High Dimensional Probability and Statistics **2nd Semester, 2022-2023**

Lecture 2: Herbst Argument and Entropy Method

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2023/03/19)*

Recap and Motivation: In Lecture 1 we have discussed the sub-Gaussian and sub-exponential distributions and the corresponding tail bounds for sums of independent random variables and functions satisfying the bounded different property. Our next goal is to extend the concentration results to other interesting functions. We will restrict our attention to the sub-Gaussian type tails while some of the techniques may also be applicable for establishing the Bernstein type bound.

Define the log-moment (or cumulant) generating function of a random variable X as

$$\psi(\lambda) = \log \mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}[X])\right)\right]. \tag{2.1}$$

The sub-Gaussian property can be equivalently expressed as

$$\psi(\lambda) \lesssim \lambda^2\nu^2 \quad \text{for all} \quad \lambda \in \mathbb{R}. \tag{2.2}$$

By the Chernoff bound we know that the sub-Gaussian property immediately implies a Gaussian tail bound (they are indeed equivalent). Moreover, the sub-Gaussian property can be established for sums of independent random variables and functions obeying the bounded difference inequality. As already seen, the proofs rely essentially on the tensorization property (or a martingale difference sequence variant) of the log-moment generating function defining the sub-Gaussian property, i.e.,

$$\log \mathbb{E}\left[\exp\left(\lambda \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k])\right)\right] \le \sum_{k=1}^{n} \log \mathbb{E}\left[\exp\left(\lambda\left(X_k - \mathbb{E}[X_k]\right)\right)\right].$$

However, for more complicated functions $f(X_1, \cdots, X_n)$ arising from the applications than sums of independent random variables, the above tensorization property hardly holds. That is, the sub-Gaussian property in terms of the (log-)moment generating function overall does not tensorize well. To mitigate this issue, one idea is to introduce an alternative formulation of the sub-Gaussian property that behaves well under tensorization.

In this lecture we will study the sub-Gaussian property based on certain entropy function and establish a concentration inequalities for more general f. To motivate this, let us recap the calculus method that is used in the proof of the sub-Gaussian property for bounded random variables. First, a simple calculation yields that

$$\psi(0) = 0 \quad \text{and} \quad \psi'(0) = 0.$$

Thus in order to establish the sub-Gaussian property (2.2), it suffices to show that

$$\psi''(\lambda) \lesssim \nu^2 \quad \text{for all} \quad \lambda \in \mathbb{R}.$$

Noting that (2.2) is equivalent to

$$\psi(\lambda)/\lambda \lesssim \lambda\nu^2 \quad \text{for all} \quad \lambda \in \mathbb{R},$$

it also suffices to show that

$$\frac{d}{d\lambda}\left(\psi(\lambda)/\lambda\right) \lesssim \nu^2. \tag{2.3}$$

Though this is a trivial reformulation, it will lead to a more powerful method for proving concentration inequalities. Moreover, it turns out that (2.3) can be related to a type of entropy function that tensorizes well.

Agenda:

- Herbst argument and Tensorization
- Modified log-Sobolev inequality and Entropy method
- Gaussian concentration

2.1 Herbst Argument and Tensorization

2.1.1 Entropy

Definition 2.1 *The entropy of a* **nonnegative** *random variable Z, denoted* $\mathrm{Ent}\left[Z\right]$*, is defined as*

$$\mathrm{Ent}\left[Z\right] = \mathbb{E}\left[\phi(Z)\right] - \phi(\mathbb{E}\left[Z\right]) = \mathbb{E}\left[Z\log\left(Z\right)\right] - \mathbb{E}\left[Z\right]\log\left(\mathbb{E}\left[Z\right]\right),$$

where $\phi(t) = t\log t$.

Remark 2.2 *Note that the entropy defined here should not be confused with the Shannon entropy which is roughly about on average how many bits are needed to store a random variable.*

Exercise 2.3 *Show that $\phi(t) = t\log t$ is a convex function and thus* $\mathrm{Ent}\left[Z\right] \geq 0$.

