COMP767 A1 Theory Questions

Dongyan Lin
McGill ID 260669424
dongyan.lin@mail.mcgill.ca

Justin Dumouchelle
McGill ID 260954626
justin.dumouchelle@mail.mcgill.ca

January 2020

1

Let $X_1^i, ..., X_{T/K}^i$ be T/K random variables for T/K samples pulled from the i-th arm. Define $\hat{\mu}_i = (X_1^i + ... + X_{T/K}^i)/(T/K)$. Define the event E^i to be the event that the T/K samples from the i-th arm deviated from the mean μ_i by at least ϵ (i.e. the event that $|\hat{\mu}_i - \mu_i| > \epsilon$). Using the union of these events we can obtain a sample complexity required to obtain a probability of δ that no arms samples deviate from their mean by an absolute difference of ϵ. The union of these events is bounded by

$$\begin{aligned}
\mathbf{Pr}[\bigcup_{i=1}^{K} E^i] &\leq \sum_{i=1}^{K} \mathbf{Pr}[E^i] \text{ , by union bound} \\
&= \sum_{i=1}^{K} \mathbf{Pr}[|\hat{\mu}_i - \mu_i| > \epsilon] \\
&\leq 2 \sum_{i=1}^{K} e^{-2\frac{T}{K}\epsilon^2} \text{ , by Hoeffding's inequality} \\
&= 2Ke^{-2\frac{T}{K}\epsilon^2}
\end{aligned}$$

This means that the probability of the empirical average reward deviating too much from the true average reward for **any** arm is bounded by $2Ke^{-2\frac{T}{K}\epsilon^2}$. In other words, with probability $1 - 2Ke^{-2\frac{T}{K}\epsilon^2}$, we guarantee that

$$|\hat{\mu}_i - \mu_i| \leq \epsilon$$

for any arm i. We can then obtain a minimum required sample complexity for T as

$$T \geq \frac{K}{2\epsilon^2} ln \frac{2K}{\delta}$$

or a big O complexity of $O(\frac{K}{\epsilon^2} ln \frac{K}{\delta})$. As ϵ is non-negative we will also trivially obtain that $\mu_i - \hat{\mu}_i \leq \epsilon$ for the same sample complexity.

Since we have that each arm will not deviate by ϵ with its true mean (with probability $1-\delta$), this implies that the arm with maximal true mean will also not deviate, thus for $i^* = \arg\max_{i \in [K]} \mu_i$, we have $\mu_{i^*} - \hat{\mu}_{i^*} \leq \epsilon$. Even if arm i^* is not selected then we will have an arm $\hat{i}$ such that if $\hat{\mu}_{\hat{i}} \geq \hat{\mu}_{i^*}$, then $\mu_{i^*} - \hat{\mu}_{\hat{i}} \leq \mu_{i^*} - \hat{\mu}_{i^*} \leq \epsilon$, again bounding the greedy choice after T pulls by ϵ with probability $1 - \delta$ as required.

2

2.1

Since $\bar{R}(s,a) = R(s,a) + \mathcal{N}(\mu, \sigma^2)$ we have that for any policy π,

$$\mathbb{E}_\pi[R(s,a)] = \mathbb{E}_\pi[\bar{R}(s,a)] - \mu \quad \forall t$$

by the linearity of expectation and the fact that the expectation of a normal distribution is the mean. Applying the same notion of linearity of expectation to Bellman's equation we get the following

$$\begin{aligned}
V_M^\pi(s) &= \mathbb{E}[G_{t+1}|S_t = s, A_t = a] \\
&= E[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...|S_t = s, A_t = a] \\
&= \big(E[\bar{R}_{t+1}] - \mu\big) + \gamma\big(E[\bar{R}_{t+2}] - \mu\big) + \gamma^2\big(E[\bar{R}_{t+3}] - \mu\big) + ... \\
&= E[\bar{R}_{t+1} + \gamma\bar{R}_{t+2} + \gamma^2\bar{R}_{t+3} + ...] - (\mu + \gamma\mu + \gamma^2\mu + ...) \\
&= V_{\bar{M}}^\pi(s) - \frac{\mu}{1-\gamma}, \text{ by the sum of a geometric series}
\end{aligned}$$

2.2

From to the vectorized form of the Bellman equation we can write $V_{\bar{M}}^\pi$ and V_M^π as,

$$V_{\bar{M}}^\pi = R + \gamma\bar{P}V_{\bar{M}}^\pi \quad \text{and} \quad V_M^\pi = R + \gamma P V_M^\pi$$

It follows that

$$\begin{aligned}
R &= V_{\bar{M}}^\pi - \gamma\bar{P}V_{\bar{M}}^\pi \\
&= V_M^\pi - \gamma P V_M^\pi
\end{aligned}$$

Since $\bar{P} = \alpha P + \beta Q$,

$$\begin{aligned}
[I - \gamma(\alpha P + \beta Q])V_{\bar{M}}^\pi &= [I - \gamma P]V_M^\pi \\
V_{\bar{M}}^\pi &= [I - \gamma(\alpha P + \beta Q)]^{-1}[I - \gamma P]V_M^\pi
\end{aligned}$$

