COMS 4772 Fall 2016 Homework 1

Si Kai Lee, sl3950

Problem 1 (25 points)**.** In this problem, "volume" refers to $(d-1)$-dimensional volume (or "surface area" in d-dimensions).

(a) Prove that there is a constant $C > 0$ (not depending on d) such that, for any set $T \subset S^{d-1}$ of $|T| = d^{100}$ unit vectors, the set

$$\bigcap_{\boldsymbol{u} \in T} \left\{ \boldsymbol{x} \in S^{d-1} : \left|\langle \boldsymbol{u}, \boldsymbol{x} \rangle\right| \leq C\sqrt{\frac{\ln d}{d}} \right\}$$

accounts for 99% of the volume of S^{d-1}. (Assume $d \geq 2$ so $\ln(d) > 0$.)

(b) Prove that there is a constant $c > 0$ (not depending on d) such that, for any $\boldsymbol{u} \in S^{d-1}$, the set

$$\left\{ \boldsymbol{x} \in S^{d-1} : \left|\langle \boldsymbol{u}, \boldsymbol{x} \rangle\right| > \frac{c}{\sqrt{d}} \right\}$$

accounts for 99% of the volume of S^{d-1}.

Solution.

a) Let $Y_\epsilon = \left\{ \boldsymbol{x} \in S^{d-1} : \left|\langle \boldsymbol{u}, \boldsymbol{x} \rangle\right| \leq \epsilon \right\}$. By definition, $\bigcap_{\boldsymbol{u} \in T} Y_\epsilon = S^{d-1} \setminus \bigcup_{\boldsymbol{u} \in T}(S^{d-1} \setminus Y_\epsilon)$. By applying the union bound on $\bigcup_{\boldsymbol{u} \in T}(S^{d-1} \setminus Y_\epsilon)$, we have the following inequality:

$$\mathrm{vol}((\bigcup_{\boldsymbol{u} \in T}(S^{d-1} \setminus Y_\epsilon)) \leq \sum_{\boldsymbol{u} \in T} \mathrm{vol}(S^{d-1} \setminus Y_\epsilon)$$

Using the technique in the notes, we define $\mathrm{vol}(S^{d-1} \setminus Y_\epsilon)$ as the points outside the 'tropics' which we define as $\bigcap_{\boldsymbol{u} \in T} Y_\epsilon$, hence the volume of the points outside the 'tropics' are bounded by the inequality shown below

$$\begin{aligned} \sum_{\boldsymbol{u} \in T} \mathrm{vol}(S^{d-1} \setminus Y_\epsilon) &\leq \sum_{\boldsymbol{u} \in T} 2(1-\epsilon^2)^{d/2} \,\mathrm{vol}(S^{d-1}) \\ &\leq \sum_{\boldsymbol{u} \in T} 2e^{-\epsilon(d-1)/2} \,\mathrm{vol}(S^{d-1}) \\ &\leq d^{100} 2e^{-\epsilon^2(d-1)/2} \,\mathrm{vol}(S^{d-1}) \end{aligned}$$

Substituting the bound into the $\bigcap_{\boldsymbol{u} \in T} Y_\epsilon$, we have the following bound:

$$\bigcap_{\boldsymbol{u} \in T} Y_\epsilon \geq (1 - d^{100} 2e^{-\epsilon^2(d-1)/2}) \,\mathrm{vol}(S^{d-1})$$

Since we want $\bigcap_{\boldsymbol{u}\in T} Y_\epsilon = 0.99\,\mathrm{vol}(S^{d-1})$, we construct the next equality demonstrating that:

$$\begin{aligned}(1 - d^{100}2e^{-\epsilon^2(d-1)/2})\,\mathrm{vol}(S^{d-1}) &= 0.99\,\mathrm{vol}(S^{d-1})\\ 100\log d + \log 2 - \epsilon^2\frac{d-1}{2} &= \log 0.01\\ 200\log d + 2\log(2/0.01) &= \epsilon^2(d-1)\\ 200\log(d) + 2\log(200) &= \epsilon^2 d \text{ as when } d >>> 1, d-1 \approx d\\ 200\log(d) &= \epsilon^2 d \text{ as when } d >>> 1, 200\log(d) \approx 200\log(d) + 2\log(200)\\ \sqrt(200)\sqrt(\log(d)/d) &= \epsilon^2\end{aligned}$$

