SDS 384-11 PS #3, Spring 2020

Instructor: Purnamrita Sarkar

Solutions thanks to **Matthew Faw**

Due: Sunday, March 29, 2020 at 11:59pm

Last modified: Tuesday 28th April, 2020 at 11:32pm

3 More concentrations, and U-Statistics

Exercise 3.1 (2+2). *We will use the Efron-Stein inequality to obtain bounds on the variances for separately convex functions whose partial derivatives exist. A separately convex function $f(x_1, \dots, x_n)$ is a convex function of its ith variable, when all else are held fixed.*

1. *Let $X_1, \dots, X_n$ be independent random variables taking values in the interval $[0,1]$ and let $f : [0,1]^n \to R$ be a separately convex function whose partial derivatives exist. Then $f(X) := f(X_1, \dots, X_n)$ satisfies*

$$\mathrm{Var}(f(X)) \leq \mathbb{E}[\|\nabla f(X)\|^2].$$

Hint: Recall that $\mathrm{Var}(Z) \leq \sum_i \mathbb{E}[Z - \mathbb{E}_i Z]^2 \leq \sum_i \mathbb{E}[Z - Z_i]^2$, *where* $\mathbb{E}_i[Z] = \mathbb{E}[Z \mid X_{1;i-1}, X_{i+1,n}]$. *Define* $Z_i = \inf_x f(X_{1:i-1}, x, X_{i+1,n})$ *and then use convexity of f.*

Solution .

Denoting $g_{X_{\neq i}}(x) = f(X_{1:i-1}, x, X_{i+1,n})$ (which is a convex function), and taking Let $X_1', \dots, X_n'$ as independent copies of $X_1, \dots, X_n$, by the Efron-Stein inequality, we have that

$$\begin{aligned}
\mathrm{Var}(f(X)) &\leq \frac{1}{2}\sum_i \mathbb{E}[(g_{X_{\neq i}}(X_i) - g_{X_{\neq i}}(X_i'))^2] \\
&= \frac{1}{2}\sum_i \mathbb{E}\left[\left((g_{X_{\neq i}}(X_i) - g_{X_{\neq i}}(X_i'))_+\right)^2 \mathbb{1}\{g_{X_{\neq i}} \geq g_{X_{\neq i}}(X_i')\}\right] \\
&\quad + \frac{1}{2}\sum_i \mathbb{E}\left[\left((g_{X_{\neq i}}(X_i) - g_{X_{\neq i}}(X_i'))_-\right)^2 \mathbb{1}\{g_{X_{\neq i}} \leq g_{X_{\neq i}}(X_i')\}\right] \\
&= \sum_i \mathbb{E}\left[\left(\left(g_{X_{\neq i}}(X_i) - g_{X_{\neq i}}(X_i')\right)_+\right)^2 \mathbb{1}\{g_{X_{\neq i}} \geq g_{X_{\neq i}}(X_i')\}\right] && \text{by symmetry} \\
&\leq \sum_i \mathbb{E}\left[\left(\left(\frac{d}{dx}g_{X_{\neq i}}(X_i)\underbrace{(X_i - X_i')}_{\in[-1,1]}\right)_+\right)^2 \underbrace{\mathbb{1}\{g_{X_{\neq i}} \geq g_{X_{\neq i}}(X_i')\}}_{\leq 1}\right] && \text{by convexity in the } i\text{th slot} \\
&\leq \mathbb{E}\left[\sum_i \left(\frac{d}{dx}g_{X_{\neq i}}(X_i)\right)^2\right] \\
&= \mathbb{E}\left[\|\nabla f(X)\|^2\right]
\end{aligned}$$

as desired. $\square$

2. *Let A be an $m \times n$ random matrix with independent entries $A_{ij} \in [0,1]$. Let*

$$Z = \sqrt{\lambda_1(A^T A)} = \sqrt{\sup_{u \in \mathbb{R}^n \,:\, \|u\|_2 = 1} u^T A^T A u} = \sup_{u \in \mathbb{R}^n \,:\, \|u\|_2 = 1} \|Au\|_2$$

Show that $\mathrm{Var}(Z) \leq 1$.

Solution .
We begin by noting that, for any $u \in \mathbb{R}^n$, $\|Au\|_2$ is a convex function in every entry $a_{i,j}$ of A. Thus, Z may be equivalently written as $Z = \sup_{\|u\|_2 \leq 1} \|Au\|_2$. Thus, since the sup is taken over a compact convex set, $f(A) = \sup_{\|u\| \leq 1} \|Au\|$ is also a convex function in every coordinate of A, and thus has a subdifferential.

Now, examining the argument from the previous part, it immediately follows that, for any element $g_A \in \partial f(A)$, the subdifferential of f at A,

$$\mathrm{Var}(f(A)) \leq \mathbb{E}[\|g_A\|^2]$$

Additionally, a standard fact is that, since f is convex, if f is L-Lipschitz, then $\|g_A\| \leq L$. Thus, it is sufficient to prove that f is 1-Lipschitz.

