Comp 767 Assignment 2

Jonathan Pearce 260672004

March 2, 2020

Problem 1.B.i We start by providing the explicit form of $V^\pi(s)$, the true value function at state s

$$V^\pi(s) = \mathbb{E}[r_0 + \gamma r_1 + ... + \gamma^{k-1} r_{k-1} + \gamma^k V^\pi(s_k)]$$

Where r_i is the ith reward received along the trajectory under π. $V^\pi(s_k)$ is the true value function at the kth state in the trajectory. γ is the discount factor. It follows from the linearity of expectation,

$$\begin{aligned} &= \mathbb{E}[r_0] + \mathbb{E}[\gamma r_1] + ... + \mathbb{E}[\gamma^{k-1} r_{k-1}] + \mathbb{E}[\gamma^k V^\pi(s_k)] \\ &= \mathbb{E}[r_0] + \gamma\mathbb{E}[r_1] + ... + \gamma^{k-1}\mathbb{E}[r_{k-1}] + \gamma^k\mathbb{E}[V^\pi(s_k)] \end{aligned}$$

We now expand and provide an upper bound for Δ_t using the formula for $\bar{V}^k_{\pi,t}(s)$ from the question and $V_\pi(s)$ from above:

$$\begin{aligned} \Delta_t &= \max_s \mid \bar{V}^k_{\pi,t}(s) - V_\pi(s) \mid \\ &= \max_s \left| \frac{1}{n}\sum_{i=1}^n \Big(\sum_{j=1}^k \gamma^{j-1} r^{(i)}_{j-1}\Big) + \gamma^k \bar{V}^k_{\pi,t-1}(s^{(i)}_k) - V_\pi(s) \right| \\ &= \max_s \left| \frac{1}{n}\sum_{i=1}^n \Big(\sum_{j=1}^k \gamma^{j-1} r^{(i)}_{j-1}\Big) + \gamma^k \bar{V}^k_{\pi,t-1}(s^{(i)}_k) - \Big(\mathbb{E}[r_0] + \gamma\mathbb{E}[r_1] + ... + \gamma^{k-1}\mathbb{E}[r_{k-1}] + \gamma^k\mathbb{E}[V^\pi(s_k)]\Big) \right| \\ &= \max_s \left| \Big(\frac{1}{n}\sum_{i=1}^n r^{(i)}_0 - \mathbb{E}[r_0]\Big) + ... + \Big(\frac{1}{n}\sum_{i=1}^n \gamma^{k-1} r^{(i)}_{k-1} - \gamma^{k-1}\mathbb{E}[r_{k-1}]\Big) + \Big(\frac{1}{n}\sum_{i=1}^n \gamma^k \bar{V}^k_{\pi,t-1}(s^{(i)}_k) - \gamma^k\mathbb{E}[V^\pi(s_k)]\Big) \right| \\ &= \Big(\frac{1}{n}\sum_{i=1}^n r^{(i)}_0 - \mathbb{E}[r_0]\Big) + ... + \Big(\frac{1}{n}\sum_{i=1}^n \gamma^{k-1} r^{(i)}_{k-1} - \gamma^{k-1}\mathbb{E}[r_{k-1}]\Big) + \max_s \left| \Big(\frac{1}{n}\sum_{i=1}^n \gamma^k \bar{V}^k_{\pi,t-1}(s^{(i)}_k) - \gamma^k\mathbb{E}[V^\pi(s_k)]\Big) \right| \\ &= \Big(\frac{1}{n}\sum_{i=1}^n r^{(i)}_0 - \mathbb{E}[r_0]\Big) + ... + \gamma^{k-1}\Big(\frac{1}{n}\sum_{i=1}^n r^{(i)}_{k-1} - \mathbb{E}[r_{k-1}]\Big) + \gamma^k \max_s \left| \Big(\frac{1}{n}\sum_{i=1}^n \bar{V}^k_{\pi,t-1}(s^{(i)}_k) - \mathbb{E}[V^\pi(s_k)]\Big) \right| \end{aligned}$$

$\left|\bar{V}^k_{\pi,t-1}(s^{(i)}_k) - \mathbb{E}[V^\pi(s_k)]\right| \leq \Delta_{t-1}$ by assumption. Therefore we obtain,

$$\leq \Big(\frac{1}{n}\sum_{i=1}^n r^{(i)}_0 - \mathbb{E}[r_0]\Big) + ... + \gamma^{k-1}\Big(\frac{1}{n}\sum_{i=1}^n r^{(i)}_{k-1} - \mathbb{E}[r_{k-1}]\Big) + \gamma^k \Delta_{t-1}$$

