COMS6998-11: Homework 2 Solutions

Akshay Krishnamurthy
akshay@cs.umass.edu

1. **Importance weighting and policy gradient.** For the first part, we have to combine to arguments we have seen previously: that "on policy" roll-outs with geometric stopping is unbiased, and that importance weighting is unbiased. The first part is shown by the following calculation (throughout we are conditioning on $s_0 = s, a_0 = a$):

$$\mathbb{E}\left[\left(\prod_{t=1}^{t_\star} \frac{\pi_2(a_t \mid s_t)}{\pi_1(a_t \mid s_t)}\right) \frac{r_{t^\star}}{1-\gamma}\right] = \sum_{T=0}^{\infty} \Pr[t^\star = T] \mathbb{E}\left[\left(\prod_{t=1}^{T} \frac{\pi_2(a_t \mid s_t)}{\pi_1(a_t \mid s_t)}\right) \frac{r_T}{1-\gamma}\right]$$
$$= \sum_{T=0}^{\infty} \mathbb{E}\left[\left(\prod_{t=1}^{T} \frac{\pi_2(a_t \mid s_t)}{\pi_1(a_t \mid s_t)}\right) \gamma^T r_T\right]$$

For the second part, let's consider just one of the terms above and expand the expectation over trajectories $\tau_t = (s_0, a_0, s_1, a_1, \ldots, s_T, a_T)$. Here we let $r(s,a)$ denote the expected reward from (s,a).

$$\mathbb{E}\left[\left(\prod_{t=1}^{T} \frac{\pi_2(a_t \mid s_t)}{\pi_1(a_t \mid s_t)}\right) \gamma^T r_T\right] = \sum_{\tau_T} \mathbb{P}^{\pi_1}[\tau_t] \left(\prod_{t=1}^{T} \frac{\pi_2(a_t \mid s_t)}{\pi_1(a_t \mid s_t)}\right) \gamma^T r(s_T, a_T)$$
$$= \sum_{\tau_T} \left(\prod_{t=1}^{T} P(s_t \mid s_{t-1}, a_{t-1}) \pi_1(a_t \mid s_t)\right) \left(\prod_{t=1}^{T} \frac{\pi_2(a_t \mid s_t)}{\pi_1(a_t \mid s_t)}\right) \gamma^T r(s_T, a_T)$$
$$= \sum_{\tau_T} \left(\prod_{t=1}^{T} P(s_t \mid s_{t-1}, a_{t-1}) \pi_2(a_t \mid s_t)\right) \gamma^T r(s_T, a_T)$$
$$= \sum_{\tau_T} \mathbb{P}^{\pi_2}[\tau_t] \gamma^T r(s_T, a_T).$$

Putting this together with the previously display, we obtain the result.

For the second part, the calculation is somewhat straightforward:

$$\frac{\pi_2(a \mid s)}{\pi_1(a \mid s)} = \frac{\exp(c(s,a))}{\sum_{a'} \pi_1(a' \mid s) \exp(c(s,a'))} \le \frac{\exp(c_\star)}{\exp(-c_\star) \sum_{a'} \pi_1(a' \mid s)} \le \exp(2c^\star).$$

A similar calculation applies for the other direction.

Finally for the third part, we focus on finding a deterministic quantity $t_{\max}$ such that $t_\star < t_{\max}$ with high probability. If this holds (formally, conditioned on $t_\star \le t_{\max}$), we know that $\hat{Q}^{\pi_2}(s,a) \le \frac{\exp(2t_{\max}c^\star)}{1-\gamma} =: Q_{\max}$ with probability 1, so we will have proved the result.

By a direct calculation

$$\Pr[t^\star \ge T] = \sum_{\tau=T}^{\infty} (1-\gamma)\gamma^\tau = \gamma^T (1-\gamma) \sum_{\tau=0}^{\infty} \gamma^\tau = \gamma^T$$

Therefore $t_{\max} \ge \log(1/\delta)/\log(1/\gamma)$ suffices.