Remark 2.4 *Given any convex function $\phi(t)$, we can define the Bregman distance (divergence) as*

$$D(y\|x) = \phi(y) - \phi(x) - \phi'(x)(y-x).$$

With this notion, it is easy to see that

$$\mathrm{Ent}\left[Z\right] = \mathbb{E}\left[D(Z\|\mathbb{E}\left[Z\right])\right] \tag{2.4}$$

for $\phi(t) = t\log t$. That is, $\mathrm{Ent}\left[Z\right]$ *is the average Bregman distance between Z and $\mathbb{E}\left[Z\right]$. Moreover, by simple calculus, one has*

$$\mathrm{Ent}\left[Z\right] = \inf_{t>0}\mathbb{E}\left[D(Z\|t)\right]. \tag{2.5}$$

Note that the definition of entropy in (2.4) *is overall similar to that of variance,*

$$\mathrm{Var}\left[Z\right] = \mathbb{E}\left[\left(Z - \mathbb{E}\left[Z\right]\right)^2\right],$$

but with a different metric. Thus, it is reasonable that entropy can tensorize well like variance.

Example 2.5 (Entropy of exponential of Gaussian) *Let $X \sim \mathcal{N}(0, \sigma^2)$. We have*

$$\begin{aligned}\text{Ent}\left[e^{\lambda X}\right] &= \mathbb{E}\left[e^{\lambda X}\log\left(e^{\lambda X}\right)\right] - \mathbb{E}\left[e^{\lambda X}\right]\log\left(\mathbb{E}\left[e^{\lambda X}\right]\right) \\ &= \mathbb{E}\left[\lambda X e^{\lambda X}\right] - \mathbb{E}\left[e^{\lambda X}\right]\log\left(e^{\frac{\lambda^2\sigma^2}{2}}\right) \\ &= \frac{1}{2}\lambda^2\sigma^2\mathbb{E}\left[e^{\lambda X}\right] \quad \textit{for all} \quad \lambda \in \mathbb{R},\end{aligned}$$

where we can use $d\mathbb{E}\left[e^{\lambda X}\right]/d\lambda = \mathbb{E}\left[Xe^{\lambda X}\right]$ to calculate the first term in the second line.

Exercise 2.6 *Show that* $\text{Ent}\left[e^{\lambda(X+c)}\right] = e^{\lambda c}\cdot\text{Ent}\left[e^{\lambda X}\right]$ *for any* $c \in \mathbb{R}$.

Lemma 2.7 (Entropy of exponential and MGF) *Let $\psi(\lambda)$ be the log-moment generating function defined in* (2.1)*. We have*

$$\frac{\text{Ent}\left[e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} = \lambda\psi'(\lambda) - \psi(\lambda).$$

Proof: The result follows from the definition and $\mathbb{E}\left[Xe^{\lambda X}\right] = \frac{d}{d\lambda}\mathbb{E}\left[e^{\lambda X}\right]$. Note that X is not necessarily mean zero though it is centered when defining $\psi(\lambda)$. ■

Example 2.8 (Entropy of exponential of bounded random variable) *Let X be mean zero and supported on $[a, b]$. By Lemma 2.7, we have*

$$\begin{aligned}\frac{\text{Ent}\left[e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} &= \lambda\psi'(\lambda) - \psi(\lambda) = [\lambda\psi'(\lambda) - \psi(\lambda)] - [0\cdot\psi'(0) - \psi(0)] \\ &= \int_0^\lambda \xi\psi''(\xi)d\xi \\ &\le \frac{(b-a)^2}{4}\int_0^\lambda \xi d\xi \\ &= \frac{\lambda^2(b-a)^2}{8},\end{aligned}$$

where the inequality follows from the bound for $\psi''(\xi)$, see Example 1.13 of Lecture 1. Thus,

$$\text{Ent}\left[e^{\lambda X}\right] \le \frac{\lambda^2(b-a)^2}{8}\mathbb{E}\left[e^{\lambda X}\right].$$

2.1.2 Herbst Argument

The last two examples reveal a connection between $\text{Ent}\left[e^{\lambda X}\right]$ and $\mathbb{E}\left[e^{\lambda X}\right]$ for certain sub-Gaussian random variables. It turns out this relation can be used to define the sub-Gaussian property, which follows from the Herbst argument.

