Problem Set 2 Solutions

姓名: 黄睿
学号: MP1933008

2019-10-23

1 Problem 1

1.1 Q1

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

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

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

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

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:

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

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:

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

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

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

Now, the upper tail can be bounded:

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

1.3 Q3

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

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

Then, we get $\Psi_X*(t)$:

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

According to Q_1, we have:

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

1.4 Q4

Bernoulli ranndom variable $\mathbf{X}$, it's probability distribution is given by $\Pr\left[X = 1\right] = 1 - \Pr\left[X = 0\right] = p$, thus we have:

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

Then, we get:

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

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 combing 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 $X_1, X_2, \cdots, X_n$ are i.i.d random variables, we have:

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

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

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

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

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

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

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

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

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

Combining all of them together:

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

2 Problem 2

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

Next, we can prove the function $f(X_1, X_2, \cdots, X_n)$ satisfying the Lipschitz condition. For any $X_1, X_2, \cdots, X_n$ and any $Y_i \in \{0, 1\}, i = 1, 2, \cdots, n$. We denote vertex $\mathbf{u}$ as $(X_1, X_2, \cdots, X_{i-1}, X_i, \cdots, X_n)$, and vertex $\mathbf{w}$ as $(X_1, X_2, \cdots, X_{i-1}, Y_i, \cdots, X_n)$.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:

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

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 $\forall 1 \leq i \leq n, \Pr\left[Y_i = 1\right] \leq p$, we can let $\Pr\left[Y_i = 1\right] = t_i$, where $t_i \leq p, i = 1, 2, \cdots, n$.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:

$$Y_i = \begin{cases} 1, if \ A \leq t_i \\ 0, otherwise \end{cases}$$

$$X_i = \begin{cases} 1, if \ A \leq p \\ 0, otherwise \end{cases}$$

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

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

According to stochastic dominance, we have:

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

3.2 Q2

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

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

4 Problem 4

4.1 Q1

The origin triangle inequality can be rewrited:

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

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:

$$\forall A, B, C \in 2^U, d(A,B) + d(B,C) \geq d(A,C) \tag{21}$$

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

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

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:

$$d(A,B) + d(B,C) = 1 - sim_{Ovl}(A,B) + 1 - sim_{Ovl}(B,C) = 1 - 1 + 1 - 1 = 0$$

$$d(A,C) = 1 - sim_{Ovl}(A,C) = 1 - 0 = 1$$

Thus, we have:

$$d(A,B) + d(B,C) < d(A,C) \quad (Contradiction)$$

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

4.3 Q3

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

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