2. **Tabular RL with generative models.** Consider some policy π and some (s,a) pair. (For notation only we consider π to be deterministic but this is not essential.) First note that since rewards are in $[0,1]$ we have that $Q^\pi_{M_2} \in [0, \frac{1}{1-\gamma}]$. Then

$$\begin{aligned}
|Q^\pi_{M_1}(s,a) - Q^\pi_{M_2}(s,a)| &= |\mathbb{E}_{R_1(s,a)}[r] + \gamma \mathbb{E}_{s'\sim P_1(s,a)} Q^\pi_{M_1}(s',\pi(s')) - \mathbb{E}_{R_2(s,a)}[r] - \gamma \mathbb{E}_{s'\sim P_2(s,a)} Q^\pi_{M_2}(s',\pi(s'))| \\
&\leq |\mathbb{E}_{R_1(s,a)}[r] - \mathbb{E}_{R_2(s,a)}[r]| + \gamma |\mathbb{E}_{s'\sim P_1(s,a)}[Q^\pi_{M_1}(s',\pi(s'))] - \mathbb{E}_{s'\sim P_2(s,a)}[Q^\pi_{M_2}(s',\pi(s'))]| \\
&\leq \varepsilon + \gamma \left|\mathbb{E}_{P_1(s,a)}[Q^\pi_{M_1}(s',\pi(s')) - Q^\pi_{M_2}(s',\pi(s'))]\right| + \gamma \left|(\mathbb{E}_{P_1(s,a)} - \mathbb{E}_{P_2(s,a)}) Q^\pi_{M_2}(s',\pi(s'))\right| \\
&\leq \varepsilon + \frac{\gamma\varepsilon}{1-\gamma} + \gamma \left|\mathbb{E}_{P_1(s,a)}[Q^\pi_{M_1}(s',\pi(s')) - Q^\pi_{M_2}(s',\pi(s'))]\right| \\
&\leq \varepsilon + \frac{\gamma\varepsilon}{1-\gamma} + \gamma \max_{s,a} \left|Q^\pi_{M_1}(s,a) - Q^\pi_{M_2}(s,a)\right|.
\end{aligned}$$

Here, the first equality uses the definitions of the Q functions. In the second we use the triangle inequality to separate the immediate reward from the next-step value functions. In the third line we use the assumed bound on the reward differences and we also add and subtract a "cross term" quantity: $\mathbb{E}_{s'\sim P_1(s,a)} Q^\pi_{M_2}(s',\pi(s'))$. This leads us to a one step error term as well as a recursive term. The one step term is bounded via $|(\mathbb{E}_P - \mathbb{E}_Q)(f(x))| \leq \sup_x |f(x)| \cdot \|P - Q\|_{\mathrm{TV}}$.

Now let $\bar{s}, \bar{a}$ be the state-action pair that maximize the difference $(\bar{s},\bar{a}) = \mathrm{argmax}_{s,a} |Q^\pi_{M_1}(s,a) - Q^\pi_{M_2}(s,a)|$. Then we have just showed that

$$|Q^\pi_{M_1}(\bar{s},\bar{a}) - Q^\pi_{M_2}(\bar{s},\bar{a})| \leq \varepsilon + \frac{\gamma\varepsilon}{1-\gamma} + \gamma \left|Q^\pi_{M_1}(\bar{s},\bar{a}) - Q^\pi_{M_2}(\bar{s},\bar{a})\right|.$$

We can re-arrange this to obtain a bound for all state-action pairs.

$$|Q^\pi_{M_1}(s,a) - Q^\pi_{M_2}(s,a)| \leq |Q^\pi_{M_1}(\bar{s},\bar{a}) - Q^\pi_{M_2}(\bar{s},\bar{a})| \leq \frac{\varepsilon}{1-\gamma} + \frac{\gamma\varepsilon}{(1-\gamma)^2} \leq \frac{2\varepsilon}{(1-\gamma)^2}.$$