3

Using Bellman's equation we can write the value functions at a given state for V^* and $\hat{V}$ as

$$V^*(s) = \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma V^*(s)]$$

$$\hat{V}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma \hat{V}(s)]$$

Let a_s^* be actions that maximizes $V^*(s)$ at state s, i.e. the optimal action. The greedy policy with respect to $\hat{V}(s)$ is given by

$$\hat{\pi}(s) = \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma \hat{V}(s)]$$

To obtain $V_{\hat{V}}(s)$ we can apply the Bellman operator for the policy $\hat{\pi}(s)$ to obtain

$$V_{\hat{V}}(s) = \sum_{s',r} p(s', r|s, \hat{\pi}(s))[r + \gamma V_{\hat{V}}(s)]$$

which follows since the greedy action is assigned probability 1. Since $\hat{\pi}(s)$ is the greedy action w.r.t. $\hat{V}(s)$ we have that

$$\sum_{s',r} p(s', r|s, a)[r + \gamma \hat{V}(s)] \leq \sum_{s',r} p(s', r|s, \hat{\pi}(s))[r + \gamma \hat{V}(s)] \tag{1}$$

for any action a, including a_s^*. Furthermore, since $|V^*(s) - \hat{V}(s)| \leq \epsilon, \forall s \in S$, it follows that

$$|\hat{V}(s) - V^*(s)| \leq \epsilon$$

$$-\epsilon \leq \hat{V}(s) - V^*(s) \leq \epsilon$$

$$V^*(s) - \epsilon \leq \hat{V}(s) \leq V^*(s) + \epsilon \tag{2}$$

By substituting 1 into 2 we can simplify the expression as

$$\sum_{s',r} p(s', r|s, a_s^*)[r + \gamma V^*(s) - \epsilon] \leq \sum_{s',r} p(s', r|s, \hat{\pi}(s))[r + \gamma V^*(s) + \epsilon]$$

As p is a probability distribution and the summation does not depend on γ and ϵ, we get

$$\sum_{s',r} p(s', r|s, a_s^*)[r + \gamma V^*(s)] - \gamma\epsilon \leq \sum_{s',r} p(s', r|s, \hat{\pi}(s))[r + \gamma V^*(s)] + \gamma\epsilon \tag{3}$$

Suppose y is the state which maximizes $L_{\hat{V}}(y)$ (i.e. $L_{\hat{V}}(y) \geq L_{\hat{V}}(s) \quad \forall s \in S$). Then we have

$$
\begin{aligned}
L_{\hat{V}}(y) &= V^*(y) - V_{\hat{V}}(y) \\
&= \sum_{s',r} p(s', r|y, a_y^*)[r + \gamma V^*(s')] - \sum_{s',r} p(s', r|y, \hat{\pi}(y))[r + \gamma V_{\hat{V}}(s')] \\
&\leq \sum_{s',r} p(s', r|s, \hat{\pi}(y))[r + \gamma V^*(s')] + 2\gamma\epsilon - \sum_{s',r} p(s', r|y, \hat{\pi}(y))[r + \gamma V_{\hat{V}}(s')] \text{ , by 3} \\
&= 2\gamma\epsilon + \sum_{s',r} p(s', r|y, \hat{\pi}(y))([r + \gamma V^*(s')] - [r + \gamma V_{\hat{V}}(s')]) \\
&= 2\gamma\epsilon + \gamma \sum_{s',r} p(s', r|y, \hat{\pi}(y))(V^*(s') - V_{\hat{V}}(s')) \\
&= 2\gamma\epsilon + \gamma \sum_{s',r} p(s', r|y, \hat{\pi}(y)) L_{\hat{V}}(s') \\
&\leq 2\gamma\epsilon + \gamma \sum_{s',r} p(s', r|y, \hat{\pi}(y)) L_{\hat{V}}(y) \text{ , since } L_{\hat{V}}(y) \geq L_{\hat{V}}(s') \\
&= 2\gamma\epsilon + \gamma L_{\hat{V}}(y)
\end{aligned}
$$

So we have

$$L_{\hat{V}}(y) \leq 2\gamma\epsilon + \gamma L_{\hat{V}}(y)$$

$$\Rightarrow L_{\hat{V}}(s) \leq L_{\hat{V}}(y) \leq \frac{2\gamma\epsilon}{1-\gamma}$$

as required.

4 Contributions

D.L. and J.D. contributed equally to this assignment. D.L. drafted the solutions to the theory questions and searched for hyper parameters in Coding Q1. J.D. edited the solutions to the theory questions and wrote the solutions to the coding questions. D.L. and J.D. contributed to the final revision and submission of the assignment.

A link to the colab notebook can be found below

https://colab.research.google.com/drive/1mvHgbGU95RfJSjQqwjQjJvdOsDmNVcfc