Using Hoeffding's inequality, we can perform a deviation analysis on the jth term in the equation above,

$$\mathbb{P}\Big[\left|\frac{1}{n}\sum_{i=1}^n r^{(i)}_j - \mathbb{E}[r_j]\right| \geq \epsilon \Big] \leq 2\ \exp\Big(\frac{-2n^2\epsilon^2}{n(1-(-1))^2}\Big) = 2\ \exp\Big(\frac{-n\epsilon^2}{2}\Big)$$

However, we want a deviation analysis that holds for all k terms (of this form) simultaneously. To do this we use the union bound

$$\mathbb{P}\Big[\bigcup_{j=1}^k \left|\frac{1}{n}\sum_{i=1}^n r^{(i)}_j - \mathbb{E}[r_j]\right| \geq \epsilon \Big] \leq \sum_{j=1}^k \mathbb{P}\Big[\left|\frac{1}{n}\sum_{i=1}^n r^{(i)}_j - \mathbb{E}[r_j]\right| \geq \epsilon \Big] \leq 2k\ \exp\Big(\frac{-n\epsilon^2}{2}\Big)$$

Equating to δ and solving for ϵ:

$$\begin{aligned} \delta &= 2ke^{\frac{-n\epsilon^2}{2}} \\ \frac{\delta}{2k} &= e^{\frac{-n\epsilon^2}{2}} \\ \ln\Big(\frac{2k}{\delta}\Big) &= \frac{n\epsilon^2}{2} \end{aligned}$$

$$\sqrt{\frac{2}{n}\ln(\frac{2k}{\delta})} = \epsilon$$

Therefore with probability at least $1-\delta$, $\left|\frac{1}{n}\sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j]\right| \leq \epsilon = \sqrt{\frac{2}{n}\ln(\frac{2k}{\delta})}$ holds for all k terms simultaneously. It follows,

$$\begin{aligned}\Delta_t &\leq \Big(\frac{1}{n}\sum_{i=1}^{n} r_0^{(i)} - \mathbb{E}[r_0]\Big) + ... + \gamma^{k-1}\Big(\frac{1}{n}\sum_{i=1}^{n} r_{k-1}^{(i)} - \mathbb{E}[r_{k-1}]\Big) + \gamma^k \Delta_{t-1} \\ &\leq \epsilon + \gamma\epsilon + ... + \gamma^{k-1}\epsilon + \gamma^k\Delta_{t-1} \\ &= \epsilon\Big(\frac{1-\gamma^k}{1-\gamma}\Big) + \gamma^k\Delta_{t-1}\end{aligned}$$

We conclude that if $\epsilon = \sqrt{\frac{2}{n}\ln(\frac{2k}{\delta})}$ then, $\Delta_t \leq \epsilon\Big(\frac{1-\gamma^k}{1-\gamma}\Big) + \gamma^k\Delta_{t-1}$ holds with probability $1-\delta$.

Problem 1.B.ii We start by providing the explicit form of $V^\pi(s)$, the true value function at state s

$$V^\pi(s) = \mathbb{E}\Big[(1-\lambda)\sum_{k=1}^{\infty}\lambda^{k-1}\Big(\sum_{j=1}^{k}\gamma^{j-1}r_{j-1} + \gamma^k V^\pi(s_k)\Big)\Big]$$

Where r_i is the ith reward received along the trajectory under π. $V^\pi(s_k)$ is the true value function at the kth state in the trajectory. γ is the discount factor. λ is the decay parameter. It follows from the linearity of expectation,

