ECE 434: Problem Set 1

Due on October 1, 2023

Professor Chi Jin

Evan Dogariu

Collaborators: None

Problem 1

Recall that a random variable X is sub-Gaussian with parameter σ^2 (denoted as $X \in \mathrm{SG}(\sigma^2)$) if:

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E} e^{\lambda(X - \mathrm{E}X)} \leq e^{\lambda^2 \sigma^2 / 2}$$

Prove the following:

(a) Gaussian is a subclass of sub-Gaussian: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $X \in \mathrm{SG}(\sigma^2)$.

(b) Sub-Gaussian random variables have good tail bound: if $X \in \mathrm{SG}(\sigma^2)$ then for any $t \geq 0$,

$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq e^{-\frac{t^2}{2\sigma^2}} \quad \text{and} \quad \mathbb{P}[X - \mathbb{E}X \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}$$

(d) Sum of independent sub-Gaussian random variables is also sub-Gaussian: if $X_1, \ldots, X_n$ are independent, and for any $i \in [n]$, $X_i \in \mathrm{SG}(\sigma_i^2)$, then $\sum_{i=1}^n X_i \in \mathrm{SG}\left(\sum_{i=1}^n \sigma_i^2\right)$.

(e) Use above results to prove the concentration inequality for the sum of independent sub-Gaussians.

Theorem 1. *Suppose $\{X_i\}_{i=1}^n$ are independent sub-Gaussian random variables with parameters $\{\sigma_i^2\}_{i=1}^n$. Then, for any $t \geq 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right] \leq e^{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}}.$$

Solution

Proof. (a) Let $X \sim \mathcal{N}(\mu, \sigma^2)$; then, $(X - \mathbb{E}X) \sim \mathcal{N}(\mu, \sigma^2)$. For any $\lambda \in \mathbb{R}$, we can explicitly compute

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} = \int_{z \in \mathbb{R}} e^{\lambda z} \frac{e^{-\frac{z^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dz = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{z \in \mathbb{R}} e^{-\left(\frac{z^2}{2\sigma^2} - \lambda z\right)} dz$$

We may complete the square to see that

$$\frac{z^2}{2\sigma^2} - \lambda z = \frac{z^2}{2\sigma^2} - \lambda z + \frac{\lambda^2\sigma^2}{2} - \frac{\lambda^2\sigma^2}{2} = \left(\frac{z}{\sigma\sqrt{2}} - \lambda\frac{\sigma}{\sqrt{2}}\right)^2 - \frac{\lambda^2\sigma^2}{2} = \frac{(z - \lambda\sigma^2)^2}{2\sigma^2} - \frac{\lambda^2\sigma^2}{2},$$

and so

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} = e^{-\lambda^2\sigma^2/2} \int_{z \in \mathbb{R}} \frac{e^{-\frac{(z - \lambda\sigma^2)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} dz$$

We recognize this integrand to simply be the pdf of the distribution $\mathcal{N}(z - \lambda\sigma^2, \sigma^2)$, and so the integral evaluates to 1. Therefore,

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} = e^{-\lambda^2\sigma^2/2},$$

which satisfies the criterion for being sub-Gaussian.

(b) Suppose that $X \in \mathrm{SG}(\sigma^2)$ and $t \geq 0$. Then, for all $\lambda > 0$ we know

$$\begin{aligned}
\mathbb{P}[X - \mathbb{E}X \geq t] &= \mathbb{P}[e^{\lambda(X - \mathbb{E}X)} \geq e^{\lambda t}] \\
&\leq \frac{\mathbb{E}e^{\lambda(X - \mathrm{E}X)}}{e^{\lambda t}} \\
&\leq \frac{e^{\lambda^2\sigma^2/2}}{e^{\lambda t}} = e^{\frac{\lambda^2\sigma^2}{2} - \lambda t},
\end{aligned}$$

where the first equality is by the monotonicity of $e^{\lambda \cdot}$, the first inequality is an application of Markov's inequality, and the second inequality applies the criterion for being sub-Gaussian. Since this holds for all $\lambda > 0$, we may apply it with the minimal λ^* given by

$$\lambda^* = \text{argmin}_{\lambda > 0} \left\{ \frac{\lambda^2 \sigma^2}{2} - \lambda t \right\} = \frac{t}{\sigma^2},$$

where we used the fact that for a parabola $ax^2 + bx + c$ with $a > 0$, the minimum occurs at the vertex $\frac{-b}{2a}$. Plugging this value of λ^* in, we get a concentration bound of

