

Problem Set 2 Solutions

姓名: 黄睿

学号: MP1933008

2019-10-23

1 Problem 1

1.1 Q1

$\Pr[X \geq t] \Leftrightarrow \Pr[e^{\lambda X} \geq e^{\lambda t}]$, for $\lambda \geq 0$. Thus, the question tranformed to the equation on the right hand side.

$$\begin{aligned}\Pr[X \geq t] &= \Pr[e^{\lambda X} \geq e^{\lambda t}] \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}} \quad (\text{Markov Inequality}) \\ &= \exp(-(\lambda t - \ln \mathbb{E}[e^{\lambda X}])) \\ &= \exp(-(\lambda t - \Psi_X(\lambda)))\end{aligned} \tag{1}$$

We have $\exp(-(\lambda t - \Psi_X(\lambda))) \geq \exp(-\Psi_X^*(\lambda))$, thus:

$$\Pr[X \geq t] = \exp(-\Psi_X^*(\lambda))$$

For $F(\lambda) = \lambda t - \Psi_X(\lambda)$, $\lambda \geq 0$. If $\Psi_X(\lambda)$ is continuously differentiable, we can perform standard analysis of this function, taking gradient of both sides:

$$\nabla_{\lambda} F(\lambda) = t - \nabla_{\lambda} \Psi_X(\lambda)$$

Let the gradient equal to 0, we can find that the unique $\lambda \geq 0$ satisfying $\Psi_X'(\lambda) = t$, according to the convexity of $\Psi_X(\lambda)$, we obtain that:

$$\Psi_X^*(t) = F(\lambda)|_{\Psi_X(\lambda)=t} = \sup_{\lambda \geq 0} (\lambda t - \Psi_X(\lambda))$$

1.2 Q2

Gaussian random variable \mathbf{X} , it's probability density function is given by $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, so we have:

$$\begin{aligned}
\Psi_X(\lambda) &= \ln \mathbb{E} [e^{\lambda X}] \\
&= \ln \int_{-\infty}^{\infty} e^{\lambda x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\lambda x - \frac{(x-\mu)^2}{2\sigma^2}} dx \right] \\
&= \ln \left[\frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(x-(\lambda\sigma^2+\mu))^2 - (\lambda^2\sigma^4 + 2\lambda\mu\sigma^2)}{2\sigma^2}} dx \right] \\
&= \ln \left[e^{\lambda\mu + \frac{\lambda^2\sigma^2}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(x-(\lambda\sigma^2+\mu))^2}{2\sigma^2}} dx \right] \\
&= \ln \left[e^{\lambda\mu + \frac{\lambda^2\sigma^2}{2}} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-t^2} d(\sqrt{2\sigma}t) \right] \quad \left(\text{let } t = \frac{x - (\lambda\sigma^2 + \mu)}{\sqrt{2}\sigma} \right) \\
&= \ln \left[e^{\lambda\mu + \frac{\lambda^2\sigma^2}{2}} \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt \right] \\
&= \ln \left[e^{\lambda\mu + \frac{\lambda^2\sigma^2}{2}} \right] \\
&= \lambda\mu + \frac{\lambda^2\sigma^2}{2}
\end{aligned} \tag{2}$$

Thus, we can calculate $\Psi_X^*(t)$

$$\begin{aligned}
\Psi_X^*(t) &= \sup_{\lambda \geq 0} (\lambda t - \Psi_X(\lambda)) \\
&= \sup_{\lambda \geq 0} \left(\lambda t - \lambda\mu - \frac{\lambda^2\sigma^2}{2} \right) \\
&= \frac{\lambda^2\sigma^2}{2}
\end{aligned} \tag{3}$$

Now, the upper tail can be bounded:

$$\Pr[X \geq t] \leq \exp \left(-\frac{\lambda^2\sigma^2}{2} \right) \tag{4}$$

1.3 Q3

Poisson random variable \mathbf{X} , it's probability distribution is given by $\Pr[X = k] = e^{-\nu} \frac{\nu^k}{k!}$ We have:

$$\begin{aligned}
\Psi_X(\lambda) &= \ln \mathbb{E} [e^{\lambda X}] \\
&= \ln \sum_{k=0}^{\infty} \Pr [X = k] e^{\lambda k} \\
&= \ln \sum_{k=0}^{\infty} e^{\lambda k - \nu} \frac{\nu^k}{k!} \\
&= \ln \left[e^{-\nu} \sum_{k=0}^{\infty} e^{\lambda k} \frac{\nu^k}{k!} \right] \\
&= \ln \left[e^{-\nu} \sum_{k=0}^{\infty} \frac{(e^{\lambda} \nu)^k}{k!} \right] \\
&= \ln [e^{-\nu} e^{e^{\lambda} \nu}] \\
&= (e^{\lambda} - 1) \nu
\end{aligned} \tag{5}$$