For the second part, consider a single (s,a) pair and obtain n samples $\{(r_i, s'_i)\}_{i=1}^n$ from the sampling oracle. Then by Hoeffding's inequality the empirical reward $\bar{R}(s,a) = \frac{1}{n}\sum_{i=1}^n r_i$ satisfies (w.p. $1-\delta$)

$$\left|\bar{R}(s,a) - \mathbb{E}_{R(s,a)}[r]\right| \lesssim \sqrt{\frac{\log(1/\delta)}{n}}.$$

Meanwhile, by Bernstein's inequality the empirical transition probability $\hat{P}(s' \mid s,a) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{s'_i = s'\}$ satisfies

$$\left|\hat{P}(s' \mid s,a) - P(s' \mid s,a)\right| \lesssim \sqrt{\frac{P(s' \mid s,a)\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}.$$

Therefore, taking a union bound over all choices of s' we have

$$\begin{aligned}
\|\hat{P}(s,a) - P(s,a)\|_{\mathrm{TV}} &\leq \frac{1}{2}\sum_{s'} \left|\hat{P}(s' \mid s,a) - P(s' \mid s,a)\right| \\
&\lesssim \sum_{s'} \left(\sqrt{\frac{P(s' \mid s,a)\log(S/\delta)}{n}} + \frac{\log(S/\delta)}{n}\right) \\
&\lesssim \sqrt{\frac{S\log(S/\delta)}{n}} + \frac{S\log(S/\delta)}{n}
\end{aligned}$$

With a union bound the above argument holds simultaneously for all SA pairs. Thus if we set $n = O(S\log(SA/\delta)/\varepsilon^2)$ we have uniform approximation, using $O(S^2 A\log(SA/\delta)/\varepsilon^2)$ samples in total.

If we find the optimal policy in the approximate MDP $\hat{M}$, by an analysis similar to that for ERM, we have

$$\begin{aligned}
J_M(\pi^\star) &\leq J_{\hat{M}}(\pi^\star) + \frac{2\varepsilon}{(1-\gamma)^2} \leq J_{\hat{M}}(\hat{\pi}) + \frac{2\varepsilon}{(1-\gamma)^2} \\
&\leq J_M(\pi^\star) + \frac{4\varepsilon}{(1-\gamma)^2}
\end{aligned}$$

Thus to obtain suboptimality ϵ in total, we require

$$O\left(\frac{S^2 A \log(SA/\delta)}{(1-\gamma)^4 \epsilon^2}\right)$$

samples in total.

3. **Generative models for linear MDPs.** We use a recursive argument, similar to the proof of the simulation lemma:

$$\left|Q^{(T)}(s,a) - Q^\star(s,a)\right| \le \left|\widehat{\mathcal{T}V^{(T-1)}}(s,a) - \mathcal{T}V^{(T-1)}(s,a)\right| + \left|\mathcal{T}V^{(T-1)}(s,a) - Q^\star(s,a)\right|$$

Using the linear MDP property, there exists a $\bar{w}$ such that $\mathcal{T}V^{(T-1)}(s,a) = \langle \phi(s,a), \bar{w}\rangle$, while we constructed the empirical backup to satisfy $\widehat{\mathcal{T}V^{(T-1)}}(s,a) = \langle \phi(s,a), \hat{w}\rangle$. Introducing the covariance matrix, we have

$$\begin{aligned}\left|\widehat{\mathcal{T}V^{(T-1)}}(s,a) - \mathcal{T}V^{(T-1)}(s,a)\right| &= |\langle \phi(s,a), \bar{w} - \hat{w}\rangle| \le \|\phi(s,a)\|_{\Sigma^{-1}} \cdot \|\bar{w} - \hat{w}\|_\Sigma \\ &\le \sqrt{d} \cdot \sqrt{\mathbb{E}_D\left[(\widehat{\mathcal{T}V^{(T-1)}}(s,a) - \mathcal{T}V^{(T-1)}(s,a))^2\right]} \\ &\le \sqrt{\frac{d\Delta \log(1/\delta)}{n}}.\end{aligned}$$

This takes care of the first term. The second term has a recursive form

$$\left|\mathcal{T}V^{(T-1)}(s,a) - Q^\star(s,a)\right| = \gamma \left|\mathbb{E}_{s' \sim P(s,a)}\left[V^{(T-1)}(s') - V^\star(s')\right]\right|$$