$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq e^{\frac{t^2}{2\sigma^2} - \frac{t^2}{\sigma^2}} = e^{-\frac{t^2}{2\sigma^2}}$$

To show the reverse bound, note that if $X \in \text{SG}(\sigma^2)$ then so too is $Y := -X$. Applying the result we just had with Y in place of X,

$$\mathbb{P}[X - \mathbb{E}X \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}$$

However, the event that $Y - \mathbb{E}Y \geq t$ is equivalent to the event $(-X) - \mathbb{E}[-X] \geq t \iff X - \mathbb{E}X \leq -t$, and so we get the other tail bound

$$\mathbb{P}[X - \mathbb{E}X \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}$$

(d) Let $X_i \in \text{SG}(\sigma_i^2)$ be independent, and let $X := \sum_{i=1}^n X_i$ be the random variable denoting their sum. Then, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} = \mathbb{E}\exp\left\{ \sum_{i=1}^n \lambda\left(X_i - \mathbb{E}X_i\right) \right\} = \mathbb{E}\prod_{i=1}^n \exp\{\lambda(X_i - \mathbb{E}X_i)\},$$

where we used the linearity of expectation for the first equality and the properties of exponents for the second. Now, since the X_i's are independent, the above expectation of a product simplifies to a product over expectations, and so we see that

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} = \prod_{i=1}^n \mathbb{E}e^{\lambda(X_i - \mathbb{E}X_i)} \leq \prod_{i=1}^n e^{\lambda^2 \sigma_i^2 / 2} = e^{\frac{\lambda^2}{2}\sum_{i=1}^n \sigma_i^2}$$

From this, we see that $X \in \text{SG}(\sum_{i=1}^n \sigma_i^2)$ as desired.

(e) The above results give us the tools to prove Theorem 1. Let $\{X_i\}_{i=1}^n$ be independent sub-Gaussian random variables with parameters $\{\sigma_i\}_{i=1}^n$. Then, for any $t \geq 0$ we know that

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right] = \mathbb{P}\left[\left(\sum_{i=1}^n X_i - \mathbb{E}\sum_{i=1}^n X_i\right) \geq t\right] \leq e^{-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}},$$

where we used the fact that $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \sigma_i^2)$-sub-Gaussian from (d) along with the concentration bound from (b) for the above inequality. ■

Problem 4

Let $X_1, \dots, X_n$ be independent real-valued random variables i.i.d. sampled from the same underlying distribution with probability density function f. We can use the following kernel density estimator $\hat{f}$ to estimate the unknown density f from data $X_1, \dots, X_n$:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Here $h > 0$ is a smoothing parameter, and K is a non-negative function that satisfies $\int K(x)dx = 1$. We measure the error in density estimation using total variation distance:

$$Z := \int \left|\hat{f}(x) - f(x)\right| dx$$

Use the bounded difference concentration inequality to prove that Z concentrates around its expectation $\mathbb{E}Z$.

Solution

Proof. We will prove the bounded difference criterion for Z, viewed as a function of the n random variables. To do so, we look at the difference when replacing X_k with X_k'; to this end, let $\hat{f}'$ denote the resulting kernel density estimator (not the derivative!), and let Z' denote the resulting total variation distance. Then,

$$\begin{aligned}
|Z' - Z| &= \left|\int \left|\hat{f}'(x) - f(x)\right| dx - \int \left|\hat{f}(x) - f(x)\right| dx\right| \\
&\leq \int \left|\left|\hat{f}'(x) - f(x)\right| - \left|\hat{f}(x) - f(x)\right|\right| dx \\
&\leq \int |\hat{f}'(x) - \hat{f}(x)| dx,
\end{aligned}$$

where the first inequality applies the triangle inequality for integrals, and the second applies the reverse triangle inequality. Now, we note that for all $x \in \mathbb{R}$, by the triangle inequality and the fact that K is nonnegative we have

$$|\hat{f}'(x) - \hat{f}(x)| = \frac{1}{nh}\left|K\left(\frac{x - X_k'}{h}\right) - K\left(\frac{x - X_k}{h}\right)\right| \leq \frac{1}{nh}\left(K\left(\frac{x - X_k'}{h}\right) + K\left(\frac{x - X_k}{h}\right)\right)$$

