_M$ and $\mathcal{T}_{\widehat{M}}$:

$$
\begin{aligned}
&|(\mathcal{T}_{\widehat{M}}Q^\star_{\widehat{M}})(s,a) - (\mathcal{T}_{\widehat{M}}Q^\star_M)(s,a) + (\mathcal{T}_M Q^\star_M)(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)| \\
&= (\widehat{R}(s,a) + \gamma\langle \widehat{P}(s,a), V^\star_{\widehat{M}}\rangle - \widehat{R}(s,a) - \gamma\langle \widehat{P}(s,a), V^\star_M\rangle \\
&\quad + R(s,a) + \gamma\langle P(s,a), V^\star_M\rangle - R(s,a) - \gamma\langle P(s,a), V^\star_{\widehat{M}}\rangle| \\
&= \gamma|\langle \widehat{P}(s,a), V^\star_{\widehat{M}} - V^\star_M\rangle - \langle P(s,a), V^\star_{\widehat{M}} - V^\star_M\rangle| \\
&= \gamma|\langle \widehat{P}(s,a) - \widehat{P}(s,a), V^\star_{\widehat{M}} - V^\star_M\rangle| \\
&\le \gamma\|P(s,a) - \widehat{P}(s,a)\|_1 \|V^\star_{\widehat{M}} - V^\star_M\|_\infty.
\end{aligned}
$$

Now we can control $\|P(s,a) - \widehat{P}(s,a)\|_1$ using the total-variation concentration bound in Section 2.2 (while paying the extra $|\mathcal{S}|$). We can separately control $\|V^\star_{\widehat{M}} - V^\star_M\|_\infty$ using the analysis in Section 2.3. Each term will scale as $O(1/\sqrt{n})$ (we are only considering the scaling with n here and ignoring the other variables such as $|\mathcal{S}|$ and $1/(1-\gamma)$), so **their product scales as $\boldsymbol{O(1/n)}$**. When n is sufficiently large, this term will be dominated by the $1/\sqrt{n}$ error coming out from $|(\mathcal{T}_{\widehat{M}})Q^\star_M(s,a) - (\mathcal{T}_M Q^\star_M)(s,a)|$ and can be omitted. (This is why it is sometimes called a "burn-in" term, as it only has significant effects in the small sample-size regime.) Note that this $O(1/n)$ term will have worse dependencies on $|\mathcal{S}|$ and $1/(1-\gamma)$ compared to the $O(1/\sqrt{n})$ term, so "sufficiently large n" means that n has to be larger than some function of $|\mathcal{S}|$ and $1/(1-\gamma)$ for the difference between n and $\sqrt{n}$ to compensate for the worse factors in $|\mathcal{S}|$ and $1/(1-\gamma)$. In such a large-sample regime, we obtain the nice error bound of $\tilde{O}(\frac{V_{\max}}{\sqrt{n}(1-\gamma)})$, i.e., there is neither the extra $|\mathcal{S}|$ factor as in Section 2.2 nor the extra $1/(1-\gamma)$ factor as in Section 2.3.

Further improvement The bound can be further improved by replacing the Hoeffding's inequalities with the Bernstein's, which provides sharper concentration bounds when the variance of the random variables are substantially smaller compared to their ranges (squared). In our setting, the range of random variables in the concentration of $\langle \hat{P}(s,a) - P(s,a), V\rangle$ is $V_{\max}$, so the worst-case variance is $O(V^2_{\max})$. However, it turns out that for certain V (e.g., $V = V^\pi_M$), such variance cannot be large for all (s,a) simultaneously as it adds up to $O(V^2_{\max})$ along the occupancy of π, and leveraging such a property leads to improved sample complexities; see [9] and [10, Section 2.3].

References

[1] Christopher John Cornish Hellaby Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge England, 1989.

[2] Satinder P Singh and Richard S Sutton. Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1-3):123–158, 1996.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[4] Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3-4):293–321, 1992.

[5] Harm van Seijen and Rich Sutton. A deeper look at planning as learning from replay. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2314–2322, 2015.

[6] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[7] Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2-3):193–208, 2002.

[8] Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

[9] 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.

[10] Alekh Agarwal, Nan Jiang, Sham Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. `https://rltheorybook.github.io/`.