Therefore $C = \sqrt(200)$.

b) Here, define $Z_\epsilon = \left\{\boldsymbol{x}\in S^{d-1} : \left|\langle \boldsymbol{u}, \boldsymbol{x}\rangle\right| > \epsilon\right\}$ and refer to Z_ϵ as the set of points outside the 'tropics'. Using the approximation obtained from the notes, we have the following inequality:

$$\begin{aligned}\mathrm{vol}(Z_\epsilon) &\le 2(1-\epsilon^2)^{d/2}\,\mathrm{vol}(S^{d-1})\\ &\le 2e^{-\epsilon^2(d-1)/2}\,\mathrm{vol}(S^{d-1})\end{aligned}$$

We want $\mathrm{vol}(Z_\epsilon) = 0.99\,\mathrm{vol}(S^{d-1})$ so set the bound obtained previously equal to $0.99\,\mathrm{vol}(S^{d-1})$ to obtain:

$$\begin{aligned}2e^{-\epsilon^2(d-1)/2}\,\mathrm{vol}(S^{d-1}) &= 0.99\,\mathrm{vol}(S^{d-1})\\ \epsilon^2(d-1) &= -2\log 0.99\\ \epsilon^2 d &= -2\log 0.99 \text{ as when } d >>> 1, d-1\approx d\\ \epsilon &= \sqrt{\frac{-2\log 0.99}{d}}\end{aligned}$$

Hence have $c = \sqrt{-2\log 0.99}$. $\square$

Problem 2 (25 points)**.** Let $B_1^d := \{\boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^d |x_i| \le 1\}$ denote the d-dimensional *cross polytope* (as explained in Ball's lecture notes).

(a) Prove that $B^d \subseteq \sqrt{d} B_1^d$.

(b) Use the fact $B^d \subseteq \sqrt{d} B_1^d$ to derive a bound on the volume of B^d of the form

$$\mathrm{vol}(B^d) \ \le \ c \cdot \left(\frac{c'}{d}\right)^{d/2}$$

for some positive constants $c, c' > 0$. Explain each step in your derivation.

Hint: Stirling's approximation implies $\sqrt{2\pi} n^{n+1/2} e^{-n} \le n! \le n^{n+1/2} e^{1-n}$ for all $n \in \mathbb{N}$.

Solution.

a) According to the notes, $B^d := \{\boldsymbol{x} \in \mathbb{R}^d : \sum_{i=1}^d ||x_i||_2 \le 1\}$, so it follows that $(B^d)^2 = \sum_{i=1}^d x_i^2$. We expand $(B_1^d)^2$ to obtain $\sum_{i=1}^d \sum_{j=1}^d |x_i||x_j| = \sum_{i=1}^d x_i^2 + 2\sum_{i \ne j} |x_i||x_j|$. By comparing the above, we see that $\sum_{i=1}^d x_i^2 \le \sum_{i=1}^d x_i^2 + 2\sum_{i \ne j} |x_i||x_j||$ as the second term ≥ 0 which shows $(B^d)^2 \le (B_1^d)^2$ as all terms are ≥ 0. Multiplying the RHS with a constant d that is ≥ 1 and taking roots on both sides, the inequality remains valid as the squared values are ≥ 1. Hence $B^d \subseteq \sqrt{d} B_1^d$.

b) From part a), we know that $B^d \subseteq \sqrt{d} B_1^d$. The result implies that $\mathrm{vol}(B^d) \le \mathrm{vol}\, B_1^d$. The volume of B_1^d is $\frac{2^d}{d!}$.