Thus, we have that

$$\begin{aligned}
|Z' - Z| &\leq \frac{1}{nh}\int \left(K\left(\frac{x - X_k'}{h}\right) + K\left(\frac{x - X_k}{h}\right)\right) dx \\
&= \frac{1}{nh}\int K\left(\frac{x - X_k'}{h}\right) dx + \frac{1}{nh}\int K\left(\frac{x - X_k}{h}\right) dx \\
&= \frac{h}{nh}\int K(u')du' + \frac{h}{nh}\int K(u)du \\
&= \frac{2}{n},
\end{aligned}$$

where for the third line we used the substitutions $u' = \frac{x - X_k'}{h}$ and $u = \frac{x - X_k}{h}$. Since this holds for any replacement of X_k with X_k', we find that Z satisfies the bounded differences criterion with $c_k = \frac{2}{n}$ for all k. Thus, by Theorem 3 from Lecture 1, for all $t \geq 0$ we know that

$$\mathbb{P}[Z - \mathbb{E}Z \geq t] \leq e^{-\frac{2t^2}{\sum_{k=1}^{n} c_k^2}} = e^{-\frac{nt^2}{2}}$$

■

Problem 5

Recall the definition of Rademacher average of a set $A \subseteq \mathbb{R}^n$:

$$\tilde{R}_n(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \epsilon_i a_i \right|$$

Denote $B_p := \{x \in \mathbb{R}^n : \ \|x\|_p \leq 1\}$ as the p-norm unit ball in $\mathbb{R}^n$. Prove the following:

(a) $\tilde{R}_n(B_2) = \frac{1}{\sqrt{n}}$

(b) $\tilde{R}_n(B_1) = \frac{1}{n}$ and $\tilde{R}_n(B_\infty) = 1$

Solution

Proof. (a) For any value ϵ that the Rademacher random variable can take, the value inside the supremum will be maximized when a points in the same direction. When this happens, we will have that $a = \frac{\epsilon}{\|\epsilon\|_2}$, and so

$$\sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \epsilon_i a_i \right| = \frac{1}{n} \cdot \frac{\|\epsilon\|_2^2}{\|\epsilon\|_2} = \frac{\|\epsilon\|_2}{n} = \frac{\sqrt{n}}{n} = \frac{1}{\sqrt{n}},$$

where for the second to last equality we used that a Rademacher random variable *always* has 2-norm of $\sqrt{n}$ since its coordinates always have magnitude 1. Since this holds for all possible values of ϵ, we may simply take the expectation over Rademacher random variables to get the result.

(b) We apply similar reasoning as above. For any ϵ value that the Rademacher RV can take, the value inside the supremum will be attained when a points in the same direction. As such, for the 1-norm case we will have $a = \frac{\epsilon}{\|\epsilon\|_1}$, and so

$$\sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \epsilon_i a_i \right| = \frac{1}{n} \cdot \frac{\|\epsilon\|_2^2}{\|\epsilon\|_1} = \frac{1}{n} \cdot \frac{n}{n} = \frac{1}{n},$$

where we used that a Rademacher random variable *always* has 2-norm of $\sqrt{n}$ and 1-norm of n since its coordinates always have magnitude 1. Taking an expectation over this constant value gives the result for the 1-norm. In the ∞-norm case, we apply the same logic again. In particular, the dot product will be maximized when $a = \frac{\epsilon}{\|\epsilon\|_\infty}$, and so

$$\sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \epsilon_i a_i \right| = \frac{1}{n} \cdot \frac{\|\epsilon\|_2^2}{\|\epsilon\|_\infty} = \frac{1}{n} \cdot \frac{n}{1} = 1,$$

where this time we used that $\|\epsilon\|_\infty = 1$ for every value that a Rademacher RV can take. Once again, taking an expectation over this value yields the desired result for the ∞-norm. ■

Problem 6

In lecture, we upper bound the excess risk by Rademacher complexity. In fact, the use of Rademacher random variable is not necessary, we can similarly define Gaussian complexity. Here we define the set version: let $\{g_i\}_{i=1}^n$ be i.i.d. standard Gaussian random variables $N(0,1)$, we define

$$\widetilde{G}_n(A) = \mathbb{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^n g_i a_i \right|$$

Prove that

$$\widetilde{R}_n(A) \leq \sqrt{\frac{\pi}{2}} \cdot \widetilde{G}_n(A)$$

Therefore, up to a constant factor, the excess risk can also be upper-bounded by Gaussian complexity plus a small concentration term.

Solution

Proof. To accomplish this, we notice that the Gaussian distribution is symmetric about 0, and so the random variables denoting the random vector of signs of the coordinates and the random vector of magnitudes of the coordinates are indeed independent. More precisely, let $M : \mathbb{R}^n \to \mathbb{R}^n$ be the magnitude map sending $(x_1, \dots, x_n) \mapsto (|x_1|, \dots, |x_n|)$ and $S : \mathbb{R}^n \to \{-1, 1\}^n$ be the sign map sending $(x_1, \dots, x_n) \mapsto (\text{sign}(x_1), \dots, \text{sign}(x_n))$ with the convention that $\text{sign}(0) = 1$. Then, for a random n-dimensional distribution D, let $M(D)$ denote the induced distribution on $\mathbb{R}^n$ via the magnitudes of coordinates of $X \sim D$, and similarly let $S(D)$ denote the induced distribution on $\mathbb{R}^n$ via the signs of coordinates of $X \sim D$.