Recall that $V^{(T-1)}(s') = \max_{a'} Q^{(T-1)}(s',a')$ while $V^\star(s') = \max_{a'} Q^\star(s',a')$. We would like to obtain a difference in the two Q-functions on the same state-action pair as this will allow us to recurse the argument. For this we introduce the policy $\tilde{\pi}(s) = \operatorname{argmax}_a \max\{Q^{(T-1)}(s,a), Q^\star(s,a)\}$. This policy leads to the inequality

$$\left|V^{(T-1)}(s') - V^\star(s')\right| = \left|\max_{a'} Q^{(T-1)}(s',a') - \max_{a'} Q^\star(s',a')\right| \le \left|Q^{(T-1)}(s',\tilde{\pi}(s')) - Q^\star(s',\tilde{\pi}(s'))\right|$$

To see why this is true, suppose that $\tilde{\pi}(s')$ is the greedy action w.r.t., $Q^{(T-1)}(s',\cdot)$. Then

$$Q^{(T-1)}(s',\tilde{\pi}(s')) \ge Q^\star(s',\pi^\star(s')) \ge Q^\star(s',\tilde{\pi}(s')),$$

as desired. A similar argument holds in the other case. Therefore,

$$\gamma \left|\mathbb{E}_{s' \sim P(s,a)}\left[V^{(T-1)}(s') - V^\star(s')\right]\right| \le \gamma \cdot \sup_{s,a} \left|Q^{(T-1)}(s,a) - Q^\star(s,a)\right|$$

Putting things together, we have

$$\sup_{s,a} |Q^{(T)}(s,a) - Q^\star(s,a)| \le \sqrt{\frac{d\Delta \log(1/\delta)}{n}} + \gamma \sup_{s,a} |Q^{(T-1)}(s,a) - Q^\star(s,a)|$$

We can apply the same argument on the last term on the right hand side and unrolling this gives

$$\sup_{s,a} |Q^{(T)}(s,a) - Q^\star(s,a)| \le \sum_{t=0}^{T} \gamma^t \sqrt{\frac{d\Delta \log(1/\delta)}{n}} + \frac{\gamma^T}{1-\gamma} \le \frac{1}{1-\gamma}\sqrt{\frac{d\Delta \log(1/\delta)}{n}} + \frac{\gamma^T}{1-\gamma}.$$

The final term uses the trivial bound that $|Q^{(1)} - Q^\star| \le \frac{1}{1-\gamma}$ simply because $Q^\star(s,a) \in [0, 1/(1-\gamma)]$.

For the second part, let the right hand side of the bound in part (a) be ε_Q. Then

$$\begin{aligned}J(\pi^\star) - J(\hat{\pi}) &= \mathbb{E}_{s_0} Q^\star(s_0, \pi^\star(s_0)) - Q^{\hat{\pi}}(s_0, \hat{\pi}(s_0)) \\ &= \mathbb{E}_{s_0} Q^\star(s_0, \pi^\star(s_0)) - Q^\star(s_0, \hat{\pi}(s_0)) + Q^\star(s_0, \hat{\pi}(s_0)) - Q^{\hat{\pi}}(s_0, \hat{\pi}(s_0)) \\ &\le \mathbb{E}_{s_0} Q^\star(s_0, \pi^\star(s_0)) - Q^{(T)}(s_0, \pi^\star(s_0)) + Q^{(T)}(s_0, \hat{\pi}(s_0)) - Q^\star(s_0, \hat{\pi}(s_0)) + Q^\star(s_0, \hat{\pi}(s_0)) - Q^{\hat{\pi}}(s_0, \hat{\pi}(s_0)) \\ &\le 2\varepsilon_Q + \mathbb{E}_{s_0} Q^\star(s_0, \hat{\pi}(s_0)) - Q^{\hat{\pi}}(s_0, \hat{\pi}(s_0))\end{aligned}$$

