SDS 384-11 PS #2, Spring 2022

Instructor: Purnamrita Sarkar*

Thursday 31st March, 2022

Exercise 1 (4 pts)**.** *Remember Hoeffding's Lemma? We proved it with a weaker constant in class using a symmetrization type argument. Now we will prove the original version. Let X be a bounded r.v. in $[a, b]$ such that $E[X] = \mu$. Let $f(\lambda) = \log E[e^{\lambda(X-\mu)}]$. Show that $f''(\lambda) \leq (b-a)^2/4$. Now use the fundamental theorem of calculus to write $f(\lambda)$ in terms of $f''(\lambda)$ and finish the argument.*

Solution.
First we find $f''(\lambda)$

$$
\begin{aligned}
f''(\lambda) &= \frac{d}{d\lambda}\frac{d}{d\lambda}\log E[e^{\lambda(X-\mu)}] \\
&= \frac{d}{d\lambda}\left[\frac{1}{E[e^{\lambda(X-\mu)}]}E[(X-\mu)e^{\lambda(X-\mu)}]\right] \\
&= \frac{E[(X-\mu)^2 e^{\lambda(X-\mu)}]}{E[e^{\lambda(X-\mu)}]} - \left(\frac{E[(X-\mu)e^{\lambda(X-\mu)}]}{E[e^{\lambda(X-\mu)}]}\right)^2.
\end{aligned}
$$

Notice that $f''(\lambda)$ is simply the variance of $X-\mu$ with respect to the probability measure $\frac{e^{\lambda(X-\mu)}}{E[e^{\lambda(X-\mu)}]}P(X)$. We know that $E[(X-t)^2]$ is minimized when $t = E[X]$, thus

$$
\begin{aligned}
f''(\lambda) &= \mathrm{var}(X) \\
&\leq E[(X-(b-a)/2)^2] \\
&\leq \frac{(b-a)^2}{4}.
\end{aligned}
$$

Now we will use the fundamental theorem of calculus to show that $f(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}$. Note from the derivations above $f'(0) = f(0) = 0$

$$
\begin{aligned}
\int_0^\lambda \int_0^\nu f''(t)dtd\nu &\leq \int_0^\lambda \int_0^\nu \frac{(b-a)^2}{4}dtd\nu \\
\int_0^\lambda f'(\nu)d\nu &\leq \int_0^\lambda \frac{\nu(b-a)^2}{4}d\nu \\
f(\lambda) &\leq \frac{\lambda^2(b-a)^2}{8}
\end{aligned}
$$

Taking the exponent of both sides gives back Hoeffding's Lemma. □

*Thanks to Matthew Faw, Brandon Carter for some of the solutions

Exercise 2. *Consider a r.v. X such that for all $\lambda \in \Re$*

$$M(\lambda) \triangleq E[e^{\lambda X}] \leq e^{\frac{\lambda^2\sigma^2}{2}+\lambda\mu}$$

If you find anyone saying f(x)<g(x) implies f'(x)<g'(x) Take 2 points off from 5. Because that is a serious mistake and point to solutions.

Prove that:

1. $\mathbb{E}[X] = \mu$.

Solution.
Let $f(\lambda) = E[e^{\lambda X}]$ and let $g(\lambda) = e^{\lambda^2\sigma^2/2+\lambda\mu}$. We have $f(0) = g(0)$.

$$f'(0) = \lim_{h\to 0} \frac{f(h) - f(0)}{h} \leq \lim_{h\to 0} \frac{g(h) - g(0)}{h} = g'(0)$$

But we also have:

$$f'(0) = \lim_{h\to 0} \frac{f(0) - f(-h)}{h} \geq \lim_{h\to 0} \frac{g(0) - g(-h)}{h} = g'(0)$$

So $f'(0) = g'(0)$. So we have $E[X] = \mu$.

□

2. $\mathrm{Var}(X) \leq \sigma^2$.

Solution.