Then, we get $\Psi_{X^*}(t)$:

$$\begin{aligned}
\Psi_{X^*}(t) &= \sup_{\lambda \geq 0} (\lambda t - \Psi_X(\lambda)) \\
&= \sup_{\lambda \geq 0} (\lambda t - (e^{\lambda} - 1) \nu) \\
&= \lambda e^{\lambda} \nu - (e^{\lambda} - 1) \nu \\
&= ((\lambda - 1) e^{\lambda} + 1) \nu
\end{aligned} \tag{6}$$

According to Q_1 , we have:

$$\Pr [X \geq t] \leq \exp (-((\lambda - 1) e^{\lambda} + 1) \nu) \tag{7}$$

1.4 Q4

Bernoulli random variable \mathbf{X} , its probability distribution is given by $\Pr [X = 1] = 1 - \Pr [X = 0] = p$, thus we have:

$$\begin{aligned}
\Psi_X(\lambda) &= \ln \mathbb{E} [e^{\lambda X}] \\
&= \ln [p e^{\lambda} + (1 - p)]
\end{aligned} \tag{8}$$

Then, we get:

$$\begin{aligned}\Psi_X^*(t) &= \sup_{\lambda \geq 0} (\lambda t - \Psi_X(\lambda)) \\ &= \sup_{\lambda \geq 0} (\lambda t - \ln [pe^\lambda + (1-p)])\end{aligned}\tag{9}$$

For the equation above, taking derivative w.r.t λ , we have:

$$t = \frac{e^\lambda p}{e^\lambda + 1 - p}$$

We may solve λ :

$$\lambda = \ln \left[\frac{(1-p)t}{(1-t)p} \right]$$

Thus, we may combining with equation 9:

$$\begin{aligned}\Psi_X^*(t) &= \ln \left[\frac{(1-p)t}{(1-t)p} \right] t - \ln \left[p \frac{(1-p)t}{(1-t)p} + (1-p) \right] \\ &= (1-t) \ln \frac{1-t}{1-p} + t \ln \frac{t}{p}\end{aligned}\tag{10}$$

1.5 Q5

As X_1, X_2, \dots, X_n are i.i.d random variables, we have:

$$\begin{aligned}\Psi_X(\lambda) &= \ln \mathbb{E} [e^{\lambda X}] \\ &= \ln \mathbb{E} \left[e^{\lambda \sum_{i=1}^n X_i} \right] \\ &= \ln \prod_{i=1}^n \mathbb{E} [e^{\lambda X_i}] \\ &= \sum_{i=1}^n \ln \mathbb{E} [e^{\lambda X_i}] \\ &= \sum_{i=1}^n \Psi_{X_i}(\lambda)\end{aligned}\tag{11}$$

Similarly, for $\Psi_X^*(t)$, we have:

$$\begin{aligned}
\Psi_X^*(t) &= \sup_{\lambda \geq 0} (\lambda t - \Psi_X(\lambda)) \\
&= \sup_{\lambda \geq 0} \left(\lambda t - \sum_{i=1}^n \Psi_{X_i}(\lambda) \right) \\
&= \sum_{i=1}^n \sup_{\lambda \geq 0} \left(\lambda \frac{t}{n} - \Psi_{X_i}(\lambda) \right) \\
&= \sum_{i=1}^n \Psi_{X_i}^*\left(\frac{t}{n}\right) \\
&= n \Psi_{X_i}^*\left(\frac{t}{n}\right) \quad (i.i.d)
\end{aligned} \tag{12}$$

For Binomial random variable $X \sim \text{Bin}(n, p)$, it can be decomposed as sum of n i.i.d random Bernoulli random variables X_1, X_2, \dots, X_n . According to what we have above, the upper bound can be measured:

$$\begin{aligned}
\Pr[X \geq t] &\leq \exp(-\Psi_X^*(t)) \\
&= \exp\left(-n \Psi_{X_i}^*\left(\frac{t}{n}\right)\right) \\
&= \exp(-n D(Y \| X_i))
\end{aligned} \tag{13}$$

Where $Y \in \{0, 1\}$ is a Bernoulli random variable with parameter $\frac{t}{n}$.