Here the main inequality uses the fact that $\hat{\pi}$ is greedy with respect to $Q^{(T)}$, so $Q^{(T)}(s, \hat{\pi}(s)) \geq Q^{(T)}(s, \pi^\star(s))$. Now, we may unroll the last term here since $Q^\star(s_0, \hat{\pi}(s_0)) = r(s_0, \hat{\pi}(s_0)) + \gamma \mathbb{E}_{s_1 \sim P(s_0, \hat{\pi}(s_0))} Q^\star(s_1, \pi^\star(s_1))$ while $Q^{\hat{\pi}}(s_0, \hat{\pi}(s_0)) = r(s_0, \hat{\pi}(s_0)) + \gamma \mathbb{E}_{s_1 \sim P(s_0, \hat{\pi}(s_0))} Q^{\hat{\pi}}(s_1, \hat{\pi}(s_1))$. This gives

$$\mathbb{E}_{s_0} Q^\star(s_0, \hat{\pi}(s_0)) - Q^{\hat{\pi}}(s_0, \hat{\pi}(s_0)) = \gamma \mathbb{E}_{s_1 \sim d_1^{\hat{\pi}}} Q^\star(s_1, \pi^\star(s_1)) - Q^{\hat{\pi}}(s_1, \hat{\pi}(s_1)),$$

which has the same form as what we started with. Thus by unrolling, we obtain the bound:

$$J(\pi^\star) - J(\hat{\pi}) \leq \frac{2\varepsilon_Q}{1-\gamma}$$

4. **Bellman rank.** The key calculation is that in a linear MDP, the Bellman backup of any function $g : \mathcal{S} \to \mathbb{R}$ is *linear* in the true features $\phi^\star$. To see this, consider some policy π and note that

$$\begin{aligned}
\mathbb{E}_{s_h \sim d_h^\pi} g(s_h) &= \mathbb{E}_{s_{h-1}, a_{h-1} \sim d_{h-1}^\pi} \int P(s_h \mid s_{h-1}, a_{h-1}) g(s_h) ds_h \\
&= \mathbb{E}_{s_{h-1}, a_{h-1} \sim d_{h-1}^\pi} \int \langle \phi^\star(s_{h-1}, a_{h-1}), \mu^\star(s_h) \rangle \, g(s_h) ds_h \\
&= \left\langle \mathbb{E}_{s_{h-1}, a_{h-1} \sim d_{h-1}^\pi} \phi^\star(s_{h-1}, a_{h-1}), \int \mu^\star(s_h) g(s_h) ds_h \right\rangle
\end{aligned}$$

This immediately shows that the Bellman error $\mathcal{E}_h(\pi, f)$ factorizes, since we can take g in the above derivation to be $g : s_h \mapsto \mathbb{E}_{a_h \sim \pi_f(s_h)}[(f - \mathcal{T} f)(s_h, a_h)]$, which is only a function of the state s_h. Then

$$\begin{aligned}
\mathcal{E}_h(\pi, f) &= \mathbb{E}_{s_h \sim d_h^\pi} \mathbb{E}_{a_h \sim \pi_f(s_h)} \left[(f - \mathcal{T} f)(s_h, a_h) \right] = \mathbb{E}_{s_h \sim d_h^\pi} \left[g(s_h) \right] \\
&= \left\langle \mathbb{E}_{s_{h-1}, a_{h-1} \sim d_{h-1}^\pi} \phi^\star(s_{h-1}, a_{h-1}), \int \mu^\star(s_h) g(s_h) ds_h \right\rangle
\end{aligned}$$

Thus, we can take $w_h(\pi) = \mathbb{E}_{s_{h-1}, a_{h-1} \sim d_{h-1}^\pi} \phi^\star(s_{h-1}, a_{h-1})$ and we can take $v_h(f) = \int \mu^\star(s_h) \mathbb{E}_{a \sim \pi_f(s_h)}[(f - \mathcal{T} f)(s_h, a_h)] ds_h$ and see that the Bellman rank is d. Note that these embeddings also satisfy reasonable normalization conditions, since we typically assume $\|\phi_h^\star\|_2$ is bounded, and it is natural to assume that both f and $\mathcal{T} f$ are bounded as well.