Let us denote

$$\begin{aligned} M_c(\lambda) &= \exp(-\lambda\mu)M(\lambda) \\ &= \mathbb{E}[\exp(\lambda(X-\mu))] \end{aligned}$$

and similarly,

$$\begin{aligned} U_c(\lambda) &= \exp(-\lambda\mu)U(\lambda) \\ &= \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \end{aligned}$$

Then, by construction, we have that $M_c(\lambda) \leq U_c(\lambda)$. Additionally, $M_c(0) = 1 = U_c(0)$, $M_c''(0) = \mathrm{Var}(X)$, and $U_c''(0) = \sigma^2$. Therefore, we have that

$$\begin{aligned} \mathrm{Var}(X) &= M_c''(0) \\ &= \lim_{\varepsilon\to 0} \frac{M_c(\varepsilon) + M_c(-\varepsilon) - 2M_c(0)}{\varepsilon^2} \\ &= \lim_{\varepsilon\to 0} \frac{M_c(\varepsilon) + M_c(-\varepsilon) - 2U_c(0)}{\varepsilon^2} \\ &\leq \lim_{\varepsilon\to 0} \frac{U_c(\varepsilon) + U_c(-\varepsilon) - 2U_c(0)}{\varepsilon^2} \\ &= U_c''(0) \\ &= \sigma^2 \end{aligned}$$

which establishes the desired inequality.

□

3. *If the smallest value of σ satisfying the above equation is chosen, is it true that* $\mathrm{Var}(X) = \sigma^2$*? Prove or give a counter example.*

Solution.
We give a counterexample to establish that $\sigma^2 \neq \mathrm{Var}(X)$. Consider $X \sim \mathrm{Bern}(p)$. Then, assuming that $\sigma^2 = p(1-p) = \mathrm{Var}(X)$, we have that

$$\begin{aligned} \mathbb{E}\left[\exp\left(\lambda\left(X-p\right)\right)\right] &= p\exp(\lambda(1-p)) + (1-p)\exp(-\lambda p) \\ &= \exp(\lambda(1-p))\left(p + (1-p)\exp(-\lambda)\right) \\ &\leq \exp\left(\frac{\lambda^2 p(1-p)}{2}\right) \qquad \text{by assumed subG bound} \\ \implies p + (1-p)\exp(-\lambda) &\leq \exp\left(\lambda(1-p)\left(\frac{\lambda p}{2} - 1\right)\right) \qquad (1) \end{aligned}$$

However, by choosing, for example, $\lambda = \frac{1}{4}$ and $p = \frac{1}{16}$, one can check that

$$p + (1-p)\exp(-\lambda) - \exp\left(\lambda(1-p)\left(\frac{\lambda p}{2} - 1\right)\right) \approx 0.0001 > 0$$

which is a *contradiction* of inequality (1). Therefore, we cannot always take $\sigma^2 = \mathrm{Var}(X)$.

□

Exercise 3. *Given a positive semidefinite matrix $Q \in \mathbb{R}^{n\times n}$, consider $Z = \sum_{i,j} Q_{ij} X_i X_j$. When $X_i \sim N(0,1)$, prove the Hanson-Wright inequality.*

$$P\left(Z \geq \mathit{trace}(Q) + t\right) \leq \exp\left(-\min\left\{c_1 t/\|Q\|_{op}, c_2 t^2/\|Q\|_F^2\right\}\right),$$

where $\|Q\|_{op}$ and $\|Q\|_F$ denote the operator and frobenius norms respectively. Hint: The rotation-invariance of the Gaussian distribution and sub-exponential nature of χ^2-variables could be useful.

Solution.
Observe that since Q is positive semidefinite, by the spectral theorem, we may decompose Q as

$$Q = V\Lambda V^T,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_i \geq 0$ is the ith largest eigenvalue of Q, and where V is an orthogonal matrix whose ith column vector is a unit eigenvector corresponding to λ_i.

Take $X \sim \mathcal{N}(\mathbf{0}, I_n)$, where I_n is the $n \times n$ identity matrix. Thus, by the linear transformation property of the normal distribution, we have that $Y = V^T X \sim \mathcal{N}(\mathbf{0}, \underbrace{V^T V}_{=I_n \text{ since } V \text{ is orthogonal}})$.