Given geometric random variable \mathbf{X} with distribution $\Pr[X = k] = (1 - p)^{k-1}p$.

$$\begin{aligned}
\Psi_X(\lambda) &= \ln \mathbb{E}[e^{\lambda X}] \\
&= \ln \sum_{k=1}^{\infty} \Pr[X = k] e^{\lambda k} \\
&= \ln \sum_{k=1}^{\infty} e^{\lambda k} (1 - p)^{k-1} p \\
&= \ln \frac{e^{\lambda} p}{e^{\lambda}(p - 1) + 1}
\end{aligned} \tag{14}$$

$$\begin{aligned}
\Psi_X^*(t) &= \sup_{\lambda \geq 0} (\lambda t - \Psi_X(\lambda)) \\
&= \sup_{\lambda \geq 0} \left(\lambda t - \ln \frac{e^\lambda p}{e^\lambda (p-1) + 1} \right) \\
&= t \ln \left(\frac{1-t}{(p-1)t} \right) - \ln \left(-\frac{p(t-1)}{p-1} \right)
\end{aligned} \tag{15}$$

Combining all of them together:

$$\begin{aligned}
\Pr[X \geq t] &\leq \exp(-\Psi_X^*(t)) \\
&= \exp \left(-n \Psi_{X_i}^* \left(\frac{t}{n} \right) \right) \\
&= \exp \left(-t \ln \left(\frac{n-t}{(p-1)t} \right) + n \ln \left(-\frac{p(t-n)}{n(p-1)} \right) \right)
\end{aligned} \tag{16}$$

2 Problem 2

We define any vertex $u \in V(Q_n)$ as n independent random variables (X_1, X_2, \dots, X_n) where $X_i \in \{0, 1\}, i = 1, 2, \dots, n$.

Next, we can prove the function $f(X_1, X_2, \dots, X_n)$ satisfying the Lipschitz condition. For any X_1, X_2, \dots, X_n and any $Y_i \in \{0, 1\}, i = 1, 2, \dots, n$. We denote vertex \mathbf{u} as $(X_1, X_2, \dots, X_{i-1}, X_i, \dots, X_n)$, and vertex \mathbf{w} as $(X_1, X_2, \dots, X_{i-1}, Y_i, \dots, X_n)$. According to the definition of shortest distance, we know that there's an edge between vertex \mathbf{u} and vertex \mathbf{w} .

Thus, we have:

$$1 + f(\mathbf{u}) \leq f(\mathbf{w})$$

Symmetrically, we have:

$$1 + f(\mathbf{w}) \leq f(\mathbf{u})$$

Combining them, we have:

$$|f(\mathbf{u}) - f(\mathbf{w})| \leq 1 \tag{17}$$

Now that f satisfying Lipschitz condition with constants 1, we can apply

Method of bounded differences:

$$\begin{aligned}
\Pr \left[|f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})]| \geq t\sqrt{n \log n} \right] &\leq 2 \exp \left(-\frac{t^2 n \log n}{2 \sum_{i=1}^n 1^2} \right) \\
&= 2 \exp \left(-\frac{t^2 \log n}{2} \right) \\
&= 2n^{-\frac{t^2}{2}} \\
&= n^{\log_n 2 - \frac{t^2}{2}}
\end{aligned} \tag{18}$$

Let $\log_n 2 - \frac{t^2}{2} = -c$, we have: $c = \frac{t^2}{2} - \log_n 2$.

3 Problem 3

3.1 Q1

It's obvious that $\forall 1 \leq i \leq n, \Pr[Y_i = 1] \leq p$, we can let $\Pr[Y_i = 1] = t_i$, where $t_i \leq p, i = 1, 2, \dots, n$. Thus, $\forall 1 \leq i \leq n$, we generate A as a uniform random variable between $[0, 1]$, we can construct the coupling \mathcal{C} as below:

$$\begin{aligned}
Y_i &= \begin{cases} 1, & \text{if } A \leq t_i \\ 0, & \text{otherwise} \end{cases} \\
X_i &= \begin{cases} 1, & \text{if } A \leq p \\ 0, & \text{otherwise} \end{cases}
\end{aligned}$$

We can easily verify that $\forall 1 \leq i \leq n, Y_i \leq X_i$, formally:

$$\Pr_{\mathcal{C}} [\forall 1 \leq i \leq n, Y_i \leq X_i] = 1$$

According to stochastic dominance, we have:

$$\forall a > 0, \Pr \left[\sum_{i=1}^n Y_i \geq a \right] \leq \Pr \left[\sum_{i=1}^n X_i \leq a \right]$$