$$\begin{aligned}&= (1-\lambda)\sum_{k=1}^{\infty}\lambda^{k-1}\Big(\sum_{j=1}^{k}\mathbb{E}[\gamma^{j-1}r_{j-1}] + \mathbb{E}[\gamma^k V^\pi(s_k)]\Big) \\ &= (1-\lambda)\sum_{k=1}^{\infty}\lambda^{k-1}\Big(\sum_{j=1}^{k}\gamma^{j-1}\mathbb{E}[r_{j-1}] + \gamma^k\mathbb{E}[V^\pi(s_k)]\Big) \\ &= (1-\lambda)\Big((\mathbb{E}[r_0] + \gamma\mathbb{E}[V^\pi(s_1)]) + \lambda(\mathbb{E}[r_0] + \gamma\mathbb{E}[r_1] + \gamma^2\mathbb{E}[V^\pi(s_2)]) + ...\Big) \\ &= \Big(\mathbb{E}[r_0] + \gamma\lambda\mathbb{E}[r_1] + (\gamma\lambda)^2\mathbb{E}[r_2] + ...\Big) + (1-\lambda)\Big(\gamma\mathbb{E}[V^\pi(s_1)] + \lambda\gamma^2\mathbb{E}[V^\pi(s_2)] + \lambda^2\gamma^3\mathbb{E}[V^\pi(s_3)] + ...\Big) \\ &= \sum_{j=0}^{\infty}(\gamma\lambda)^j\mathbb{E}[r_j] + (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\mathbb{E}[V^\pi(s_{j+1})]\end{aligned}$$

We can now expand and simplify $\bar{V}_{\pi,t}^{\lambda}(s)$:

$$\bar{V}_{\pi,t}^{\lambda}(s) = \frac{1}{n}\sum_{i=1}^{n}(1-\lambda)\sum_{k=1}^{\infty}\lambda^{k-1}\Big(\sum_{j=1}^{k-1}\gamma^{j-1}r_{j+1}^{(i)} + \gamma^k\bar{V}_{\pi,t-1}^{\lambda}(s_{k+1}^{(i)})\Big)$$

Using the rewriting of $V^\pi(s)$ above, we obtain:

$$\begin{aligned}&= \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=0}^{\infty}(\gamma\lambda)^j r_j^{(i)} + (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\bar{V}_{\pi,t-1}^{\lambda}(s_{j+1}^{(i)})\Big) \\ &= \sum_{j=0}^{\infty}(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}r_j^{(i)} + (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}\bar{V}_{\pi,t-1}^{\lambda}(s_{j+1}^{(i)})\end{aligned}$$

We now expand and provide an upper bound for Δ_t using the formula for $\bar{V}_{\pi,t}^{\lambda}(s)$ above and $V_\pi(s)$ also above:

$$\begin{aligned}\Delta_t &= \max_s \mid \bar{V}_{\pi,t}^{\lambda}(s) - V_\pi(s) \mid \\ &= \max_s\left|\sum_{j=0}^{\infty}(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}r_j^{(i)} + (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}\bar{V}_{\pi,t-1}^{\lambda}(s_{j+1}^{(i)}) - V_\pi(s)\right| \\ &= \max_s\left|\sum_{j=0}^{\infty}(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}r_j^{(i)} + (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}\bar{V}_{\pi,t-1}^{\lambda}(s_{j+1}^{(i)}) - \sum_{j=0}^{\infty}(\gamma\lambda)^j\mathbb{E}[r_j] - (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\mathbb{E}[V^\pi(s_{j+1})]\right| \\ &= \max_s\left|\sum_{j=0}^{\infty}(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}r_j^{(i)} - \sum_{j=0}^{\infty}(\gamma\lambda)^j\mathbb{E}[r_j] + (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\frac{1}{n}\sum_{i=1}^{n}\bar{V}_{\pi,t-1}^{\lambda}(s_{j+1}^{(i)}) - (1-\lambda)\sum_{j=0}^{\infty}\gamma(\gamma\lambda)^j\mathbb{E}[V^\pi(s_{j+1})]\right|\end{aligned}$$

$$= \max_s \left| \sum_{j=0}^{\infty} (\gamma\lambda)^j \Big(\frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j] \Big) + (1-\lambda) \sum_{j=0}^{\infty} \gamma(\gamma\lambda)^j \Big(\frac{1}{n} \sum_{i=1}^{n} \bar{V}^{\lambda}_{\pi,t-1}(s^{(i)}_{j+1}) - \mathbb{E}[V^{\pi}(s_{j+1})] \Big) \right|$$

$$= \sum_{j=0}^{\infty} (\gamma\lambda)^j \Big(\frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j] \Big) + (1-\lambda) \sum_{j=0}^{\infty} \gamma(\gamma\lambda)^j \max_s \left| \Big(\frac{1}{n} \sum_{i=1}^{n} \bar{V}^{\lambda}_{\pi,t-1}(s^{(i)}_{j+1}) - \mathbb{E}[V^{\pi}(s_{j+1})] \Big) \right|$$