To see this, we observe that

$$\begin{aligned}
|f(A) - f(B)| &= \left| \sup_{\|u\|=1} \|Au\| - \sup_{\|v\|=1} \|Bv\| \right| \\
&= |\|Au^*\| - \|Bv^*\|| \\
&\leq \max_{w \in \{u^*, v^*\}} |\|Aw\| - \|Bw\|| \\
&\leq \max_{w \in \{u^*, v^*\}} \|(A - B)w\| \\
&\leq \sup_{\|w\|=1} \|(A - B)w\| \\
&= \|(A - B)w^*\| \\
&= \sqrt{\lambda_1((A - B)^T (A - B))} \\
&\leq \sqrt{\sum_i \lambda_i((A - B)^T (A - B))} \\
&= \sqrt{\mathrm{trace}((A - B)^T (A - B)} \\
&= \|A - B\|_F
\end{aligned}$$

Which establishes that f is 1-Lipschitz, and thus, by the generalization to the previous problem, that

$$\mathrm{Var}(Z) \leq 1$$

as desired. □

Exercise 3.2 (2+6)**.** *In this question we will look at the Gaussian Lipschitz Theorem. Consider* $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(0,1)$.

1. *Prove that the order statistics are* 1*-Lipschitz.*

   Solution .
   We begin by noting that the claim holds trivially for the first and last order statistics, since, for any sample path $\omega \in \Omega$,

$$\begin{aligned}
|X_{(1)}(\omega) - Y_{(1)}(\omega)| &= \max\{\max_i X_i(\omega) - \max_j Y_j(\omega), \max_j Y_j(\omega) - \max_i X_i(\omega)\} \\
&= \max\{X_{i^*}(\omega) - \underbrace{Y_{j^*}(\omega)}_{\geq Y_{i^*}}, Y_{j^*}(\omega) - \underbrace{X_{i^*}(\omega)}_{\geq X_{j^*}(\omega)}\} \\
&\leq \max\{X_{i^*}(\omega) - Y_{i^*}(\omega), Y_{j^*}(\omega) - X_{j^*}\} \\
&\leq \|X(\omega) - Y(\omega)\|
\end{aligned}$$

   and similarly,

$$\begin{aligned}
|X_{(n)}(\omega) - Y_{(n)}(\omega)| &= \max\{\min_i X_i(\omega) - \min_j Y_j(\omega), \min_j Y_j(\omega) - \min_i X_i(\omega)\} \\
&= \max\{\underbrace{X_{i_*}(\omega)}_{\leq X_{j_*}} - Y_{j_*}(\omega), \underbrace{Y_{j_*}(\omega)}_{\leq Y_{i_*}} - X_{i_*}(\omega)\} \\
&\leq \max\{X_{j_*}(\omega) - Y_{j_*}(\omega), Y_{i_*}(\omega) - X_{i_*}\} \\
&\leq \|X(\omega) - Y(\omega)\|
\end{aligned}$$

   Now, for any $k \in (0, n)$, (assuming WLOG along this sample path, $X_{(k)}(\omega) \geq Y_{(k)}(\omega)$), there are $k-1$ entries in $X(\omega)$ that are larger than $X_{(k)}(\omega)$, and $n-k$ entries in $Y(\omega)$ that are smaller than $Y_{(k)}(\omega)$. Thus, by the pigeonhole principle, either at least one index of these entries is shared, or the index of the kth order statistic on this sample path is the same for $X(\omega)$ and $Y(\omega)$ – indeed, if the $n-1$ indices are distinct, then the index of the kth largest entry for $X(\omega)$ and $Y(\omega)$ must be shared. In either case, we may conclude that

$$|X_{(k)}(\omega) - Y_{(k)}(\omega)| \leq \|X(\omega) - Y(\omega)\|.$$

   Therefore, the order statistics are 1-Lipschitz almost surely. □

2. *(2+1+1+1+1) Now show that, for large enough* n,

$$c\sqrt{\log n} \leq \mathbb{E}[\max_i X_i] \leq \sqrt{2 \log n}$$

   for some universal constant c.

   (a) *For the upper bound, let* $Y = \max_i X_i$. *First, show that* $\exp(t\mathbb{E}[Y]) \leq \sum_i \mathbb{E}[\exp(tX_i)]$. *Now pick a* t *to get the right form.*

   Solution .
   Observe that

$$\begin{aligned}
\exp(t\mathbb{E}Y) &\leq \mathbb{E}[\exp(t \max_i X_i)] && \text{Jensen's – } \exp(t\cdot) \text{ is convex for all } t \in \mathbb{R} \\
&\leq \mathbb{E} \max_i \exp(tX_i) && \exp(t\cdot) \text{ is bijective} \\
&\leq \mathbb{E} \sum_i \exp(tX_i) && \exp(t\cdot) \text{ is nonnegative} \\
&= n\mathbb{E}\exp(tX_i) && X_i \text{ are i.i.d} \\
&= \exp(t^2/2 + \log n) && X_i \sim \mathcal{N}(0,1)
\end{aligned}$$

In particular, rearranging the resulting expression and taking logs on both sides, we have that

$$\mathbb{E}Y \leq \frac{t}{2} + \frac{\log n}{t}$$

and plugging in $t = \sqrt{2 \log n}$ (by optimizing the expression for t), we have

$$\mathbb{E}Y \leq \sqrt{2 \log n}$$

as desired. $\square$

(b) *For the lower bound, do the following steps.*

i. *Show that* $\mathbb{E}[Y] \geq \delta \mathbb{P}(Y \geq \delta) + \mathbb{E}[\min\{Y, 0\}]$.