3.2 Q2

Now that X_1, X_2, \dots, X_n are mutually independent, by linearity of expectation, it holds that $\mathbb{E}[\sum_{i=1}^n X_i] = np$.

$$\begin{aligned}
 \Pr \left[\sum_{i=1}^n Y_i \geq np + t \right] &\leq \Pr \left[\sum_{i=1}^n X_i \geq np + t \right] \\
 &= \Pr \left[\sum_{i=1}^n X_i - \mathbb{E} \left[\sum_{i=1}^n X_i \right] \geq t \right] \\
 &\leq \exp \left(-\frac{2t^2}{n} \right) \quad (\text{Chernoff Bound})
 \end{aligned} \tag{19}$$

4 Problem 4

4.1 Q1

The origin triangle inequality can be rewritten:

$$\begin{aligned}
 d(A, B) + d(B, C) &\geq d(A, C) \\
 \Leftrightarrow 1 - \text{sim}(A, B) + 1 - \text{sim}(B, C) &\geq 1 - \text{sim}(A, C) \\
 \Leftrightarrow \text{sim}(A, B) + \text{sim}(B, C) - \text{sim}(A, C) &\leq 1 \\
 \Leftrightarrow \Pr_{h \in \mathcal{F}} [h(A) = h(B)] + \Pr_{h \in \mathcal{F}} [h(B) = h(C)] - \Pr_{h \in \mathcal{F}} [h(A) = h(C)] &\leq 1 \\
 \Leftrightarrow \Pr_{h \in \mathcal{F}} [h(A) = h(B)] + \Pr_{h \in \mathcal{F}} [h(B) = h(C)] - \Pr_{h \in \mathcal{F}} [(h(A) = h(B)) \wedge (h(B) = h(C))] &\leq 1 \\
 \Leftrightarrow \Pr_{h \in \mathcal{F}} [(h(A) = h(B)) \vee (h(B) = h(C))] &\leq 1
 \end{aligned} \tag{20}$$

It's obvious that probability value is less or equal than 1.

4.2 Q2

Assume that we have locality sensitive hash function family corresponding to Dice's coefficient, by triangle inequality we have:

$$\begin{aligned} \forall A, B, C \in 2^U, d(A, B) + d(B, C) &\geq d(A, C) \\ \Leftrightarrow \frac{|A| + |B| - 2|A \cap B|}{|A| + |B|} + \frac{|B| + |C| - 2|B \cap C|}{|B| + |C|} &\geq \frac{|A| + |C| - 2|A \cap C|}{|A| + |C|} \end{aligned} \quad (21)$$

Let $|A| = |C| = \frac{|U|}{2}$, $A = U - C$, $B = U$, we get:

$$\frac{\frac{|U|}{2}}{\frac{3|U|}{2}} + \frac{\frac{|U|}{2}}{\frac{3|U|}{2}} \geq 1 \Leftrightarrow \frac{2}{3} \geq 1 \quad (\text{Contradiction})$$

Thus, we can prove that no locality sensitive hash function family corresponding to Dice's coefficient.

Accordingly, let $|A| = |C| = \frac{|U|}{2}$, $A = U - C$, $B = U$, we have:

$$d(A, B) + d(B, C) = 1 - \text{sim}_{Ovl}(A, B) + 1 - \text{sim}_{Ovl}(B, C) = 1 - 1 + 1 - 1 = 0$$

$$d(A, C) = 1 - \text{sim}_{Ovl}(A, C) = 1 - 0 = 1$$

Thus, we have:

$$d(A, B) + d(B, C) \leq d(A, C)$$

We get contradiction, thus we proved that no locality sensitive hash function family corresponding to Overlap's coefficient.

4.3 Q3

Assuming we have $A, B \in \{0, 1\}^m$:

$$\begin{aligned} \Pr_{h' \in \mathcal{F}'} [h'(A) = h'(B)] &= \Pr_{h \in \mathcal{F}} [h(A) = h(B)] \cdot \Pr_{f \in \mathcal{B}} [f(h(A)) = f(h(B))] \\ &\quad + \Pr_{h \in \mathcal{F}} [h(A) \neq h(B)] \cdot \Pr_{f \in \mathcal{B}} [f(h(A)) = f(h(B))] \\ &= \text{sim}(A, B) \cdot 1 + (1 - \text{sim}(A, B)) \cdot \frac{1}{2} \\ &= \frac{1 + \text{sim}(A, B)}{2} \end{aligned} \quad (22)$$