$\left| \bar{V}^{k}_{\pi,t-1}(s^{(i)}_{k}) - \mathbb{E}[V^{\pi}(s_k)] \right| \leq \Delta_{t-1}$ by assumption. Therefore we obtain,

$$\leq \sum_{j=0}^{\infty} (\gamma\lambda)^j \Big(\frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j] \Big) + (1-\lambda) \sum_{j=0}^{\infty} \gamma(\gamma\lambda)^j \Delta_{t-1}$$

$$= \sum_{j=0}^{k-1} (\gamma\lambda)^j \Big(\frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j] \Big) + \sum_{j=k}^{\infty} (\gamma\lambda)^j \Big(\frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j] \Big) + (1-\lambda) \sum_{j=0}^{\infty} \gamma(\gamma\lambda)^j \Delta_{t-1}$$

Here we bound the terms in the summation from $j = 0...(k-1)$ using Hoeffding's inequality and the union bound in the exact same way as in part i. Therefore with probability at least $1-\delta$, $\left| \frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j] \right| \leq \epsilon = \sqrt{\frac{2}{n} \ln(\frac{2k}{\delta})}$ holds for all k terms simultaneously. For the terms in the summation from $j = k, ...\infty$ we will assume maximum variance. Because $r_i \in [-1,1]\ \forall i$, $\left| \frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} - \mathbb{E}[r_j] \right| \leq 1-(-1) = 2$. In order to make our bound on Δ_t as tight as possible we select the value k that minimizes the sum of these two summations; $j = 0, ...(k-1)$ and $j = k, ...\infty$.

$$\Delta_t \leq \min_k \Big(\sum_{j=0}^{k-1} (\gamma\lambda)^j \epsilon + \sum_{j=k}^{\infty} (\gamma\lambda)^j 2 \Big) + (1-\lambda) \sum_{j=0}^{\infty} \gamma(\gamma\lambda)^j \Delta_{t-1}$$

$$= \min_k \Big(\epsilon \sum_{j=0}^{k-1} (\gamma\lambda)^j + 2 \sum_{j=k}^{\infty} (\gamma\lambda)^j \Big) + (1-\lambda)\gamma\Delta_{t-1} \sum_{j=0}^{\infty} (\gamma\lambda)^j$$

$$= \min_k \Big(\epsilon \sum_{j=0}^{k-1} (\gamma\lambda)^j + 2(\gamma\lambda)^k \sum_{j=0}^{\infty} (\gamma\lambda)^j \Big) + (1-\lambda)\gamma\Delta_{t-1} \sum_{j=0}^{\infty} (\gamma\lambda)^j$$

$$= \min_k \Big(\epsilon \frac{1-(\gamma\lambda)^k}{1-\gamma\lambda} + 2(\gamma\lambda)^k \frac{1}{1-\gamma\lambda} \Big) + (1-\lambda)\gamma\Delta_{t-1} \frac{1}{1-\gamma\lambda}$$

$$= \min_k \Big(\frac{1-(\gamma\lambda)^k}{1-\gamma\lambda} \epsilon + 2 \frac{(\gamma\lambda)^k}{1-\gamma\lambda} \Big) + \frac{(1-\lambda)\gamma}{1-\gamma\lambda} \Delta_{t-1}$$

We conclude that if $\epsilon = \sqrt{\frac{2}{n} \ln(\frac{2k}{\delta})}$ then, $\Delta_t \leq \min_k \Big(\frac{1-(\gamma\lambda)^k}{1-\gamma\lambda} \epsilon + 2 \frac{(\gamma\lambda)^k}{1-\gamma\lambda} \Big) + \frac{(1-\lambda)\gamma}{1-\gamma\lambda} \Delta_{t-1}$ holds with probability $1-\delta$.

Note for part ii, the second term in the minimum expression has a factor of 2 that does not appear in the original paper. The paper did not appear to specify any information on how the rewards are distributed in [-1,1]. Therefore that factor of 2 assumes a worst case of an empirical average being -1 and the expectation being 1 or vice versa. However if the rewards are sampled uniformly from [-1,1] then the expectation of any reward would be 0, therefore the maximum deviation would be 1 and my result would be the same as the result in the paper.