Solution .
We have that, for any $\delta > 0$,

$$\begin{aligned}
\mathbb{E}Y &= \mathbb{E}[\underbrace{Y \mathbb{1}\{Y \geq \delta\}}_{\geq \delta \mathbb{1}\{Y \geq \delta\}}] + \mathbb{E}[\underbrace{Y \mathbb{1}\{Y < \delta\}}_{\geq Y \mathbb{1}\{Y \leq 0\}}] \\
&\geq \delta \mathbb{P}(Y \geq \delta) + \mathbb{E}[Y \mathbb{1}\{Y \leq 0\}] \\
&= \delta \mathbb{P}(Y \geq \delta) + \mathbb{E}[\min\{Y, 0\}]
\end{aligned}$$

as desired. $\square$

ii. *Now show that* $\mathbb{E}[\min\{Y, 0\}] \geq \mathbb{E}[\min\{X_1, 0\}]$.

Solution .

$$\begin{aligned}
\mathbb{E}[\min\{Y, 0\}] &= \mathbb{E}[\min\{Y, 0\} \mid Y = X_1]\mathbb{P}(Y = X_1) + \mathbb{E}[\underbrace{\min\{Y, 0\}}_{\geq \min\{X_1, 0\}} \mid Y > X_1]\mathbb{P}(Y > X_1) \\
&\geq \mathbb{E}[\min\{X_1, 0\}]\mathbb{P}(Y = X_1) + \mathbb{E}[\min\{X_1, 0\}]\mathbb{P}(Y > X_1) \\
&= \mathbb{E}[\min\{X_1, 0\}] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} x \exp(-x^2/2) dx \\
&= -\frac{1}{\sqrt{2\pi}}
\end{aligned}$$

$\square$

iii. *Finally, relate* $\mathbb{P}(Y \geq \delta)$ *to* $\mathbb{P}(X_1 \geq \delta)$ *by using independence.*

Solution .

$$
\begin{aligned}
\mathbb{P}(Y \geq \delta) &= \mathbb{P}(\max_i X_i \geq \delta) \\
&= \mathbb{P}\left(\exists i \in [n] : X_i \geq \delta\right) \\
&= 1 - \mathbb{P}\left(\forall i \in [n] : X_i < \delta\right) \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}\left(X_i < \delta\right) \\
&= 1 - \mathbb{P}\left(X_1 < \delta\right)^n \\
&= 1 - \left(1 - \mathbb{P}\left(X_1 \geq \delta\right)\right)^n
\end{aligned}
$$

□

iv. *Now show that $\mathbb{P}(X_1 \geq \delta) \geq \exp(-\delta^2/\sigma^2)/c$ for some universal constant c.*

Solution .
Note that, for $X_1 \sim \mathcal{N}(0, \sigma^2)$,

$$
\begin{aligned}
\mathbb{P}(X_1 \geq \delta) &= \frac{1}{\sigma\sqrt{2\pi}} \int_\delta^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right) dx \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty \exp\left(-\underbrace{\frac{(x+\delta)^2}{2\sigma^2}}_{\leq \frac{x^2+\delta^2}{\sigma^2}}\right) dx \\
&\geq \frac{\exp\left(-\frac{\delta^2}{\sigma^2}\right)}{\sigma\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{x^2}{\sigma^2}\right) dx \\
&= \frac{\exp\left(-\frac{\delta^2}{\sigma^2}\right)}{\sqrt{2}} \left(\frac{1}{\frac{\sigma}{\sqrt{2}}\sqrt{2\pi}} \int_0^\infty \exp\left(-\frac{x^2}{2\left(\frac{\sigma^2}{\sqrt{2}}\right)^2}\right) dx\right) \\
&= \frac{\exp\left(-\frac{\delta^2}{\sigma^2}\right)}{2\sqrt{2}}
\end{aligned}
$$

In particular, since in our case, $\sigma = 1$, we have that

$$
\mathbb{P}(X_1 \geq \delta) \geq \frac{\exp\left(-\delta^2\right)}{2\sqrt{2}}
$$

as desired, with the universal constant $c = 2\sqrt{2}$ □

v. *Choose the parameter δ carefully to have $\mathbb{P}(X_1 \geq \delta) \geq \frac{1}{n}$ for large enough n.*