Therefore, we have that

$$\begin{aligned} Z &= X^T Q X \\ &= X^T V \Lambda V^T X \\ &= Y^T \Lambda Y \\ &= \sum_{i=1}^{n} \lambda_i Y_i^2 \end{aligned}$$

Now, each $Y_i^2 \overset{i.i.d}{\sim} \chi_1^2$. Thus,

$$\begin{aligned} \mathbb{E}[Z] &= \sum_{i=1}^{n} \lambda_i \underbrace{\mathbb{E}[Y_i^2]}_{=1} \\ &= \sum_{i=1}^{n} \lambda_i \\ &= \text{trace}(Q) \end{aligned}$$

Now, following the same arguments used in class to obtain the subexponential parameters for χ^2 random variables, we have that $Z_i = \lambda_i Y_i^2$ is $(\nu_i = 2\lambda_i, b_i = 4\lambda_i)$-subexponential, and thus, by concentration of subexponential r.v.'s, we have that

$$\mathbb{P}(Z_i - \mathbb{E}[Z_i] \geq t) \leq \begin{cases} \exp\left(-\frac{t^2}{8\lambda_i^2}\right) & \text{if } 0 \leq t \leq \frac{4\lambda_i^2}{4\lambda_i} = \lambda_i \\ \exp\left(-\frac{t}{8\lambda_i}\right) & \text{if } t > \lambda_i \end{cases}$$

Now, we have that Z is $(\sqrt{n}\nu_*, b_*)$-subexponential, where

$$\begin{aligned} \nu_*^2 &= \frac{1}{n}\sum_i \nu_i^2 = \frac{4}{n}\sum_i \lambda_i^2 = \frac{4}{n}\|Q\|_F^2 \\ b_* &= \max_k b_k = 4\lambda_1 = 4\|Q\|_{op} \end{aligned}$$

and so

$$\mathbb{P}(Z - \text{trace}(Q) \geq t) \leq \begin{cases} \exp\left(-\frac{t^2}{8\|Q\|_F^2}\right) & \text{if } 0 \leq t \leq \frac{\|Q\|_F^2}{\|Q\|_{op}} \\ \exp\left(-\frac{t}{8\|Q\|_{op}}\right) & \text{if } t > \frac{\|Q\|_F^2}{\|Q\|_{op}} \end{cases}$$

as desired. □

Exercise 4. *We will prove properties of subgaussian random variables here. Prove that:*

1. *Moments of a mean zero subgaussian r.v. X with variance proxy σ^2 satisfy:*

$$E[|X^k|] \leq k 2^{k/2} \sigma^k \Gamma(k/2), \tag{2}$$

where Γ is the gamma function.

Solution.
We have that, by the Subgaussian assumption,

$$\begin{aligned}\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}(|X|^k > t)dt \\ &= \int_0^\infty \mathbb{P}(|X| > t^{1/k})dt \\ &\leq 2\int_0^\infty \exp\left(-\frac{t^2}{2\sigma^2}\right)dt\end{aligned}$$

Now, recalling that

$$\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)dt,$$

we may perform the change of variables $t = ax^b$ to obtain:

$$\begin{aligned}\Gamma(z) &= \int_0^\infty a^{z-1}x^{bz-b}\exp(-ax^b)abx^{b-1}dx \\ &= a^z b\int_0^\infty x^{bz-1}\exp(-ax^b)dx\end{aligned}$$

Thus,

$$\Gamma\left(\frac{1}{b}\right) = a^{\frac{1}{b}}b\int_0^\infty \exp(-ax^b)dx$$

Now, choosing $a = \frac{1}{2\sigma^2}$ and $b = \frac{2}{k}$, we combine our results to obtain:

$$\begin{aligned}\mathbb{E}[|X|^k] &\leq 2\int_0^\infty \exp\left(-\frac{t^2}{2\sigma^2}\right)dt \\ &= \frac{2}{a^{\frac{1}{b}}b}\Gamma\left(\frac{1}{b}\right) \\ &= \frac{2}{\left(\frac{1}{2\sigma^2}\right)^{\frac{k}{2}}\frac{2}{k}}\Gamma\left(\frac{k}{2}\right) \\ &= k2^{k/2}\sigma^k\Gamma\left(\frac{k}{2}\right)\end{aligned}$$

as desired. ∎

2. *If X is a mean 0 subgaussian r.v. with variance proxy σ^2, prove that, $X^2 - E[X^2]$ is a subexponential $(c_1\sigma^2, c_2\sigma^2)$ (we are using the (ν, b) parametrization of subexponentials we did in class, so ν^2 is the variance proxy). Here c_1, c_2 are positive constants.*

I am going to give two different solutions here. And point out common mistakes you may make. The first uses Bernstein's moment condition. In class we did a

very easy bounded random variable example to show it is subexponential since it satisfies the bernstein m.c. Here is a far less trivial example of its use. The second solution gets to the answer through the definition of sub-exp r.v.s as we saw in class.

