# CS 542 Stats RL: Homework 2

October 1, 2024

**Submission deadline: Oct 16 (Wednesday) before class.**

**1.** We used (the V-function variant of) the following result when proving the simulation lemma: for any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and any $\pi : \mathcal{S} \to \Delta(\mathcal{A})$,

$$\mathbb{E}_{s\sim d_0}[f(s,\pi)] - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{d^\pi}[f - \mathcal{T}^\pi f]. \tag{1}$$

Note that if $f = Q^\pi$, then $\mathbb{E}_{s\sim d_0}[f(s,\pi)] = \mathbb{E}_{s\sim d_0}[Q(s,\pi)] = J(\pi)$, so an interpretation is that if we treat $f$ as an approximation to $Q^\pi$ and use it to estimate $J(\pi)$, the error can be written as the Bellman error of $f$ — that is, how much it violates the Bellman equation satisfied by $Q^\pi$ — on $d^\pi$.

Now we know that we can also obtain $J(\pi)$ via $J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)]$. One can now ask an analogous question: if we use an arbitrary distribution $d \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as an approximation of $d^\pi$ to form an estimate of $J(\pi)$ as $\frac{1}{1-\gamma}\mathbb{E}_d[R]$, can we also express the error as the violation of $d$ w.r.t. the Bellman *flow* equation satisfied by $d^\pi$? The answer is yes, which is the following identity you are asked to prove: for any $d \in \Delta(\mathcal{S}\times\mathcal{A})$,[1]

$$\frac{1}{1-\gamma}\mathbb{E}_d[R] - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d, s'\sim P(\cdot|s,a)}[Q^\pi(s,a) - \gamma Q^\pi(s',\pi)] - \mathbb{E}_{s\sim d_0}[Q^\pi(s,\pi)]. \tag{2}$$

The RHS can be viewed as the Bellman flow error $d - \gamma(P^\pi)^\top d - d_0 \times \pi$[2] "tested" on $Q^\pi$ as a discriminator.

[1] The bound still holds when $d$ is not a valid distribution, and we just need to change $\mathbb{E}_d[\cdot]$ to be the dot product between $d$ and the function inside.

[2] Here $P^\pi$ is the state-action transition matrix: $P^\pi(s',a'|s,a) = P(s'|s,a) \times \pi(a'|s')$, and $d_0 \times \pi$ is the joint distribution $s \sim d_0, a \sim \pi(\cdot|s)$.

**2.** Recall that in value iteration we have $f_k = \mathcal{T} f_{k-1}$ for $k = 1, 2, \ldots, K$, with an arbitrary initialization of $f_0$. For simplicity let's take $f_0 \equiv 0$.

Now imagine that we are running some approximate version of value iteration where $f_k \approx \mathcal{T} f_{k-1}$ (i.e., we expect $f_k - \mathcal{T} f_{k-1}$ to be small for all $k$),[3] and output a non-stationary policy $\widehat{\pi}$: $a_1 \sim \pi_{f_K}$, $a_2 \sim \pi_{f_{K-1}}$, ..., $a_{K+1} \sim \pi_{f_0}$, and $a_{K+2:\infty}$ are decided arbitrarily.[4] We will also write this as $\widehat{\pi} = \widehat{\pi}_{1:\infty}$ with $\widehat{\pi}_t = \pi_{f_{K-t+1}}$ for $t \le K+1$, i.e., $\widehat{\pi}_t$ refers to the $t$-th "slice" of $\widehat{\pi}$ which is a stationary policy that maps $\mathcal{S}$ to $\mathcal{A}$.

Given an initial distribution $d_0 \in \Delta(\mathcal{S})$, let $J(\pi) := \mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t | \pi, s_1 \sim d_0]$. (Note that this definition applies to non-stationary $\pi$.) **Show that** for any (possibly non-stationary) $\pi$,

$$J(\pi) - J(\widehat{\pi}) \le \sum_{t=1}^{K} \gamma^{t-1} \left( \mathbb{E}_{d_t^{\pi}}[\mathcal{T} f_{K-t} - f_{K-t+1}] + \mathbb{E}_{d_t^{\widehat{\pi}}}[f_{K-t+1} - \mathcal{T} f_{K-t}] \right) + \gamma^K V_{\max}. \tag{3}$$

Here $d_t^{\pi} \in \Delta(\mathcal{S} \times \mathcal{A})$ is the $t$-step state-action distribution induced by starting from $d_0$ and executing $\pi$, and is well-defined for non-stationary policies. The terms in the form of $\mathbb{E}_{\mu}[f]$ are the shorthand for $\mathbb{E}_{(s,a)\sim\mu}[f(s,a)]$.

**In addition, derive** the following as a **direct** corollary of Eq.(3): Now consider the scenario where we are given an arbitrary function $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and output a stationary policy $\pi_f$. Show that

$$J(\pi) - J(\pi_f) \le \frac{1}{1-\gamma} (\mathbb{E}_{d^{\pi}}[\mathcal{T} f - f] + \mathbb{E}_{d^{\pi_f}}[f - \mathcal{T} f]). \tag{4}$$

By "direct corollary" you are asked to invoke Eq.(3) with a specific choice of $K$ and $f_{1:K}$.