Problem 2.B.i Before performing policy evaluation or control, certainty-equivalence has to process all of the trajectories in $\mathcal{D}$ in order to build (or update) the model of the environment. In contrast methods that estimate the q-value function are capable of running online, which means they can refine their estimates of the q-value function immediately after receiving more data, instead of collecting data and processing at the end like certainty-equivalence. Certainty-equivalence estimates a model of the environment (transitions and rewards). Therefore certainty-equivalence has space complexity $O(|\mathcal{S}|^2|\mathcal{A})$. Methods that estimate the q-value functions are more space efficient and have only $O(|\mathcal{S}||\mathcal{A}|)$ space complexity. Therefore it appears methods that estimate the q-value function are more space and time efficient than certainty-equivalence. However, the main benefit of certainty-equivalence (and model based RL algorithms in general) is that it is more sample efficient than model free methods that estimate the q-value functions directly [1].

Problem 2.B.ii Assuming $\hat{R}(s,a) \in [0, R_{\max}]$. By Hoeffding's inequality we have,

$$\mathbb{P}\Big[\Big|\hat{R}(s,a) - R(s,a)\Big| \geq \epsilon\Big] \leq 2\,\exp\Big(\frac{-2n^2\epsilon^2}{n(R_{\max}-0)^2}\Big) = 2\,\exp\Big(\frac{-2n\epsilon^2}{R_{\max}^2}\Big)$$

However we want a deviation analysis that holds for all $|\mathcal{S}\times\mathcal{A}|$ state-action pairs simultaneously. To do this we use the union bound,

$$\mathbb{P}\Big[\bigcup_{(s,a)\in|\mathcal{S}\times\mathcal{A}|}\Big|\hat{R}(s,a) - R(s,a)\Big| \geq \epsilon\Big] \leq \sum_{(s,a)\in|\mathcal{S}\times\mathcal{A}|}\mathbb{P}\Big[\ \Big|\hat{R}(s,a) - R(s,a)\Big| \geq \epsilon\Big] \leq 2\,|\mathcal{S}\times\mathcal{A}|\ \exp\Big(\frac{-2n\epsilon^2}{R_{\max}^2}\Big)$$

Equating to δ and solving for ϵ:

$$\begin{aligned}
\delta = 2\,|\mathcal{S}\times\mathcal{A}|\ \exp\Big(\frac{-2n\epsilon^2}{R_{\max}^2}\Big)\\
\frac{\delta}{2\,|\mathcal{S}\times\mathcal{A}|} &= \exp\Big(\frac{-2n\epsilon^2}{R_{\max}^2}\Big)\\
\ln\frac{2\,|\mathcal{S}\times\mathcal{A}|}{\delta} &= \frac{2n\epsilon^2}{R_{\max}^2}\\
\sqrt{\frac{R_{\max}^2}{2n}\ln\frac{2\,|\mathcal{S}\times\mathcal{A}|}{\delta}} &= \epsilon\\
R_{\max}\sqrt{\frac{1}{2n}\ln\frac{2\,|\mathcal{S}\times\mathcal{A}|}{\delta}} &= \epsilon
\end{aligned}$$

Therefore with probability at least $1-\delta$, $\Big|\hat{R}(s,a) - R(s,a)\Big| \leq \epsilon = R_{\max}\sqrt{\frac{1}{2n}\ln\frac{2|\mathcal{S}\times\mathcal{A}|}{\delta}}$ holds for all state-action pairs simultaneously. Equivalently, with probability $1-\delta$,

$$\max\nolimits_{s,a}\Big|\hat{R}(s,a) - R(s,a)\Big| \leq R_{\max}\sqrt{\frac{1}{2n}\ln\frac{2\,|\mathcal{S}\times\mathcal{A}|}{\delta}} \tag{1}$$

Now, assuming $\hat{P}(s'|s,a) \in [0,1]$. By Hoeffding's inequality we have,

$$\mathbb{P}\Big[\Big|\hat{P}(s'|s,a) - P(s'|s,a)\Big| \geq \epsilon\Big] \leq 2\,\exp\Big(\frac{-2n^2\epsilon^2}{n(1-0)^2}\Big) = 2\,\exp(-2n\epsilon^2)$$

However we want a deviation analysis that holds for all $|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|$ state-action-state triples simultaneously. To do this we use the union bound

$$\mathbb{P}\Big[\bigcup_{(s,a,s')\in|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}\Big|\hat{P}(s'|s,a) - P(s'|s,a)\Big| \geq \epsilon\Big] \leq \sum_{(s,a,s')\in|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}\mathbb{P}\Big[\ \Big|\hat{P}(s'|s,a) - P(s'|s,a)\Big| \geq \epsilon\Big] \leq 2\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|\ \exp(-2n\epsilon^2)$$