Solution.
Here, we wish to apply the Bernstein condition. Observe that

$$\begin{aligned}
&\left|\mathbb{E}(X^2 - \mathbb{E}X^2)^k\right| \\
&\leq \mathbb{E}\left|X^2 - \mathbb{E}X^2\right|^k && \text{Jensen's} \\
&= \mathbb{E}\left(\left|X^2 - \mathbb{E}X^2\right|^k \mathbb{1}\{X^2 \geq \mathbb{E}X^2\}\right) + \mathbb{E}\left(\left|X^2 - \mathbb{E}X^2\right|^k \mathbb{1}\{X^2 < \mathbb{E}X^2\}\right)
\end{aligned}$$

Now, observe that, almost surely,

$$\begin{aligned}
|X^2 - \mathbb{E}X^2|^k \mathbb{1}\{X^2 \geq \mathbb{E}X^2\} &\leq |X^2|^k \mathbb{1}\{X^2 \geq \mathbb{E}X^2\} && \text{since } \mathbb{E}X^2 \geq 0 \\
&\leq |X|^{2k} && \text{since } \mathbb{1}\{\cdot\} \leq 1 \text{ a.s.}
\end{aligned}$$

and similarly,

$$\begin{aligned}
|X^2 - \mathbb{E}X^2|^k \mathbb{1}\{X^2 < \mathbb{E}X^2\} &\leq |\mathbb{E}X^2|^k \mathbb{1}\{X^2 < \mathbb{E}X^2\} && \text{since } X^2 \geq 0 \text{ a.s.} \\
&\leq |\mathbb{E}X^2|^k && \text{since } \mathbb{1}\{\cdot\} \leq 1 \text{ a.s.} \\
&\leq \mathbb{E}|X|^{2k} && \text{by Jensen's, since } |\cdot|^k \text{ is convex}
\end{aligned}$$

Note the treatment above. Many of you may bound $E[(X^2 - [EX]^2)^k] \leq E[X^{2k}]$. This is incorrect, because $|X^2 - E[X^2]| \leq \max(X^2, E[X^2])$. I am gong to take a point off for this, just so that this sticks in our minds.

Finally, note that

$$\begin{aligned}
\mathrm{Var}(X^2) &\leq \mathbb{E}X^4 \\
&\leq 4 \cdot 2^2 \sigma^4 \Gamma(2) \\
&= 2^4 \sigma^4 \\
&< 2^5 \sigma^4
\end{aligned}$$

Combining these bounds, we have that

$$\begin{aligned}
\left|\mathbb{E}(X^2 - \mathbb{E}X^2)^k\right| &\leq 2\mathbb{E}|X|^{2k} \\
&\leq 4k 2^k \sigma^{2k} \underbrace{\Gamma(k)}_{=(k-1)!} && \text{by the previous exercise} \\
&= \frac{1}{2} k! 2^5 \sigma^4 \left(2\sigma^2\right)^{k-2}
\end{aligned}$$

Therefore, by the Bernstein condition, we have that X^2 is subexponential with paratmeters $(\nu = 8\sigma^2, b = 4\sigma^2)$, as desired.