Then, it is a property of any rotationally symmetric distribution D that

$$\mathbb{E}_{x \sim D}[f(x)] = \mathbb{E}_{m \sim M(D)} \left[\mathbb{E}_{s \sim S(D)}[f(m,s) \,|\, m] \right] = \mathbb{E}_{m \sim M(D)} \left[\mathbb{E}_{s \sim S(D)}[f(m,s)] \right],$$

where we used rotational symmetry to say that s and m are independent, and so $\mathbb{E}_{s \sim S(D)}[f(m,s) \,|\, m] = \mathbb{E}_{s \sim S(D)}[f(m,s)]$ for all m. Note that we are allowed to write $f(m,s)$ since one can always recreate x from m and s via $x = m \odot s$. The last cute observation that we will need is that, when D is rotationally symmetric and each coordinate is i.i.d., then the induced distribution $S(D)$ is precisely the Rademacher distribution (for each coordinate, there is an equal chance of being positive or negative and a 0 chance of being 0, and so the coordinate is 1 w.p. $\frac{1}{2}$ and -1 w.p. $\frac{1}{2}$).

The above observations allow us to see that if $D = \mathcal{N}(0, I_n)$ is the standard multivariate Gaussian, then

$$\tilde{G}_n(A) = \mathbb{E}_{g \sim D} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^n g_i a_i \right| = \mathbb{E}_{m \sim M(D)} \mathbb{E}_{s \sim S(D)} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^n m_i s_i a_i \right|$$

For any m, let us denote

$$\Psi(m) := \mathbb{E}_{s \sim S(D)} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^n m_i s_i a_i \right|$$

Then, Ψ is a convex function of m since it is an integral over a supremum of convex functions of m. Also, $\tilde{G}_n(A) = \mathbb{E}_{m \sim M(D)} \Psi(m)$. This sets us up perfectly for Jensen's inequality, which reveals that

$$\mathbb{E}_{m \sim M(D)} \Psi(m) \geq \Psi(\mathbb{E}_{m \sim M(D)}[m])$$

We recognize that for each coordinate $g_i \sim \mathcal{N}(0,1)$, the expectation of $|g_i|$ is $\sqrt{\frac{2}{\pi}}$. So, $\mathbb{E}_{m\sim M(D)}[m] = \left(\sqrt{\frac{2}{\pi}}, \dots, \sqrt{\frac{2}{\pi}}\right)$ since each coordinate is drawn i.i.d.. Therefore, relabeling $s \sim S(D)$ to $\epsilon \sim \text{Rad}(1/2)$ to emphasize that $S(D)$ is the Rademacher distribution, we find that

$$\begin{aligned}
\tilde{G}_n(A) \geq \Psi\left(\sqrt{\frac{2}{\pi}}, \dots, \sqrt{\frac{2}{\pi}}\right) &= \mathbb{E}_{\epsilon\sim\text{Rad}(1/2)} \sup_{a\in A} \frac{1}{n}\left|\sum_{i=1}^n \sqrt{\frac{2}{\pi}}\epsilon_i a_i\right| \\
&= \sqrt{\frac{2}{\pi}} \cdot \mathbb{E}_{\epsilon\sim\text{Rad}(1/2)} \sup_{a\in A} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i a_i\right| \\
&= \sqrt{\frac{2}{\pi}}\tilde{R}_n(A)
\end{aligned}$$

Rearranging,

$$\tilde{R}_n(A) \leq \sqrt{\frac{\pi}{2}}\tilde{G}_n(A)$$

as desired. From here, we may conclude with a bound on the excess risk for $g \in \mathcal{G}$ with bounded images ($\text{range}(g) \subseteq [0, B]$) using Proposition 1 and Theorem 2 from Lecture 2. This tells us that for for $g \in \mathcal{G}$ with bounded images ($\text{range}(g) \subseteq [0, B]$) and any $\delta > 0$, with probability at least $1 - \delta$ we have

$$\text{excess risk} \leq 2\sqrt{2\pi} \cdot G_n(\mathcal{G}) + B\sqrt{\frac{2\log(1/\delta)}{n}}$$

■