Equating to δ and solving for ϵ:

$$\begin{aligned}
\delta = 2\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|\ \exp(-2n\epsilon^2)\\
\frac{\delta}{2\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|} &= \exp(-2n\epsilon^2)\\
\ln\frac{2\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta} &= 2n\epsilon^2\\
\sqrt{\frac{1}{2n}\ln\frac{2\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}} &= \epsilon
\end{aligned}$$

Therefore with probability at least $1-\delta$, $\left|\hat{P}(s'|s,a) - P(s'|s,a)\right| \leq \epsilon = \sqrt{\frac{1}{2n}\ln\frac{2|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}}$ holds for all state-action-state triples simultaneously. Equivalently, with probability $1-\delta$,

$$\max_{s,a,s'} \left|\hat{P}(s'|s,a) - P(s'|s,a)\right| \leq \sqrt{\frac{1}{2n}\ln\frac{2\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}} \tag{2}$$

Substituting $\delta = \frac{\delta}{2}$ in (1) and (2), we obtain the following two inequalities that hold simultaneously with probability $1-\delta$,

$$\max_{s,a} \left|\hat{R}(s,a) - R(s,a)\right| \leq R_{\max}\sqrt{\frac{1}{2n}\ln\frac{4\,|\mathcal{S}\times\mathcal{A}|}{\delta}}$$

$$\max_{s,a,s'} \left|\hat{P}(s'|s,a) - P(s'|s,a)\right| \leq \sqrt{\frac{1}{2n}\ln\frac{4\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}}$$

We now provide a proof of the simulation lemma (Kearns & Singh 1998; Near-Optimal Reinforcement Learning in Polynomial Time).

We begin by making this observation,

$$\begin{aligned}
\mathbb{E}[V^{\pi}_{\hat{M}}(s)] &= \mathbb{E}[r_0 + \gamma r_1 + \gamma^2 r_2 + ...] \\
&= \mathbb{E}[r_0] + \mathbb{E}[\gamma r_1] + \mathbb{E}[\gamma^2 r_2] + ... \\
&= \mathbb{E}[r_0] + \gamma\mathbb{E}[r_1] + \gamma^2\mathbb{E}[r_2] + ...
\end{aligned}$$

Assuming the reward distribution is uniform from 0 to $R_{\max}$ we obtain,

$$\begin{aligned}
&= \frac{R_{\max}}{2} + \gamma\frac{R_{\max}}{2} + \gamma^2\frac{R_{\max}}{2} + ... \\
&= \frac{R_{\max}}{2}(1 + \gamma + \gamma^2 + ...) \\
&= \frac{R_{\max}}{2}\frac{1}{1-\gamma} \\
&= \frac{R_{\max}}{2(1-\gamma)}
\end{aligned}$$

Next we develop a bound for $\left|\left|\hat{P}^{\pi}(s,\cdot) - P^{\pi}(s,\cdot)\right|\right|_1$ using some properties from [2],

$$\begin{aligned}
\max_{s,a} \left|\left|\hat{P}^{\pi}(s,a) - P^{\pi}(s,a)\right|\right|_1 &\leq \max_{s,a}|\mathcal{S}|\left|\left|\hat{P}^{\pi}(s,a) - P^{\pi}(s,a)\right|\right|_\infty \\
&= |\mathcal{S}|\max_{s,a,s'} \left|\hat{P}(s'|s,a) - P(s'|s,a)\right| \\
&\leq |\mathcal{S}|\sqrt{\frac{1}{2n}\ln\frac{4\,|\mathcal{S}\times\mathcal{A}\times\mathcal{S}|}{\delta}} \\
&= \epsilon_P
\end{aligned}$$