$\square$

Solution.
Now we will prove the subexponentiality using the MGF. Note that we have $E[X^2] \le \sigma^2$.

$$
\begin{aligned}
E[\exp(\lambda(X^2 - E[X^2]))] &\le \exp(-\lambda E[X^2]) E[\exp(\lambda X^2)] \\
&= \exp(-\lambda E[X^2]) \left(1 + \lambda E[X^2] + \sum_{k\ge 2} \lambda^k \frac{E[X^{2k}]}{k!}\right) \\
&= \exp(-\lambda E[X^2]) \left(1 + \lambda E[X^2] + 2\sum_{k\ge 2} 2^k \sigma^{2k} |\lambda|^k\right) \\
\text{(For } |\lambda| < 1/2\sigma^2\text{, we have)} &= \underbrace{\exp(-\lambda E[X^2])\left(1 + \lambda E[X^2]\right)}_{\text{This is } \le 1 \text{ since } \exp(x)\ge 1+x} + \frac{8\sigma^4\lambda^2 \exp(-\lambda E[X^2])}{1 - 2\sigma^2|\lambda|} \\
\text{(For } |\lambda| < 1/4\sigma^2\text{, we have)} &\le 1 + \underbrace{16\sigma^4\lambda^2 \exp(|\lambda|\sigma^2)}_{E[X^2]\le\sigma^2} \le 1 + \underbrace{16\sigma^4\lambda^2\exp(1/4)}_{|\lambda|\le 1/4\sigma^2} \\
&\le 1 + 2^5\sigma^4\lambda^2 \le \underbrace{\exp((4\sqrt{2}\sigma^2)^2\lambda^2)}_{\exp(x)\ge 1+x}
\end{aligned}
$$

So we have $X^2 - E[X^2]$ is sub exponential $(8\sigma^2, 4\sigma^2)$.

$\square$

3. *Consider two independent mean zero subgaussian r.v.s X_1 and X_2 with variance proxies σ_1^2 and σ_2^2 respectively. Show that X_1X_2 is a subexponential r.v. with parameters $(d_1\sigma_1\sigma_2, d_2\sigma_1\sigma_2)$. Here d_1, d_2 are positive constants.*

Solution.
Observe that,

$$
\begin{aligned}
\mathbb{E}[(X_1X_2 - \mathbb{E}[X_1X_2])^k] &= \mathbb{E}[(X_1X_2 - \mathbb{E}[X_1]\mathbb{E}[X_2])^k] && \text{by independence} \\
&= \mathbb{E}[(X_1X_2)^k] && \text{mean } 0 \\
&\le \mathbb{E}[|X_1X_2|^k] \\
&= \mathbb{E}[|X_1|^k]\mathbb{E}[|X_2|^k] && \text{independence} \\
&\le \left(k2^{k/2}\sigma_1^k\Gamma\left(\frac{k}{2}\right)\right)\left(k2^{k/2}\sigma_2^k\Gamma\left(\frac{k}{2}\right)\right) && \text{by part 1} \\
&= \left(k\Gamma\left(\frac{k}{2}\right)\right)^2 2^k(\sigma_1\sigma_2)^2
\end{aligned}
$$

Now, recall that, for k an odd integer,

$$\begin{aligned}\Gamma\left(\frac{k}{2}\right) &= \Gamma\left(\left\lfloor\frac{k}{2}\right\rfloor + \frac{1}{2}\right)\\ &= \sqrt{\pi}\frac{\left(2\left\lfloor\frac{k}{2}\right\rfloor\right)!}{4^{\lfloor k/2\rfloor}\lfloor k/2\rfloor!}\\ &= \sqrt{\pi}\frac{2\,(k-1)!}{4^{k/2}\lfloor k/2\rfloor!}\end{aligned}$$

Thus, we have that

$$\begin{aligned}\left(k\Gamma\left(\frac{k}{2}\right)\right)^2 &= \pi k^2\frac{((k-1)!)^2}{4^k\left(\lfloor k/2\rfloor\right)^2} && (3)\\ &\leq k!\\ \iff \pi k! &\leq 4^k\lfloor k/2\rfloor! && (4)\end{aligned}$$

Now, note that (3) is true for sufficiently large k. Similarly, when k is even,

$$\Gamma\left(\frac{k}{2}\right) = \left(\frac{k}{2}-1\right)!$$

so we have that

$$\begin{aligned}\left(k\Gamma\left(\frac{k}{2}\right)\right)^2 &= k^2\left(\left(\frac{k}{2}-1\right)!\right)^2\\ &\leq k! && (5)\\ \iff k\left(\frac{k}{2}-1\right)! &\leq \prod_{i=1}^{\frac{k}{2}-1}(k-i)\\ \iff 1 &\leq \frac{k-1}{k}\prod_{i=2}^{\frac{k}{2}-1}\frac{k-i}{\frac{k}{2}+1-i} && (6)\end{aligned}$$