Theorem 2.9 (Herbst) *Suppose that*

$$\text{Ent}\left[e^{\lambda X}\right] \le \frac{\lambda^2\nu^2}{2}\mathbb{E}\left[e^{\lambda X}\right] \quad \textit{for all} \quad \lambda \ge 0. \tag{2.6}$$

Then X satisfies the bound

$$\psi(\lambda) = \log\mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}\left[X\right])\right)\right] \le \frac{1}{2}\lambda^2\nu^2 \quad \textit{for all} \quad \lambda \ge 0. \tag{2.7}$$

Proof: The proof is indeed based on the argument sketched around (2.3). First note that

$$\lim_{\lambda \to 0} \frac{\psi(\lambda)}{\lambda} = 0. \qquad (\textbf{check this!})$$

Consequently,

$$\frac{\psi(\lambda)}{\lambda} = \int_0^\lambda \frac{d}{d\xi} \left(\frac{\psi(\xi)}{\xi} \right) d\xi.$$

Moreover, condition (2.6) can be used to provide an upper bound for $\frac{d}{d\xi}\left(\frac{\psi(\xi)}{\xi}\right)$ since there holds

$$\frac{d}{d\xi}\left(\frac{\psi(\xi)}{\xi}\right) = \frac{1}{\xi^2}\left(\xi\psi'(\xi) - \psi(\xi)\right) = \frac{1}{\xi^2}\frac{\mathrm{Ent}\left[e^{\xi X}\right]}{\mathbb{E}\left[e^{\xi X}\right]} \leq \frac{1}{\xi^2}\frac{\xi^2\nu^2}{2} = \frac{\nu^2}{2} \quad \text{for all} \quad \xi \geq 0,$$

where the second equality follows from Lemma 2.7 and the inequality follows from (2.6). Inserting this bound into the integral yields that

$$\frac{\psi(\lambda)}{\lambda} \leq \frac{\lambda\nu^2}{2} \Rightarrow \psi(\lambda) \leq \frac{\lambda^2\nu^2}{2},$$

as claimed. ■

Remark 2.10 *The fact* $\mathrm{Ent}\left[e^{\lambda(X+c)}\right] = e^{\lambda c} \cdot \mathrm{Ent}\left[e^{\lambda X}\right]$ *implies that if* X *satisfies* (2.6)*, so does* $X+c$*. That is why we do not need to center the random variable in* (2.6)*, but are still able to obtain a result for a centered random variable in* (2.7)*. Indeed, the random variables we are interested are in the form of* $f(X_1, \cdots, X_n)$ *which are generally not mean zero.*

The following proposition follows immediately from Theorem 2.9, showing the sub-Gaussian property can be defined based on the relation between $\mathrm{Ent}\left[e^{\lambda X}\right]$ and $\mathbb{E}\left[e^{\lambda X}\right]$.

Proposition 2.11 (sub-Gaussian property via Entropy) *Suppose*

$$\mathrm{Ent}\left[e^{\lambda X}\right] \leq \frac{\lambda^2\nu^2}{2}\mathbb{E}\left[e^{\lambda X}\right] \quad \textit{for all} \quad \lambda \in \mathbb{R}. \tag{2.8}$$

Then, X *is sub-Gassuain with parameter* ν*.*

Exercise 2.12 *Prove Proposition 2.11. (**Hint:** apply Theorem 2.9 to* $-X$ *and* $-\lambda$ *in the case when* (2.8) *holds for* $\lambda \leq 0$*.)*

Remark 2.13 *The above proposition provides a new perspective for sub-Gaussian distribution through the comparison of entropy of exponential and MGF, which enables us to avoid bounding MGF directly. This is very useful in establishing sub-Gaussian property of nonlinear functions since entropy tensorizes well which allows the comparison of entropy of exponential and MGF in a coordinate way.*