$$\mathrm{vol}\, B^d \le \frac{2^d}{d!}$$

Applying Stirling's approximation to the denominator and collecting relevant terms

$$\mathrm{vol}\, B^d \le \frac{2^d}{\sqrt{2\pi} d^{d+1/2} e^{-d}} = \frac{2^{d/2+d/2} \cdot e^{d/2+d/2}}{\sqrt{2\pi} d^{d+1/2}} = \frac{2^{d/2} \cdot e^{d/2}}{\sqrt{2\pi} d^d} \cdot \frac{2e}{d}^{d/2}$$

As $d > 0$, we know all RHS terms are greater than 0. This gives us $c = \frac{2^{d/2} \cdot e^{d/2}}{\sqrt{2\pi} d^d}$ and $c' = 2e$.

$\square$

Problem 3 (25 points)**.** Let X be an $[a, b]$-valued random variable (i.e., $\mathbb{P}(X \in [a, b]) = 1$) with $\mathbb{E}(X) = 0$. For simplicity, assume X has a probability density function f. In this problem, you will prove $\psi_X(\lambda) \leq \lambda^2(b-a)^2/8$ using a technique due to McAllester and Ortiz (2003).

(a) Consider the family of density functions $\{g_\lambda : \lambda \in \mathbb{R}\}$, where

$$g_\lambda(x) \ := \ \frac{e^{\lambda x}}{\mathbb{E}e^{\lambda X}} f(x) \quad \text{for all } x \in \mathbb{R}\,.$$

Briefly verify that if $Y_\lambda \sim g_\lambda$, then

$$\begin{aligned} \mathbb{E}(Y_\lambda) \ &= \ \psi'_X(\lambda)\,, \\ \operatorname{var}(Y_\lambda) \ &= \ \psi''_X(\lambda)\,, \end{aligned}$$

where ψ'_X is the first-derivative of ψ_X, and ψ''_X is the second-derivative of ψ_X. (You don't need to write much at all for this part.)

(b) Prove that any $[a, b]$-valued random variable has variance at most $(b-a)^2/4$.

(c) The fundamental theorem of calculus implies

$$\psi_X(\lambda) \ = \ \int_0^\lambda \int_0^\mu \psi''_X(\gamma)\,\mathrm{d}\gamma\,\mathrm{d}\mu\,.$$

Use this identity and the results of parts (a) and (b) to prove that $\psi_X(\lambda) \leq \lambda^2(b-a)^2/8$.

Solution.

a) We start by setting out $\mathbb{E}X$ and $\operatorname{var} X$ with $X \in [a, b]$:

$$\begin{aligned} \mathbb{E}Y_\lambda &= \int_a^b x \frac{e^{\lambda x}}{\mathbb{E}e^{\lambda X}} f(x)dx \\ \operatorname{var} Y_\lambda &= \mathbb{E}Y_\lambda^2 - (\mathbb{E}Y_\lambda)^2 \end{aligned}$$

Have $\psi_X \lambda = \log \mathbb{E}e^{\lambda x} = \int_a^b \log e^{\lambda x} f(x)dx$. Differentiate it twice:

$$\begin{aligned} \psi'_X(\lambda) &= \int_a^b \frac{d}{d\lambda} \log e^{\lambda x} f(x)dx \\ &= \int_a^b \frac{xe^{\lambda x}}{\mathbb{E}e^{\lambda X}} f(x)dx = \mathbb{E}(Y_\lambda) \\ \psi''_X(\lambda) &= \int_a^b \frac{d}{d\lambda} \frac{xe^{\lambda x}}{\mathbb{E}e^{\lambda X}} f(x)dx \\ &= \int_a^b \frac{x^2e^{\lambda x}}{\mathbb{E}e^{\lambda X}} f(x)dx - \int_a^b \frac{xe^{\lambda x}}{(\mathbb{E}e^{\lambda X})^2} f(x)dx * \int_a^b xe^{\lambda x} f(x)dx \\ &= \mathbb{E}Y_\lambda^2 - (\int_a^b x \frac{e^{\lambda x}}{\mathbb{E}e^{\lambda X}} f(x)dx)^2 \text{ as } \mathbb{E}e^{\lambda X} \text{ is a constant} \\ &= \mathbb{E}Y_\lambda^2 - (\mathbb{E}Y_\lambda)^2 \end{aligned}$$