Observe that (5) is true for sufficiently large k. Therefore, there exists a universal constant C such that

$$\begin{aligned}\mathbb{E}[(X_1X_2-\mathbb{E}[X_1X_2])^k] &= \left(k\Gamma\left(\frac{k}{2}\right)\right)^2 2^k(\sigma_1\sigma_2)^2\\ &\leq Ck!2^k(\sigma_1\sigma_2)^k\\ &\leq \frac{1}{2}k!(\sigma_1\sigma_2)^2(\tilde{C}\sigma_1\sigma_2)^k\end{aligned}$$

For sufficiently large $\tilde{C}$. Therefore, since $\mathrm{Var}(X_1X_2) \leq \sigma_1^2\sigma_2^2$, by Bernstein's theorem, X_1X_2 is subexponential with parameters $(\nu = \sqrt{2}\sigma_1\sigma_2, b = 2\tilde{C}\sigma_1\sigma_2)$. This establishes the desired result.

$\square$

4. *In class we proved McDiarmid's inequality for bounded random variables. But now we will look at extensions for unbounded R.V's. Take a look at "Concentration in unbounded metric spaces and algorithmic stability" by Aryeh Kontorovich, `https://arxiv.org/pdf/1309.1007.pdf`. Reproduce the proof of theorem 1. The steps of this proof is very similar to the martingale based inequalities we looked at in class. (Thanks to Shentao Yang)*

Solution.
Denote the following,

$$\begin{aligned} V_i &\triangleq \mathbb{E}\left[\phi \mid X_1^i\right]-\mathbb{E}\left[\phi \mid X_1^{i-1}\right], \\ \mathbb{E}\left[\phi \mid X_1^i\right] &= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} P\left(x_{i+1}^n\right) \phi\left(X_1^i x_{i+1}^n\right), \\ \mathbb{E}\left[\phi \mid X_1^{i-1}\right] &= \sum_{x_i^n \in \mathcal{X}_i^n} P\left(x_i^n\right) \phi\left(X_1^{i-1} x_i^n\right) \\ P\left(x_i^n\right) &= P\left(x_{i+1}^n\right) P\left(x_i\right). \end{aligned} \tag{7}$$

Given X_1^{i-1}, we have,

$$V_i \mid X_1^{i-1}=\sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} P\left(x_{i+1}^n\right) \phi\left(X_1^{i-1} X_i x_{i+1}^n\right)-\sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} P\left(x_{i+1}^n\right) \sum_{y^{\prime} \in \mathcal{X}_i} P_i\left(y^{\prime}\right) \phi\left(X_1^{i-1} y^{\prime} x_{i+1}^n\right) \tag{8}$$

We would like to bound the conditional moment generating function of V_i given X_1^{i-1}. Using Jensen's inequality and the fact the $\mathcal{B}(X_1^{i-1}) \subseteq \mathcal{B}(X_1^i)$, where $\mathcal{B}(\cdot)$ denotes the Borel algebra generated by the random variable, we have,

$$\begin{aligned} \mathbb{E}\left(e^{\lambda V_i} \mid X_1^{i-1}\right) &= \mathbb{E}\left(e^{\lambda\left(\mathbb{E}\left[\phi \mid X_1^i\right]-\mathbb{E}\left[\phi \mid X_1^{i-1}\right]\right)} \mid X_1^{i-1}\right) \\ &= e^{-\lambda \mathbb{E}\left[\phi \mid X_1^{i-1}\right]} \mathbb{E}\left(e^{\lambda \mathbb{E}\left[\phi \mid X_1^i\right]} \mid X_1^{i-1}\right) \\ &\leq \mathbb{E}\left(e^{-\lambda \phi} \mid X_1^{i-1}\right) \cdot \mathbb{E}\left(e^{\lambda \phi} \mid X_1^{i-1}\right) \\ &= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} P\left(x_{i+1}^n\right)\left[\sum_{y^{\prime} \in \mathcal{X}_i} P_i\left(y^{\prime}\right) e^{-\lambda \phi\left(X_1^{i-1} y^{\prime} x_{i+1}^n\right)} \cdot \sum_{y \in \mathcal{X}_i} P_i(y) e^{\lambda \phi\left(X_1^{i-1} y x_{i+1}^n\right)}\right]. \end{aligned} \tag{9}$$