2.1.3 Tensorization of Entropy

The entropy has a nice tensorization property for functions of independent variables which enables us to bound the entropy of random variables in the form of $g(X_1, \cdots, X_n)$ in a coordinate-wise manner. To present this property, let us first introduce a new notation. Let $X_1, \cdots, X_n$ be independent random variables. Given $g : \mathcal{X}^n \to [0, \infty)$, we define $\mathrm{Ent}_k [g(X_1, \cdots, X_n)]$ as

$$\mathrm{Ent}_k [g(X_1, \cdots, X_n)] = \mathrm{Ent} [g(x_1, \cdots, x_{k-1}, X_k, x_{k+1}, \cdots, x_n)]|_{x_1 = X_1, \cdots, x_{k-1} = X_{k-1}, x_{k+1} = X_{k+1}, \cdots, x_n = X_n}.$$

In other words, $\mathrm{Ent}_k [g(X_1, \cdots, X_n)]$ is the entropy of $g(X_1, \cdots, X_n)$ with respect to the variable to X_k only, while the others keep fixed. Note that $\mathrm{Ent}_k [g(X_1, \cdots, X_n)]$ is still a *random variable*, a function of $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$.

Theorem 2.14 (Tensorization of entropy) *We have*

$$\mathrm{Ent} [g(X_1, \cdots, X_n)] \leq \mathbb{E} \left[\sum_{k=1}^{n} \mathrm{Ent}_k [g(X_1, \cdots, X_n)] \right],$$

where $X_1, \cdots, X_n$ are independent.

The tensorization property of entropy is an exact analogue of the tensorization property of the variance, see for example Chapter 2.1 of [2]. This property allows us to deduce a bound for functions of independent random variables from bounds for functions of each individual random variable, thus is very helpful for studying high dimensional problems. If we think $\mathrm{Ent}_k [\cdot]$ as the way of quantifying the random in the k-th mode (when average out the other random variables), the theorem implies that the randomness of the joint distribution is less than or equal to the sum of the randomness of all the modes. **Compared to MGF, this property implies that for any form of g, we can control the entropy of $g(X_1, \cdots, X_n)$ by considering the entropy of each coordinate, which means entropy tensorizes better than MGF.** The proof of this theorem is based on the following variational form of entropy[1].

Lemma 2.15 (Variational formula of entropy) *Let $Z \geq 0$ be a nonnegative random variable. Then,*

$$\begin{aligned} \mathrm{Ent} [Z] &= \sup \left\{ \mathbb{E} [ZX] : \ X \text{ is a random variable satisfying } \mathbb{E} \left[e^X \right] = 1 \right\} \\ &= \sup \{ \mathbb{E} [Z (\log Y - \log \mathbb{E} [Y])] : \ Y \geq 0 \}. \end{aligned}$$

Proof: Note if letting $X = \log (Z / \mathbb{E} [Z])$, then it is not hard to show that $\mathbb{E} \left[e^X \right] = 1$ and $\mathbb{E} [ZX] = \mathrm{Ent} [Z]$. Thus, it suffices to show that

$$\mathrm{Ent} [Z] - \mathbb{E} [ZX] \geq 0$$

for all X satisfying $\mathbb{E} \left[e^X \right] = 1$. Note that $\mathrm{Ent} [Z] - \mathbb{E} [ZX]$ can be expressed as

$$\mathrm{Ent} [Z] - \mathbb{E} [ZX] = \mathbb{E} \left[\left(e^{-X} Z \log \left(e^{-X} Z \right) \right) e^X \right] - \mathbb{E} \left[\left(e^{-X} Z \right) e^X \right] \log \mathbb{E} \left[\left(e^{-X} Z \right) e^X \right].$$

[1] Expressing a quantity in terms of a variational form is very widely used in maths. Typical examples include variational forms for norms, conjugate duality in optimization. We will see more of this technique later.