Solution .
Solving for

$$
\frac{\exp(-\delta^2)}{2\sqrt{2}} \geq \frac{1}{n}
$$

we have that, by choosing

$$\delta \leq \sqrt{\log \frac{n}{2\sqrt{2}}} \qquad \text{assuming } n \geq 2\sqrt{2}$$

then we have that

$$\mathbb{P}(X_1 \geq \delta) \geq \frac{1}{n}$$

In particular, note that for any $c > 1$, for sufficiently large n, we can take $\delta = \sqrt{\frac{\log n}{c}}$, since

$$\log \frac{n}{2\sqrt{2}} \geq \frac{\log n}{c}$$
$$\iff n^c \geq (2\sqrt{2})^2 n$$

holds for sufficiently large n. □

Solution .
Putting these results together, we have that, for $\delta = \sqrt{\frac{\log n}{c}}$, $c > 1$, for sufficiently large n,

$$\begin{aligned}
\mathbb{E}Y &\geq \sqrt{\frac{\log n}{c}} \left(1 - \left(1 - \frac{1}{n}\right)^n\right) - \frac{1}{\sqrt{2\pi}} \\
&\geq \sqrt{\frac{\log n}{c}} \exp(-1) - \frac{1}{\sqrt{2\pi}} \\
&\geq \sqrt{\frac{\log n}{2c \exp(2)}} && \text{assuming } n \text{ large enough}
\end{aligned}$$

which is the desired result. □

Exercise 3.3. *(6 pts) 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 "Concentrations in unbounded metric spaces and algorithmic stability" by Aryeh Kontrovich. Reproduce the proof of Theorem 1. The steps of this proof are very similar to the Martingale based inequalities we looked at in class.*

Solution .
We wish to prove the following theorem:

Theorem 1. *If $\varphi : \mathcal{X}^n \to \mathbb{R}^n$ is 1−Lipschitz, then $\mathbb{E}\varphi < \infty$, and*

$$\mathbb{P}(|\varphi - \mathbb{E}\varphi| > t) \leq 2 \exp\left(-\frac{t^2}{2\sum_{i=1}^n \Delta_{\text{SG}}^2(\mathcal{X}_i)}\right)$$

Proof. As in the paper, we will denote

$$V_i = \mathbb{E}[\varphi \mid X_1^i] - \mathbb{E}[\varphi \mid X_1^{i-1}]$$

Now, we may rewrite this expression as

$$
\begin{aligned}
V_i &= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} \mathbb{P}(x_{i+1}^n) \left(\varphi(X_1^{i-1}, X_i, x_{i+1}^n) - \sum_{\tilde{x}_i \in \mathcal{X}_i} \mathbb{P}(\tilde{x}_i) \varphi(X_1^{i-1}, x_i, x_{i+1}^n) \right) \\
&= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} \mathbb{P}(x_{i+1}^n) \left(\sum_{x_i \in \mathcal{X}_i} \mathbb{1}\{X_i = x_i\} \varphi(X_1^{i-1}, x_i, x_{i+1}^n) - \sum_{\tilde{x}_i \in \mathcal{X}_i} \mathbb{P}(\tilde{x}_i) \varphi(X_1^{i-1}, \tilde{x}_i, x_{i+1}^n) \right) \\
&= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} \mathbb{P}(x_{i+1}^n) \sum_{\tilde{x}_i \in \mathcal{X}_i} \sum_{x_i \in \mathcal{X}_i} \mathbb{P}(\tilde{x}_i) \mathbb{1}\{X_i = x_i\} \left(\varphi(X_1^{i-1}, x_i, x_{i+1}^n) - \varphi(X_1^{i-1}, \tilde{x}_i, x_{i+1}^n) \right)
\end{aligned}
$$

Hence, by three applications of Jensen's inequality applied to each of the three convex combinations, we may conclude that

$$
\begin{aligned}
&\mathbb{E}[\exp(\lambda V_i) \mid X_1^{i-1}] \\
&\leq \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} \mathbb{P}(x_{i+1}^n) \sum_{x_i, \tilde{x}_i \in \mathcal{X}_i} \mathbb{E} \left[\mathbb{P}(\tilde{x}_i) \mathbb{1}\{X_i = x_i\} \exp \left(\lambda \left(\varphi(X_1^{i-1}, x_i, x_{i+1}^n) - \varphi(X_1^{i-1}, \tilde{x}_i, x_{i+1}^n) \right) \right) \mid X_1^{i-1} \right] \\
&= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} \mathbb{P}(x_{i+1}^n) \sum_{x_i, \tilde{x}_i \in \mathcal{X}_i} \mathbb{P}(\tilde{x}_i) \underbrace{\mathbb{E}[\mathbb{1}\{X_i = x_i\} \mid X_1^{i-1}]}_{= \mathbb{E}[\mathbb{1}\{X_i = x_i\}] = \mathbb{P}(x_i)} \exp \left(\lambda \left(\varphi(X_1^{i-1}, x_i, x_{i+1}^n) - \varphi(X_1^{i-1}, \tilde{x}_i, x_{i+1}^n) \right) \right) \\
&= \sum_{x_{i+1}^n \in \mathcal{X}_{i+1}^n} \mathbb{P}(x_{i+1}^n) \sum_{x_i, \tilde{x}_i \in \mathcal{X}_i} \mathbb{P}(\tilde{x}_i) \mathbb{P}(x_i) \exp \left(\lambda \left(\varphi(X_1^{i-1}, x_i, x_{i+1}^n) - \varphi(X_1^{i-1}, \tilde{x}_i, x_{i+1}^n) \right) \right)
\end{aligned}
$$

Now, for fixed $X_1^{i-1} \in \mathcal{X}_1^{i-1}$ and $x_{i+1}^n \in \mathcal{X}_{i+1}^n$, define $F : \mathcal{X}_i \to \mathbb{R}$ as $F(y) = \varphi(X_1^{i-1}, y, x_{i+1}^n)$. Since φ is 1-Lipschitz with respect to the ℓ_1 product measure ρ^n, F is 1-Lipschitz with respect to ρ_i.