For all states $s \in S$

$$\begin{aligned}
|V^{\pi}_{\hat{M}}(s) - V^{\pi}_{M}(s)| &= \left|\left(\hat{R}^{\pi}(s,\cdot) + \gamma\hat{P}^{\pi}(s,\cdot)V^{\pi}_{\hat{M}}\right) - \left(R^{\pi}(s,\cdot) + \gamma P^{\pi}(s,\cdot)V^{\pi}_{M}\right)\right| \\
&= \left|\left(\hat{R}^{\pi}(s,\cdot) - R^{\pi}(s,\cdot)\right) + \left(\gamma\hat{P}^{\pi}(s,\cdot)V^{\pi}_{\hat{M}} - \gamma P^{\pi}(s,\cdot)V^{\pi}_{M}\right)\right| \\
&\leq \left|\hat{R}^{\pi}(s,\cdot) - R^{\pi}(s,\cdot)\right| + \left|\gamma\hat{P}^{\pi}(s,\cdot)V^{\pi}_{\hat{M}} - \gamma P^{\pi}(s,\cdot)V^{\pi}_{M}\right| \\
&\leq \epsilon_R + \left|\gamma\hat{P}^{\pi}(s,\cdot)V^{\pi}_{\hat{M}} - \gamma P^{\pi}(s,\cdot)V^{\pi}_{M}\right| \\
&= \epsilon_R + \gamma\left|\hat{P}^{\pi}(s,\cdot)V^{\pi}_{\hat{M}} - P^{\pi}(s,\cdot)V^{\pi}_{M}\right| \\
&= \epsilon_R + \gamma\left|\hat{P}^{\pi}(s,\cdot)V^{\pi}_{\hat{M}} - P^{\pi}(s,\cdot)V^{\pi}_{\hat{M}} + P^{\pi}(s,\cdot)V^{\pi}_{\hat{M}} - P^{\pi}(s,\cdot)V^{\pi}_{M}\right| \\
&= \epsilon_R + \gamma\left|\left(\hat{P}^{\pi}(s,\cdot) - P^{\pi}(s,\cdot)\right)V^{\pi}_{\hat{M}} + P^{\pi}(s,\cdot)\left(V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right)\right| \\
&\leq \epsilon_R + \gamma\left|\left(\hat{P}^{\pi}(s,\cdot) - P^{\pi}(s,\cdot)\right)V^{\pi}_{\hat{M}}\right| + \gamma\left|P^{\pi}(s,\cdot)\left(V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right)\right| \\
&\leq \epsilon_R + \gamma\left|\left(\hat{P}^{\pi}(s,\cdot) - P^{\pi}(s,\cdot)\right)V^{\pi}_{\hat{M}}\right| + \gamma\left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_\infty
\end{aligned}$$

In our setting we have n large and $\hat{R} \approx R$, therefore $V^{\pi}_{\hat{M}} \approx \mathbb{E}[V^{\pi}_{\hat{M}}(s)] \cdot \mathbf{1}$. Since $\hat{P}^{\pi}(s,\cdot)$ and $P^{\pi}(s,\cdot)$ are valid probability distributions both summing to 1, their difference is a distribution that sums up to 0. Therefore we can shift $V^{\pi}_{\hat{M}}$ down by $\mathbb{E}[V^{\pi}_{\hat{M}}(s)] \cdot \mathbf{1}$ without changing the value of our current bound.

$$\begin{aligned}
&= \epsilon_R + \gamma \left| \left(\hat{P}^{\pi}(s,\cdot) - P^{\pi}(s,\cdot)\right)\left(V^{\pi}_{\hat{M}} - \mathbb{E}[V^{\pi}_{\hat{M}}] \cdot \mathbf{1}\right) \right| + \gamma \left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty} \\
&= \epsilon_R + \gamma \left| \left(\hat{P}^{\pi}(s,\cdot) - P^{\pi}(s,\cdot)\right)\left(V^{\pi}_{\hat{M}} - \frac{R_{\max}}{2(1-\gamma)} \cdot \mathbf{1}\right) \right| + \gamma \left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty} \\
&\leq \epsilon_R + \gamma \left|\left|\hat{P}^{\pi}(s,\cdot) - P^{\pi}(s,\cdot)\right|\right|_1 \left|\left|V^{\pi}_{\hat{M}} - \frac{R_{\max}}{2(1-\gamma)}\right|\right|_{\infty} + \gamma \left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty} \\
&\leq \epsilon_R + \gamma \epsilon_P \left|\left|V^{\pi}_{\hat{M}} - \frac{R_{\max}}{2(1-\gamma)}\right|\right|_{\infty} + \gamma \left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty} \\
&= \epsilon_R + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)} + \gamma \left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty}
\end{aligned}$$

Because the above holds for all $s \in \mathcal{S}$ we therefore have,

$$\left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty} \leq \epsilon_R + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)} + \gamma \left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty}$$