Since $\mathbb{E}\left[e^X\right] = 1$, if we define the new probability $d\mathbb{Q} = e^X d\mathbb{P}$ where $\mathbb{P}$ is the probability distribution defining Z and X, then $\mathrm{Ent}\left[Z\right] - \mathbb{E}\left[ZX\right]$ is indeed the entropy of $e^{-X}Z$ under the probability distribution, i.e.,

$$\mathrm{Ent}\left[Z\right] - \mathbb{E}\left[ZX\right] = \mathrm{Ent}_{\mathbb{Q}}\left[e^{-X}Z\right] \geq 0, \tag{2.9}$$

where the inequality follows from the nonnegative property of entropy.

The second equality follows simply from the fact that

$$\mathbb{E}\left[e^X\right] = 1 \Leftrightarrow \exists\ Y \geq 0 \text{ such that } X = \log Y - \log \mathbb{E}\left[Y\right].$$

The proof is complete now. ■

Exercise 2.16 *If you are not happy with the $d\mathbb{Q} = e^X d\mathbb{P}$ argument in* (2.9)*. Try to show it directly by following a similar argument for Jensen's inequality. That is, show that*

$$\mathbb{E}\left[f(Z)e^X\right] \geq f(\mathbb{E}\left[Ze^X\right])$$

hold for any convex function f and random variable X satisfying $\mathbb{E}\left[e^X\right] = 1$.

Proof: [of Theorem 2.14] Let $Z = g(X_1, \cdots, X_n)$ and define

$$U_k = \log \mathbb{E}\left[Z|X_1, \cdots, X_k\right] - \log \mathbb{E}\left[Z|X_1, \cdots, X_{k-1}\right].$$

Then we have

$$\mathrm{Ent}\left[Z\right] = \mathbb{E}\left[Z\left(\log\left(Z\right) - \log \mathbb{E}\left[Z\right]\right)\right] = \sum_{k=1}^{n} \mathbb{E}\left[ZU_k\right].$$

Thus, it suffices to show that $\mathbb{E}\left[ZU_k|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n\right] \leq \mathrm{Ent}_k\left[Z\right]$. To this end, let us fix

$$X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$$

and consider

$$ZU_k = Z\left(\log \mathbb{E}\left[Z|X_1, \cdots, X_k\right] - \log \mathbb{E}\left[Z|X_1, \cdots, X_{k-1}\right]\right)$$

as a function of X_k. Noting that $\mathbb{E}_{X_k}\left[\mathbb{E}\left[Z|X_1, \cdots, X_k\right]\right] = \mathbb{E}\left[Z|X_1, \cdots, X_{k-1}\right]$ due to the independence of all the X_k, the application of Lemma 2.15 with respect to X_k immediately that

$$\mathbb{E}\left[ZU_k|X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n\right] \leq \mathrm{Ent}_k\left[Z\right],$$

which completes the proof. ■

Exercise 2.17 *Verify that the equality in the tensorization property holds for*

$$g(X_1, \cdots, X_n) = \exp\left(\lambda \sum_{k=1}^{n} X_k\right),$$

where $X_1, \cdots, X_n$ are independent.

Example 2.18 (Bounded difference inequality revisited) *In this example, we are going to show that the bounded difference inequality can be proved in an alternative way based on the Herbst argument (Theorem 2.9) and the tensorization property (Theorem 2.14). Recall that a function f satisfies the bounded difference property if*

$$|f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n)| \leq L_k$$

with parameters $(L_1, \cdots, L_n)$ over the range of the independent random variables $X = (X_1, \cdots, X_n)$. Thus, when fixing $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$, $f(X_1, \cdots, X_n)$ can be viewed as a bounded random variable which locates in an interval of length at most L_k. Then it follows from Example 2.8 that

$$\mathrm{Ent}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right] \leq \frac{L_k^2}{8}\mathbb{E}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right],$$

where $\mathbb{E}_k[\cdot]$ means taking expectation with respect to X_k only. Furthermore, letting $g(X_1, \cdots, X_n) = e^{\lambda f(X_1, \cdots, X_n)}$, the tensorization property implies that

$$\begin{aligned}\mathrm{Ent}\left[e^{\lambda f(X_1, \cdots, X_n)}\right] &\leq \sum_{k=1}^n \frac{L_k^2}{8}\mathbb{E}\left[\mathbb{E}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right]\right] \\ &= \left(\sum_{k=1}^n \frac{L_k^2}{8}\right)\mathbb{E}\left[e^{\lambda f(X_1, \cdots, X_n)}\right].\end{aligned}$$