**Hint 1:** You may find the following lemma useful (you need to prove it before using it): for any $\pi = \pi_{1:\infty}$,

$$\mathbb{E}_{s\sim d_0}[f_K(s, \pi_1)] - J(\pi) = \left( \sum_{t=1}^{K} \gamma^{t-1} \mathbb{E}_{d_t^{\pi}}[f_{K-t+1} - \mathcal{T}^{\pi_{t+1}} f_{K-t}] \right) - \mathbb{E}[\textstyle\sum_{t=K+1}^{\infty} \gamma^{t-1} r_t | \pi, d_0]. \tag{5}$$

**Hint 2:** Eq.(4) is proved in note3, which you can use as a hint. It relies on Eq.(1) (which is analogous to Eq.(5)), and its V-function variant is what we used to prove the simulation lemma.

**Remark** In the class we showed that the Bellman error $\|f - \mathcal{T} f\|_\infty$ can control $\|f - Q^\star\|_\infty$ up to a factor of horizon $1/(1-\gamma)$, which then controls the suboptimality of $\pi_f$ with another $2/(1-\gamma)$ factor. Put together, we have

$$\|V^\star - V^{\pi_f}\|_\infty \le \frac{2\|f - \mathcal{T} f\|_\infty}{(1-\gamma)^2}.$$

It turns out this is loose by a factor of horizon $1/(1-\gamma)$, and Eq.(4) gives this improved result.

In addition, although we consider the infinite-horizon setting here, it is not difficult to see that the result easily extends to the finite-horizon setting.

[3] Note that this "$\approx$" is not a "hard" assumption but rather to provide intuition. In fact, the sequence of functions $f_{1:K}$ can be *anything*.

[4] Even when VI is exact, outputting such a non-stationary policy actually yields *better* guarantees (see note1). It is also easier to analyze in some scenarios.

**3.** In the class we went through two different analyses in the tabular case to provide guarantees on $\|V_M^\star - V_M^{\pi^\star_{\widehat{M}}}\|_\infty$: either by bounding $\max_\pi \|V_M^\pi - V_{\widehat{M}}^\pi\|_\infty$ (Sec 2.1 and 2.2 of note3) or by bounding $\|Q_M^\star - Q_{\widehat{M}}^\star\|_\infty$ (Sec 2.3). In the former, we bound the concentration of rewards and transitions separately; in the latter, we bound the concentration of empirical Bellman update $r + \gamma V_M^\star(s')$ as a whole.

Here, you are asked to still take the first route, but without separately controlling the concentration of rewards and transitions. Instead, control the concentration of $r + \gamma V_M^\pi(s')$ for all $\pi$. In the analysis we did in the class, what showed up through the simulation lemma is the average of $r + \gamma V_{\widehat{M}}^\pi(s')$ over the dataset, where Hoeffding's inequality is not applicable (why?). Think about how to replace $V_{\widehat{M}}^\pi$ with $V_M^\pi$ here.

Once you obtain the bound, compare it to the results in Sec 2.2 of note3. They should only differ in minor ways (i.e., logarithmic terms). If you are seeing a substantial improvement (especially a $\sqrt{S}$), your concentration analysis is likely missing something important.

1.

From RHS:

$$\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d,\, s'\sim P(\cdot|s,a)}\left[Q^\pi(s,a)-\gamma Q^\pi(s',\pi)\right]-\mathbb{E}_{s\sim d_0}\left[Q^\pi(s,\pi)\right]$$

$$=\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d,\, s'\sim P(\cdot|s,a)}\left[Q^\pi(s,a)+R(s\quad)-R(s\quad)-\gamma Q^\pi(s',\pi)\right]-\mathbb{E}_{s\sim d_0}\left[Q^\pi(s,\pi)\right]$$

$$=\frac{1}{1-\gamma}\mathbb{E}_{s\sim d}\left[R\right]-J(\pi)$$

$$\mathbb{E}_{s\sim d_0}\left[f(s,\pi)\right]-J(\pi)=\mathbb{E}_{s\sim d_0}\left[f(s,\pi)-Q^\pi(s,\pi)\right]$$

$$=\mathbb{E}_{s\sim d_0}\left[f(s,\pi)-\mathcal{T}^\pi f(s,\pi)+\mathcal{T}^\pi f(s,\pi)-Q^\pi(s,\pi)\right]$$

$$=\mathbb{E}_{d^\pi}\left[f-\mathcal{T}^\pi f\right]$$

2.