We can now solve for $\left|\left|V^{\pi}_{\hat{M}} - V^{\pi}_{M}\right|\right|_{\infty}$

$$\begin{aligned}
||V^{\pi}_{\hat{M}} - V^{\pi}_{M}||_{\infty} - \gamma ||V^{\pi}_{\hat{M}} - V^{\pi}_{M}||_{\infty} &\leq \epsilon_R + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)} \\
||V^{\pi}_{\hat{M}} - V^{\pi}_{M}||_{\infty}(1-\gamma) &\leq \epsilon_R + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)} \\
||V^{\pi}_{\hat{M}} - V^{\pi}_{M}||_{\infty} &\leq \frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)^2}
\end{aligned} \tag{3}$$

Problem 2.B.iii

$$V^{*}_{M}(s) - V^{\pi^*_{\hat{M}}}_{M}(s) = V^{\pi^*_M}_{M}(s) - V^{\pi^*_M}_{\hat{M}}(s) + V^{\pi^*_{\hat{M}}}_{\hat{M}}(s) - V^{\pi^*_{\hat{M}}}_{M}(s)$$

Because $V^{\pi^*_M}_{\hat{M}}(s) \leq V^{\pi^*_{\hat{M}}}_{\hat{M}}(s)$, we obtain

$$\begin{aligned}
&\leq V^{\pi^*_M}_{M}(s) - V^{\pi^*_M}_{\hat{M}}(s) + V^{\pi^*_{\hat{M}}}_{\hat{M}}(s) - V^{\pi^*_{\hat{M}}}_{M}(s) \\
&\leq ||V^{\pi^*_M}_{M} - V^{\pi^*_M}_{\hat{M}}||_{\infty} + ||V^{\pi^*_{\hat{M}}}_{\hat{M}} - V^{\pi^*_{\hat{M}}}_{M}||_{\infty} \\
&\leq 2 \sup_{\pi: S \to A} ||V^{\pi}_{\hat{M}} - V^{\pi}_{M}||_{\infty}
\end{aligned}$$

Problem 2.B.iv

$$\begin{aligned}
V^{*}_{M}(s) - V^{\pi^*_{\hat{M}}}_{M}(s) &\leq 2 \sup_{\pi: S \to A} ||V^{\pi}_{\hat{M}}(s) - V^{\pi}_{M}(s)||_{\infty} \\
&\leq 2\Big(\frac{\epsilon_R}{1-\gamma} + \frac{\gamma \epsilon_P R_{\max}}{2(1-\gamma)^2}\Big) \\
&= 2\Big(\frac{R_{\max}\sqrt{\ln\frac{4|S \text{ X } A|}{\delta}}}{\sqrt{2n}(1-\gamma)} + \frac{\gamma R_{\max}|S|\sqrt{\ln\frac{4|S \text{ X } A \text{ X } S|}{\delta}}}{2\sqrt{2n}(1-\gamma)^2}\Big)
\end{aligned}$$

Because $\frac{\gamma}{2} \in [0, \frac{1}{2}]$

$$\leq 2\Big(\frac{R_{\max}\sqrt{\ln\frac{4|S \text{ X } A|}{\delta}}}{\sqrt{2n}(1-\gamma)} + \frac{R_{\max}|S|\sqrt{\ln\frac{4|S \text{ X } A \text{ X } S|}{\delta}}}{\sqrt{2n}(1-\gamma)^2}\Big)$$

Because, $\sqrt{\ln\frac{4|S \text{ X } A|}{\delta}} \leq \sqrt{\ln\frac{4|S \text{ X } A \text{ X } S|}{\delta}}$, $(1-\gamma) \in [0,1]$ and $|S| \geq 1$.

$$\leq 4\Big(\frac{R_{\max}|S|\sqrt{\ln\frac{4|S \text{ X } A \text{ X } S|}{\delta}}}{\sqrt{2n}(1-\gamma)^2}\Big)$$

It follows,

$$V_M^*(s) - V_M^{\pi^*_{\hat{M}}}(s) = O\Big(\frac{R_{\max}|S|\sqrt{\ln\frac{4|S\text{ x }A\text{ x }S|}{\delta}}}{\sqrt{n}(1-\gamma)^2}\Big)$$

If we suppress $R_{\max}$ and the logarithm, which grows slowly relative to $|S|$, we obtain the desired result

$$V_M^*(s) - V_M^{\pi^*_{\hat{M}}}(s) = O\Big(\frac{|S|}{\sqrt{n}(1-\gamma)^2}\Big)$$

References

1. `https://people.eecs.berkeley.edu/~cbfinn/_files/mbrl_bootcamp.pdf`
2. `https://www.cs.ubc.ca/~schmidtm/Courses/540-F14/norms.pdf`