b) Suppose we have a distribution with support $\in [a,b]$ is that with $P(Z=a)=P(Z=b)=0.5$. The above distribution has $\mathbb{E}(Z)=\frac{a+b}{2}$ and $\mathbb{E}(Z^2)=\frac{a^2+b^2}{2}$. Hence $\text{var}(Z)=\frac{a^2+b^2}{2}-(\frac{a+b}{2})^2=\frac{2a^2+2b^2-a^2-b^2-2ab}{4}=\frac{a^2+b^2-2ab}{4}=\frac{(b-a)^2}{4}$.

To prove that the above distribution has the largest variance, we consider distributions with less concentrated point densities. If we take 0.5ϵ (where ϵ is some small number) away from a and b and place it at the point $\frac{a+b}{2}$, $\text{var}(Z')$ of the new distribution Z' decreases. This is due to $\min(X_i-\mathbb{E}X)<\max(X_i-\mathbb{E}X)=\frac{b-a}{2}$ giving $\text{var}(Z')=p(\max(X_i-\mathbb{E}X))^2+(1-p)(\frac{b-a}{2})^2\leq\text{var}(Z)=(\frac{b-a}{2})^2$. Hence, as soon as we spread probability mass around to other discrete points, the variance $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mathbb{E}X_i)^2$ deceases and will be $\leq\frac{(b-a)^2}{4}$. By continuing to spread probability mass to all $X\in[a,b]$ we end up with a continuous distribution Z'' with probability mass is even more diffuse than before so var(Z") will be even smaller than before.

In the case of having more mass on b than a in distribution Z''', $\text{var}(Z''')=p(b-(pb+(1-p)a))^2+(1-p)(a-(pb+(1-p)a))^2$ where $p>0.5+\epsilon$. Expanding var Z''', we get

$$\begin{aligned}
p(b-(pb+(1-p)a))^2+(1-p)(a-(pb+(1-p)a))^2 &= p((1-p)(b-a))^2+(1-p)(p(a-b))^2\\
&= p(1-p)^2(b-a)^2+(1-p)(-p)^2(b-a)^2\\
&= (p-2p^2+p^3+p^2-p^3)(b-a)^2\\
&= p(1-p)(b-a)^2
\end{aligned}$$

Since we know that the max for $p(1-p)$ is $1/4$ when $p=0.5$, therefore the variance of any distribution of point masses on a and b is $\leq\frac{(b-a)^2}{4}$.

Therefore, we can see that any $[a,b]$-valued random variable has variance at most $(b-a)^2/4$.

c) Substituting the results from a) and b) into $\psi_X\lambda=\int_0^\lambda\int_0^\mu\psi''_X(\gamma)\,\mathrm{d}\gamma\,\mathrm{d}\mu$, we have:

$$\begin{aligned}
\psi_X(\lambda) &= \int_0^\lambda\int_0^\lambda\psi''_X(\gamma)\,\mathrm{d}\gamma\,\mathrm{d}\mu\\
&\leq\int_0^\lambda\int_0^\lambda\frac{(b-a)^2}{4}\,\mathrm{d}\gamma\,\mathrm{d}\mu\\
&\leq\int_0^\lambda\lambda\frac{(b-a)^2}{4})\,\mathrm{d}\gamma\\
&\leq\lambda^2\frac{(b-a)^2}{8}
\end{aligned}$$

$\square$

Problem 4 (25 points)**.** Let $\boldsymbol{U}$ be a random unit vector with the uniform density on S^{d-1}, and let $X := \langle \boldsymbol{v}, \boldsymbol{U} \rangle$, where $\boldsymbol{v}$ is a fixed unit vector in S^{d-1}.

(a) Prove that $\psi_{X^2 - \mathbb{E}(X^2)}(\lambda) \leq \psi_{Z^2-1}(\lambda/d)$ for all $\lambda \in \mathbb{R}$, where $Z \sim \mathrm{N}(0,1)$.