Now, recalling that $\exp(t) + \exp(-t) = 2\cosh(t)$, and that $\cosh(t) \leq \cosh(s)$ for all $|t| \leq s$, we have that

$$
\begin{aligned}
\exp(\lambda(F(x_i) - F(\tilde{x}_i))) + \exp(-\lambda(F(x_i) - F(\tilde{x}_i))) &\leq 2\cosh(\lambda |F(x_i) - F(\tilde{x}_i)|) \\
&\leq 2\cosh(\lambda \rho_i(x_i, \tilde{x}_i)) \\
&= \exp(\lambda \rho_i(x_i, \tilde{x}_i)) + \exp(-\lambda \rho_i(x_i, \tilde{x}_i))
\end{aligned}
$$

Denote ϵ_i as a Rademacher random variable that is independent of all other randomness. Then, applying our symmetrization argument, we have that

$$\begin{aligned}
&\sum_{x_i, \tilde{x}_i \in \mathcal{X}_i} \mathbb{P}(x_i)\mathbb{P}(\tilde{x}_i) \exp(\lambda(F(x_i) - F(\tilde{x}_i))) \\
&= \mathbb{E}_{X_i, \tilde{X}_i}[\exp(\lambda(F(X_i) - F(\tilde{X}_i)))] \\
&= \mathbb{E}_{X_i, \tilde{X}_i}[\mathbb{E}_{\epsilon_i}[\exp(\lambda\epsilon_i(F(X_i) - F(\tilde{X}_i)))]] \\
&= \mathbb{E}_{X_i, \tilde{X}_i}\left[\frac{1}{2}\exp(\lambda(F(X_i) - F(\tilde{X}_i))) + \frac{1}{2}\exp(-\lambda(F(X_i) - F(\tilde{X}_i)))\right] \\
&= \mathbb{E}_{X_i, \tilde{X}_i}\left[\frac{1}{2}\exp(\lambda\rho_i(X_i, \tilde{X}_i)) + \frac{1}{2}\exp(-\lambda\rho_i(X_i, \tilde{X}_i))\right] \\
&= \mathbb{E}_{X_i, \tilde{X}_i}\left[\mathbb{E}_{\epsilon_i}[\exp(\lambda\epsilon_i\rho_i(X_i, \tilde{X}_i))]\right] \\
&= \mathbb{E}\left[\exp(\lambda\Xi(\mathcal{X}_i)\right] \\
&\leq \exp(\lambda^2\Delta_{\text{SG}}^2(\mathcal{X}_i)/2)
\end{aligned}$$

where

$$\Xi(\mathcal{X}_i) = \epsilon_i\rho_i(X_i, \tilde{X}_i)$$

and $\Delta_{\text{SG}}(\mathcal{X}_i)$ is the smallest subGaussian parameter for the random variable $\Xi(\mathcal{X}_i)$.

Now, combining results, we conclude that

$$\mathbb{E}[\exp(\lambda V_i) \mid X_1^{i-1}] \leq \exp(\lambda^2\Delta_{\text{SG}}^2(\mathcal{X}_i)/2)$$

Finally, conclude by applying the standard exponentiated Markov's inequality:

$$\begin{aligned}
\mathbb{P}(\varphi - \mathbb{E}\varphi > t) &= \mathbb{P}\left(\sum_{i=1}^n V_i > t\right) \\
&= \mathbb{P}\left(\exp(\lambda\sum_{i=1}^n V_i) > \exp(\lambda t)\right) \\
&\leq \exp(-\lambda t)\mathbb{E}\left[\exp(\lambda\sum_{i=1}^n V_i)\right] \\
&= \exp(-\lambda t)\mathbb{E}\left[\prod_{i=1}^n \exp(\lambda V_i)\right] \\
&= \exp(-\lambda t)\mathbb{E}\left[\prod_{i=1}^n \mathbb{E}[\exp(\lambda V_i) \mid X_1^{i-1}]\right] \\
&\leq \exp(-\lambda t)\mathbb{E}\left[\prod_{i=1}^n \exp(\lambda^2\Delta_{\text{SG}}^2(\mathcal{X}_i)/2)\right] \\
&= \exp\left(\frac{1}{2}\lambda^2\sum_{i=1}^n \Delta_{\text{SG}}(\mathcal{X}_i) - \lambda t\right)
\end{aligned}$$

Choosing $\lambda = \frac{t}{\sum_{i=1}^n \Delta_{\text{SG}}^2(\mathcal{X}_i)}$ yields

$$\mathbb{P}(\varphi - \mathbb{E}\varphi > t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n \Delta_{\text{SG}}^2(\mathcal{X}_i)}\right)$$