Finally we conclude that,

$$\mathbb{E}\left(e^{\lambda V_i} \mid X_1^{i-1}\right) \leq \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} P\left(x_{i+1}^n\right) \sum_{y, y^{\prime} \in \mathcal{X}_i} P_i(y) P_i\left(y^{\prime}\right) e^{\lambda\left(\phi\left(X_1^{i-1} y x_{i+1}^n\right)-\phi\left(X_1^{i-1} y^{\prime} x_{i+1}^n\right)\right)}, \tag{10}$$

For fixed $X_1^{i-1} \in \mathcal{X}_1^{i-1}$, $x_{i+1}^n \in \mathcal{X}_{i+1}^n$, define $F: \mathcal{X}_i \rightarrow \mathbb{R}$, $F(y)=\phi\left(X_1^{i-1} y x_{i+1}^n\right)$. We have the following two lemmas.

Lemma 1. *$F(y)$ is Lipschitz-1 on $\mathcal{X}_i$ w.r.t. metric ρ_i.*

Proof. Using the Lipschitz-1 property of ϕ, we have,

$$\begin{aligned} \left|F(y)-F\left(y'\right)\right| &= \left|\phi\left(X_1^{i-1} y x_{i+1}^n\right)-\phi\left(X_1^{i-1} y' x_{i+1}^n\right)\right| \\ &\leq \rho\left(X_1^{i-1} y x_{i+1}^n, X_1^{i-1} y' x_{i+1}^n\right) \\ &= \sum_{j=1}^{i-1} \rho_j\left(X_j, X_j\right)+\rho_i\left(y, y'\right)+\sum_{j=i+1}^n \rho_j\left(X_j, X_j\right) \\ &= \rho_i\left(y, y'\right). \end{aligned} \tag{11}$$

Therefore, $F(y)$ is Lipschitz-1 on $\mathcal{X}_i$ *w.r.t.* metric ρ_i □

Lemma 2. $\mathbb{E}_{y,y'\sim P_i}\left(e^{\lambda(F(y)-F(y'))}\right)=\mathbb{E}_{y,y'\sim P_i}\left(e^{\lambda(F(y')-F(y))}\right)$.

Proof. Using change of variable, we have,

$$\begin{aligned} \mathbb{E}_{y,y'\sim P_i}\left(e^{\lambda(F(y)-F(y'))}\right) &= \sum_{y,y'\in\mathcal{X}_i} P_i(y) P_i\left(y'\right) e^{\lambda(F(y)-F(y'))} \\ &= \sum_{y,y'\in\mathcal{X}_i} P_i\left(y'\right) P_i(y) e^{\lambda(F(y')-F(y))} \\ &= \mathbb{E}_{y,y'\sim P_i}\left(e^{\lambda(F(y')-F(y))}\right), \end{aligned} \tag{12}$$

which proves the claim. □

Note that $e^t+e^{-t}=2\cosh(t)\leq 2\cosh(s)=e^s+e^{-s}, \forall\, s\geq |t|$. Using Lemma 1, we have,

$$e^{\lambda(F(y)-F(y'))}+e^{\lambda(F(y')-F(y))}\leq e^{\lambda\rho_i(y,y')}+e^{-\lambda\rho_i(y,y')} \tag{13}$$

Using Lemma 2, we have,

$$\begin{aligned} \mathbb{E}_{y,y'\sim P_i}\left(e^{\lambda(F(y)-F(y'))}\right) &\leq \frac{1}{2}\cdot\mathbb{E}_{y,y'\sim P_i}\left(e^{\lambda\rho_i(y,y')}+e^{-\lambda\rho_i(y,y')}\right) \\ &= \mathbb{E}_{y,y'\sim P_i}\mathbb{E}_{\epsilon_i}\left(e^{\lambda\epsilon_i\rho_i(y,y')}\right) \\ &= \mathbb{E}_{y,y'\sim P_i,\epsilon_i}\left(e^{\lambda\cdot\Xi(\mathcal{X}_i)}\right) \\ &\leq e^{\frac{\lambda^2\sigma^*(\Xi(\mathcal{X}_i))^2}{2}} \\ &= e^{\frac{\lambda^2\Delta_{\mathrm{SG}}^2(\mathcal{X}_i)}{2}}. \end{aligned} \tag{14}$$