Hint: You may use the fact that if $Y_d \sim \chi^2(d)$ and $\boldsymbol{U}$ are independent, then $\sqrt{Y_d}\boldsymbol{U} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$ (standard multivariate Gaussian in $\mathbb{R}^d$). Jensen's inequality may also be useful.

(b) Use the result of part (a) to derive a tail bound for $\left|X^2 - \mathbb{E}(X^2)\right|$. Explain each step in your derivation.

Solution. □

Problem 5 (25 points)**.** Let $\Phi \colon \mathbb{R} \to [0,1]$ denote the cumulative distribution function for $\mathrm{N}(0,1)$, i.e., $\Phi(t) = \mathbb{P}(Z \le t)$ where $Z \sim \mathrm{N}(0,1)$. Prove that for any $d \in \mathbb{N}$, if

1. $X_1, X_2, \ldots, X_d$ are independent random variables;
2. $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 = 1$ for all $i \in [d]$;

then for a $1 - o(1)$ fraction of unit vectors $\boldsymbol{u} \in S^{d-1}$, the random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_d)$ satisfies

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\langle \boldsymbol{u}, \boldsymbol{X} \rangle \le t\right) - \Phi(t) \right| \;\le\; O\left(\frac{\rho}{d^{0.49}}\right),$$

where $\rho := \max_{i\in[d]} \mathbb{E}|X_i|^3$.

You can use the following theorem (which you do not need to prove):

> **Theorem 1** (Berry-Esséen theorem)**.** *There is an absolute positive constant $C > 0$ such that the following holds. Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with $\mathbb{E}Y_i = 0$, $\sigma_i^2 := \mathbb{E}Y_i^2 < \infty$. Define $v_n := \sum_{i=1}^n \sigma_i^2$ and $\rho_n := \sum_{i=1}^n \mathbb{E}|Y_i|^3$. Then*
>
> $$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\frac{Y_1 + Y_2 + \cdots + Y_n}{\sqrt{v_n}} \le t\right) - \Phi(t) \right| \;\le\; \frac{C\rho_n}{v_n^{3/2}}.$$

Solution.

We compare $\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\langle \boldsymbol{u}, \boldsymbol{X} \rangle \le t\right) - \Phi(t) \right| \;\le\; O\left(\frac{\rho}{d^{0.49}}\right)$ and Berry-Esséen and observe that they are the same apart from the inner probability terms and the bounds. We can use Berry-Esséen to prove the above inequality if the theorem's assumptions holds.

Since we know $\boldsymbol{X}$ that are independent random variables and $\boldsymbol{u}$ is a fixed unit vector, hence the inner product of the two are also independent random variables. If we let Y_i be the inner product of $\boldsymbol{u}$ and $\boldsymbol{X}$, $\mathbb{E}Y_i = \mathbb{E}[u^T X_i] = u^T \mathbb{E}X_i = 0$ and $\mathbb{E}Y_i^2 = \mathbb{E}[u^T X_i^T X_i u^T] = u^2 \mathbb{E}[X^2] = 1 * 1 = 1$ ($u^T u = 1$ since u is a unit vector). Armed with the above facts, we know that $\langle \boldsymbol{u}, \boldsymbol{X} \rangle$ is a valid Y_i as it fulfils all the assumptions required for Berry-Esséen to hold.

By definition, we have $v_n = \sum_{i=1}^d 1 = d$ and $\rho_n = \sum_{i=1}^n \mathbb{E}|Y_i|^3 \le d \max_{i\in[d]} \mathbb{E}|X_i|^3$. Therefore $\frac{C\rho_n}{v_n^{3/2}} = \frac{d\rho}{d^{3/2}} = \frac{\rho}{d^{1/2}} = O(\frac{\rho}{d^{0.49}})$ which proves $\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left(\langle \boldsymbol{u}, \boldsymbol{X} \rangle \le t\right) - \Phi(t) \right| \;\le\; O\left(\frac{\rho}{d^{0.49}}\right)$

□

References

D. McAllester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911, 2003.