Now, applying the same argument to $\mathbb{E}\varphi - \varphi$ gives the desired claim. $\square$

$\square$

Exercise 3.4 (2+2+(1+1)+(1+1)+3+1)**.** *Consider an i.i.d sample of size n from a discrete distribution parametrized by $p_1, \ldots, p_m$ on m atoms. A common test for uniformity of the distribution is to look at the fraction of pairs that collide, or are equal. Call this statistic U.*

1. *Is U a U-statistic? When is it degenerate?*

Solution .
Let us denote the samples as $\{X_i\}_{i=1}^n$. We may write U as

$$U = \frac{1}{\binom{n}{2}} \sum_{i>j\in[n]} \mathbb{1}\{X_i = X_j\}$$

Hence, by definition, U is a U-statistic of order 2 with (symmetric) kernel $h(x,y) = \mathbb{1}\{x = y\}$. Note that

$$\begin{aligned}
\theta &= \mathbb{E}[\mathbb{1}\{X_1 = X_2\}] \\
&= \sum_{a\in[m]} p_a^2 \\
&= \mathbb{E}[U]
\end{aligned}$$

U is called degenerate when the limiting variance of the U-statistic, $4\xi_1 = 0$. We have that

$$
\begin{aligned}
\xi_1 &= \operatorname{Var}(\mathbb{E}[h(X_1, X_2) \mid X_1]) \\
&= \operatorname{Var}(\mathbb{E}[\mathbb{1}\{X_1 = X_2\} \mid X_1]) \\
&= \operatorname{Var}(\sum_{a \in [m]} p_a \mathbb{1}\{X_1 = a\}) \\
&= \mathbb{E}\left[\left(\sum_{a \in [m]} p_a(\mathbb{1}\{X_1 = a\} - p_a)^2\right)\right] \\
&= \sum_{a \in [m]} p_a^2(p_a - p_a^2) + \sum_{a,b: a \neq b} \mathbb{E}[p_a p_b(\mathbb{1}\{X_1 = a\} - p_a)(\mathbb{1}\{X_1 = b\} - p_b)] \\
&= \sum_{a \in [m]} p_a^3(1 - p_a) - \sum_{a,b: a \neq b} p_a^2 p_b^2 \\
&= \sum_{a \in [m]} \sum_{b: b \neq a} (p_a^3 p_b - p_a^2 p_b^2) \\
&= \sum_{a \in [m]} \sum_{b: b \neq a} p_a^2 p_b (p_a - p_b) \\
&= \sum_{\substack{a,b \in [m] \\ p_a > p_b}} (p_a^2 p_b(p_a - p_b) - p_b^2 p_a(p_a - p_b)) \\
&= \sum_{\substack{a,b \in [m] \\ p_a > p_b}} (p_a - p_b) p_a p_b (p_a - p_b) \\
&= \sum_{\substack{a,b \in [m] \\ p_a > p_b}} (p_a - p_b)^2 p_a p_b
\end{aligned}
$$

Thus, $\xi_1 = 0$ iff $p_a = p_b$ for all $a, b \in \{a \in [m] : p_a > 0\}$. That is, U is degenerate exactly when the distribution is uniform over some subset of the atoms in $[m]$ (and assigns 0 mass to the rest of the atoms). □

2. *What is the variance of U? Please give the exact answer, without approximation.*

Solution .
Recall that, from the previous part,

$$
\xi_1 = \sum_{\substack{a,b \in [m] \\ p_a > p_b}} (p_a - p_b)^2 p_a p_b
$$

Additionally, we may compute

$$\begin{aligned}\xi_2 &= \mathrm{Var}(\mathbb{E}[h(X_1, X_2) \mid X_1, X_2]) \\ &= \mathrm{Var}(h(X_1, X_2)) \\ &= \mathrm{Var}(\mathbb{1}\{X_1 = X_2\}) \\ &= \mathbb{E}(\mathbb{1}\{X_1 = X_2\}) - (\mathbb{E}[\mathbb{1}\{X_1 = X_2\}])^2 \\ &= \sum_{a\in[m]} p_a^2 - \left(\sum_{a\in[m]} p_a^2\right)^2 \\ &= \sum_{a\in[m]} p_a^2(1-p_a^2) - \sum_{a\neq b\in[m]} p_a^2 p_b^2\end{aligned}$$