①

$$J(\pi) - J(\hat\pi) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, \pi, s_1 \sim d_0\right] - \mathbb{E}\left[\sum_{t=1}^{K} \gamma^{t-1} r_t \,\middle|\, \hat\pi_t, s_1 \sim d_0\right]$$

$$\leq \sum_{t=1}^{K} \gamma^{t-1} \left( \mathbb{E}_{d_t^\pi} R(s,a) - \mathbb{E}_{d_t^{\hat\pi}} R(s,a) \right) + \gamma^K V_{max}$$

$$\leq \sum_{t=1}^{K} \gamma^{t-1} \Big( \mathbb{E}_{d_t^\pi} R(s,a) + \mathbb{E}_{d_t^\pi}\left[\gamma \langle P(\cdot \mid s,a), V_{f_{K-t}} \rangle - f_{K-t+1}\right]$$ check $\leq$

$$- \mathbb{E}_{d_t^{\hat\pi}}\left[\gamma \langle P(\cdot \mid s,a), V_{f_{K-t}} \rangle - f_{K-t+1}\right] - \mathbb{E}_{d_t^{\hat\pi}} R(s,a) \Big) + \gamma^K V_{max}$$

$$= \sum_{t=1}^{K} \gamma^{t-1} \left( \mathbb{E}_{d_t^\pi} (\mathcal{T} f_{K-t} - f_{K-t+1}) + \mathbb{E}_{d_t^{\hat\pi}} (f_{K-t+1} - \mathcal{T} f_{K-t}) \right) + \gamma^K V_{max}$$

②

lemma:

Start from RHS,

$$\left( \sum_{t=1}^{K} \gamma^{t-1} \mathbb{E}_{d_t^\pi} \left[ f_{K-t+1} - \mathcal{T}^{\pi_{t+1}} f_{K-t} \right] \right) - \mathbb{E}\left[ \sum_{t=K+1}^{\infty} \gamma^{t-1} r_t \,\middle|\, \pi, d_0 \right]$$

$$= \left( \mathbb{E}_{d_1^{\pi_1}} \left[ f_K - R(s,a) - \gamma \cancel{f_{K-1}(s, \pi_2)} \right] + \gamma \mathbb{E}_{d_t^{\pi_2}} \left[ \cancel{f_{K-1}} - R(s,a) - \gamma \cancel{f_{K-2}(s, \pi_3)} \right] + \cdots \right) - (\quad)$$

$$= \mathbb{E}_{s \sim d_0} [f_K(s, \pi_1)] - \mathbb{E}\left[ \sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, \pi, s \sim d_0 \right] = \mathbb{E}_{s \sim d_0} [f_K(s, \pi_1)] - J(\pi)$$

Proof: (1) $\mathbb{E}_{s \sim d_0} [f(s, \pi)] - J(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{d^\pi} [f - \mathcal{T}^\pi f]$

$$J(\pi) - J(\pi_f) = V^\pi(s_0) - V^{\pi_f}(s_0) \leq V^\pi(s_0) - f(s_0, \pi) + f(s_0, \pi_f) - V^{\pi_f}(s_0)$$

by (1)

$$= \frac{1}{1-\gamma} \left( \mathbb{E}_{d^\pi} [\mathcal{T} f - f] + \mathbb{E}_{d^{\pi_f}} [f - \mathcal{T} f] \right)$$ note: $\mathcal{T}^\pi f \leq \mathcal{T} f$ and $\mathcal{T}^{\pi_f} f = \mathcal{T} f$

3,

$$\forall \pi,\ \| V_M^\pi - V_{\hat{M}}^\pi \|_\infty = \| V_M^\pi - \hat{R} - \gamma \hat{P}^T V_{\hat{M}}^\pi \|_\infty$$

$$= \| V_M^\pi - \hat{R} - \gamma \hat{P}^T V_M^\pi + \gamma \hat{P}^T V_M^\pi - \gamma \hat{P}^T V_{\hat{M}}^\pi \|_\infty$$

$$\leq \| V_M^\pi - \hat{R} - \gamma \hat{P}^T V_M^\pi \| + \gamma \| \hat{P}^T (V_M^\pi - V_{\hat{M}}^\pi) \|_\infty$$

$$\leq \| V_M^\pi - \hat{R} - \gamma \hat{P}^T V_M^\pi \| + \gamma \| V_M^\pi - V_{\hat{M}}^\pi \|_\infty$$

$$\Rightarrow \| V_M^\pi - V_{\hat{M}}^\pi \|_\infty \leq \frac{1}{1-\gamma} \underbrace{\Big\| V_M^\pi - \sum_{(r,s')\sim D} (r + \gamma V_M^\pi(s')) \Big\|_\infty}_{(i)}$$

Hoeffding's $\downarrow$

$$\forall s,\ \text{w.p. } 1-\delta\ \ (i) \leq \frac{1}{1-\gamma} V_{max} \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$$

$$\Downarrow$$

$$\text{w.p. } 1-\delta\ \ \forall s,\ \ (i) \leq \frac{1}{1-\gamma} V_{max} \sqrt{\frac{1}{2n} \ln \frac{2|S|}{\delta}}$$

Therefore, $\max_\pi \| V_M^\pi - V_{\hat{M}}^\pi \|_\infty \leq \frac{1}{1-\gamma} V_{max} \sqrt{\frac{1}{2n} \ln \frac{2|S|}{\delta}}$ w.p. $1-\delta$