Thus,

$$\mathbb{E}\left(e^{\lambda V_i}\mid X_1^{i-1}\right)\leq\sum_{x_{i+1}^n\in\mathcal{X}_{i+1}^n} P\left(x_{i+1}^n\right) e^{\frac{\lambda^2\Delta_{\mathrm{SG}}^2(\mathcal{X}_i)}{2}}=e^{\frac{\lambda^2\Delta_{\mathrm{SG}}^2(\mathcal{X}_i)}{2}}. \tag{15}$$

Finally, note that $\phi = \mathbb{E}\left[\phi \mid X_1^n\right], \mathbb{E}\phi = \mathbb{E}\left[\phi \mid X_1^0\right] \implies \phi - \mathbb{E}\phi = \sum_{i=1}^n V_i$. Using Markov inequality,

$$\begin{aligned}
\Pr\left(\phi - \mathbb{E}\phi > t\right) &= \Pr\left(\sum_{i=1}^n V_i > t\right) \\
&= \Pr\left(\lambda \sum_{i=1}^n V_i > \lambda t\right), \quad \forall\, \lambda > 0 \\
&= \Pr\left(\exp\left(\lambda \sum_{i=1}^n V_i\right) > \exp\left(\lambda t\right)\right) \\
&\le e^{-\lambda t} \mathbb{E} \exp\left(\lambda \sum_{i=1}^n V_i\right) \\
&= e^{-\lambda t} \mathbb{E}\left(\prod_{i=1}^n e^{\lambda V_i}\right).
\end{aligned} \tag{16}$$

Note that V_i's are not independent, but using (15) and notice that $V_1, \ldots, V_{n-1} \sim \sigma\left(X_1^{n-1}\right)$, we have,

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda V_1} \cdots e^{\lambda V_n}\right] &= \mathbb{E}\mathbb{E}\left[e^{\lambda V_1} \cdots e^{\lambda V_n} \mid X_1^{n-1}\right] \\
&= \mathbb{E}\left[e^{\lambda V_1} \cdots e^{\lambda V_{n-1}} \cdot \mathbb{E}\left[e^{\lambda V_n} \mid X_1^{n-1}\right]\right] \\
&\le \mathbb{E}\left[e^{\lambda V_1} \cdots e^{\lambda V_{n-1}} \cdot e^{\frac{\lambda^2 \Delta_{\mathrm{SG}}^2(\mathcal{X}_n)}{2}}\right] \\
&= e^{\frac{\lambda^2 \Delta_{\mathrm{SG}}^2(\mathcal{X}_n)}{2}} \cdot \mathbb{E}\left[e^{\lambda V_1} \cdots e^{\lambda V_{n-1}}\right] \\
&\le \cdots \\
&\le \prod_{i=1}^n e^{\frac{\lambda^2 \Delta_{\mathrm{SG}}^2(\mathcal{X}_i)}{2}} = e^{\frac{\lambda^2}{2} \sum_{i=1}^n \Delta_{\mathrm{SG}}^2(\mathcal{X}_i)}.
\end{aligned} \tag{17}$$

Plug into (16), we get,

$$\Pr\left(\phi - \mathbb{E}\phi > t\right) \le e^{\frac{\lambda^2}{2} \sum_{i=1}^n \Delta_{\mathrm{SG}}^2(\mathcal{X}_i) - \lambda t}. \tag{18}$$

Optimize with respect to $\lambda > 0$, the minimum of RHS is achieved when $\lambda = \frac{t}{\sum_{i=1}^n \Delta_{\mathrm{SG}}^2(\mathcal{X}_i)}$, and thus,

$$\Pr\left(\phi - \mathbb{E}\phi > t\right) \le e^{-\frac{t^2}{2 \sum_{i=1}^n \Delta_{\mathrm{SG}}^2(\mathcal{X}_i)}}. \tag{19}$$

The other side of the bound is the same. Apply the union bound, we get the desire result.

$\square$