Therefore, the variance of U is given by

$$\begin{aligned}\mathrm{Var}(U) &= \frac{1}{\binom{n}{2}^2}\sum_{c=0}^{2}\binom{n}{2}\binom{2}{c}\binom{n-2}{2-c}\xi_c \\ &= \frac{1}{\binom{n}{2}^2}\left(\binom{n}{2}\binom{2}{1}\binom{n-2}{1}\xi_1 + \binom{n}{2}\binom{2}{2}\binom{n-2}{0}\xi_2\right) \\ &= \frac{1}{\binom{n}{2}^2}\left(\binom{n}{2}\binom{2}{1}\binom{n-2}{1}\xi_1 + \binom{n}{2}\binom{2}{2}\binom{n-2}{0}\xi_2\right) \\ &= \frac{1}{\binom{n}{2}^2}\left(\binom{n}{2}\binom{2}{1}\binom{n-2}{1}\xi_1 + \binom{n}{2}\binom{2}{2}\binom{n-2}{0}\xi_2\right) \\ &= \frac{1}{\binom{n}{2}}\left(2\binom{n-2}{1}\xi_1 + \xi_2\right) \\ &= \frac{1}{\binom{n}{2}}\left(2(n-2)\sum_{\substack{a,b\in[m]\\ p_a>p_b}}(p_a-p_b)^2 p_a p_b + \sum_{a\in[m]} p_a^2(1-p_a^2) - \sum_{a\neq b\in[m]} p_a^2 p_b^2\right)\end{aligned}$$

□

3. *For a hypothesis test, we will consider alternative distributions which have $p_i = \frac{1+a}{m}$ for half of the atoms of the distribution and $\frac{1-a}{m}$ for the other half $(0 \le a \le 1)$, for some $a > 0$. Assume that there are an even number of atoms.*

   (a) *What are the mean and variance of this statistic under the null?*

   Solution .
   From the previous parts, we note that

$$\begin{aligned}\mathbb{E}[U \mid H_0] &= \frac{1}{\binom{n}{2}}\sum_{i>j\in[n]}\sum_{a\in[m]} p_a^2 \\ &= m\frac{1}{m^2} \\ &= \frac{1}{m}\end{aligned}$$

and additionally,

$$
\begin{aligned}
\operatorname{Var}(U \mid H_0) &= \frac{1}{\binom{n}{2}} \xi_2 \\
&= \frac{1}{\binom{n}{2}} \left(\sum_{a \in [m]} \frac{1}{m^2} \left(1 - \frac{1}{m^2}\right) - 2\binom{m}{2} \frac{1}{m^4} \right) \\
&= \frac{1}{\binom{n}{2}} \left(\frac{1}{m} - \frac{1}{m^3} - \frac{m-1}{m^3} \right) \\
&= \frac{1}{\binom{n}{2}} \left(\frac{1}{m} - \frac{1}{m^2} \right) \\
&= \frac{p_1(1-p_1)}{\binom{n}{2}}
\end{aligned}
$$

□

(b) *What are the mean and variance of this under the alternative?*

Solution .
Similarly, plugging into our equations,

$$
\begin{aligned}
\mathbb{E}[U \mid H_a] &= \frac{1}{\binom{n}{2}} \sum_{i>j} \sum_{a \in [m]} p_a^2 \\
&= \frac{m}{2} \left(\frac{(1-a)^2}{m^2} + \frac{(1+a)^2}{m^2} \right) \\
&= \frac{1+a^2}{m}
\end{aligned}
$$

and, noting that

$$
\begin{aligned}
\xi_1 &= \left(\frac{m}{2}\right) \left(\frac{2a}{m}\right)^2 \left(\frac{(1-a)(1+a)}{m^2} \right) \\
&= \frac{a^2(1-a^2)}{m^2}
\end{aligned}
$$

and

$$
\begin{aligned}
\xi_2 = &\frac{m}{2} \left(\frac{(1-a)^2}{m^2} \left(1 - \frac{(1-a)^2}{m^2}\right) + \frac{(1+a)^2}{m^2} \left(1 - \frac{(1+a)^2}{m^2}\right) \right) \\
&- 2\left(\binom{m/2}{2} \left(\frac{(1-a)^4}{m^4} + \frac{(1+a)^4}{m^4} \right) + \frac{m^2}{4} \frac{(1-a)^2(1+a)^2}{m^4} \right) \\
= &\frac{m}{2} \left(\frac{2+2a^2}{m^2} - \frac{2+12a^2+2a^4}{m^4} \right) - 2\left(\frac{m}{4}\left(\frac{m}{2}-1\right) \frac{2+12a^2+2a^4}{m^4} + \frac{m^4}{4} \frac{1-2a^2+a^4}{m^4} \right) \\
= &\frac{1+a^2}{m} \left(1 - \frac{1+a^2}{m}\right)
\end{aligned}
$$

Therefore, we have that

$$
\operatorname{Var}(U \mid H_a) = \frac{2}{n(n-1)} \left(2(n-2)\frac{a^2(1-a^2)}{m^2} + \frac{1+a^2}{m} \left(1 - \frac{1+a^2}{m}\right) \right)
$$

□

(c) *What is the asymptotic distribution of* U *under the null hypothesis that* $p_i = \frac{1}{m}$*? Hint: you can use the fact that for* $X_1, \ldots, X_N \overset{iid}{\sim} \text{multinomial}(q_1, \ldots, q_k)$, $\sum_{i=1}^{k} \frac{(N_i - Nq_i)^2}{Nq_i} \xrightarrow{d} \chi^2_{k-1}$.

Solution .
Observe that

$$
\begin{aligned}
U &= \frac{1}{\binom{n}{2}} \sum_{i>j} \mathbb{1}\{X_i = X_j\} \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{1}\{X_i = X_j\} \\
&= \frac{1}{n(n-1)} \sum_{a \in [m]} \sum_{i \neq j} \mathbb{1}\{X_i = a\} \mathbb{1}\{X_j = a\} \\
&= \frac{1}{n(n-1)} \sum_{a \in [m]} \left(\left(\underbrace{\sum_{i=1}^{n} \mathbb{1}\{X_i = a\}}_{=N_a} \right)^2 - \sum_{i=1}^{n} \mathbb{1}\{X_i = a\} \right) \\
&= \frac{1}{n(n-1)} \sum_{a \in [m]} N_a^2 - N_a \\
&= \frac{1}{n(n-1)} \left(\sum_{a \in [m]} N_a^2 \right) - \frac{1}{n-1} && \text{since } \sum_a N_a = n \text{ a.s.} \\
&= \frac{1}{(n-1)} \left(\sum_{a \in [m]} \left(\frac{\sqrt{n} N_a}{n} \right)^2 \right) - \frac{1}{n-1} \\
&= \frac{1}{(n-1)} \left(\sum_{a \in [m]} \left(\sqrt{n} F_a \right)^2 \right) - \frac{1}{n-1} && \text{taking } F_a = \frac{N_a}{n} \\
&= \frac{n}{(n-1)} \left(\sum_{a \in [m]} F_a^2 \right) - \frac{1}{n-1}
\end{aligned}
$$

Additionally, under the null hypothesis, by the hint,

$$
\begin{aligned}
V &= \sum_{a\in[m]} \frac{(N_i - np_a)^2}{np_a} \\
&= \sum_{a\in[m]} \frac{1}{p_a}\left(\frac{\sqrt{n}N_a}{n} - \sqrt{n}p_a\right)^2 \\
&= \sum_{a\in[m]} \frac{n}{p_a}\left(F_a - p_a\right)^2 \\
&= \sum_{a\in[m]} \frac{n}{p_a}\left(F_a^2 + p_a^2 - 2p_aF_a\right) \\
&= \sum_{a\in[m]} \frac{n}{p_a}F_a^2 + np_a - 2nF_a \\
&= \sum_{a\in[m]} \frac{n}{p_a}F_a^2 + n\underbrace{\sum_{a\in[m]} p_a}_{=1} - 2n\underbrace{\sum_{a\in[m]} F_a}_{=1} \\
&= \left(\sum_{a\in[m]} \frac{n}{p_a}F_a^2\right) - n \\
&= nm\left(\sum_{a\in[m]} F_a^2\right) - n \\
&\xrightarrow{d} \chi^2_{m-1}
\end{aligned}
$$

Now, observe that, by the above calculations, we have the following:

$$
\begin{aligned}
m(n-m)\left(U - \mathbb{E}[U \mid H_0]\right) &= m(n-m)\left(U - \frac{1}{m}\right) \\
&= m[(n-1)U + 1] - n \\
&= nm\left(\sum_{a\in[m]} F_a^2\right) - n \\
&= V \\
&\xrightarrow{d} \chi^2_{m-1}
\end{aligned}
$$

□

(d) *Under the alternative hypothesis, is it always the case that U has a limiting normal distribution? Can you give a sufficient condition on the sample size n so that this is true?*

Solution .
By normal convergence of U-statistics, since $\mathbb{E}[h^2] = \mathbb{E}[\mathbb{1}\{X_1 = X_2\}] = \frac{1+a^2}{m} < \infty$, we

have that, for $a \in (0,1)$, and treating m as a constant that does not scale with n,

$$\sqrt{n}\,(U - \mathbb{E}[U \mid H_a]) \xrightarrow{d} \mathcal{N}(0, 4\xi_1^2),$$

where

$$\xi_1^2 = \frac{a^2(1-a^2)}{m^2}$$

Note that, when $a \in \{0,1\}$, the U-statistic is *degenerate*. In these cases, we note that the distribution is uniform over $\frac{m}{2}$ atoms when $a = 1$ and uniform over m atoms when $a = 0$. Thus, in both of these cases, the convergence result from the previous section is applicable (with m replaced by $\frac{m}{2}$ for $a = 1$). Thus, in these cases, the limiting distribution is *not* normal.

Additionally, let us refer back to the variance under H_a:

$$\mathrm{Var}(U \mid H_a) = \frac{2}{n(n-1)} \left(2(n-2)\frac{a^2(1-a^2)}{m^2} + \frac{1+a^2}{m}\left(1 - \frac{1+a^2}{m}\right)\right)$$

Note that we can break this term into two terms, the first term (to which ξ_1 contributes) scales as $\frac{1}{nm^2}$, and the second term (to which ξ_2 contributes), scales as $\frac{1}{n^2 m}$ Consider a regime in which $m = n^c$ for some positive constant c. Then, in order for the ξ_1 term to dominate, we need that $c < 1$. Indeed, this ensures that $1 + 2c < 2 + c$, and thus, that the first term will dominate. As noted in the updated homework document, this implies normal convergence in the case where $m = o(n)$.

$\square$