Thus, by the Herbst argument, we know that $f(X_1, \cdots, X_n)$ is sub-Gaussian with parameter $\nu^2 = \frac{\sum_{k=1}^n L_k^2}{4}$ and the tail bound in the bounded difference inequality follows immediately.

2.2 Modified Log-Sobolev Inequality and Entropy Method

As demonstrated in the last example, in order to apply the Herbst argument and the tensorization property to establish the sub-Gaussian tail, it remains to bound $\mathrm{Ent}_k\left[e^{\lambda f(X_1, \cdots, X_n)}\right]$. In a more general setting, this can be achieved by the modified log-Sobolev inequality[2] (MLS), which is the last piece of the entropy method. In a nutshell, MLS controls the entropy of $e^{\lambda f(X)}$ by the fluctuation/gradient of the function f.

Lemma 2.19 (MLS) *Let $f : \mathcal{X} \to \mathbb{R}$ be a single variable function. Define*

$$D^- f(x) = f(x) - \inf_{z \in \mathcal{X}} f(z).$$

Then for any $\lambda \geq 0$ we have

$$\begin{aligned}\mathrm{Ent}\left[e^{\lambda f(X)}\right] &\leq \mathbb{E}\left[\phi\left(\lambda D^- f(X)\right)e^{\lambda f(X)}\right] \\ &\leq \frac{1}{2}\mathbb{E}\left[|\lambda D^- f(X)|^2 e^{\lambda f(X)}\right].\end{aligned}$$

where $\phi(x) = e^{-x} + x - 1$.

[2]Overall, Sobolev inequalities are a family of inequalities which control the energy of functions by that of their derivatives.

Proof: By (2.5), we have

$$\mathrm{Ent}\left[Z\right] = \inf_{t>0} \mathbb{E}\left[Z\log Z - Z\log t - Z + t\right].$$

Thus, letting $Z = e^{\lambda f(X)}$ yields that

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda f(X)}\right] &= \inf_{t>0} \mathbb{E}\left[\lambda f(X)e^{\lambda f(X)} - e^{\lambda f(X)}\log t - e^{\lambda f(X)} + t\right] \\
&\le \mathbb{E}\left[\lambda f(X)e^{\lambda f(X)} - e^{\lambda f(X)}\log\left(e^{\lambda \inf_z f(z)}\right) - e^{\lambda f(X)} + e^{\lambda \inf_z f(z)}\right] \\
&= \mathbb{E}\left[\left\{\lambda f(X) - \lambda \inf_z f(z) - 1 + e^{-\lambda f(X) + \lambda \inf_z f(z)}\right\} e^{\lambda f(X)}\right] \\
&= \mathbb{E}\left[\phi\left(\lambda D^- f(X)\right) e^{\lambda f(X)}\right].
\end{aligned}$$

The second inequality in the lemma simply follows from the fact $\phi(x) \le \frac{1}{2}x^2$ for $x \ge 0$. ∎

When $f : \mathcal{X}^n \to \mathbb{R}$ is a multivariable function, applying the MLS conditionally to each $\mathrm{Ent}_k\left[e^{\lambda f(X_1,\cdots,X_n)}\right]$ leads to the following theorem which can be viewed as an generalization of the bounded difference inequality.

Theorem 2.20 (General bounded difference inequality) *Let $x = (x_1, \cdots, x_n)$ and define*

$$\begin{aligned}
D_k^- f(x) &= f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - \inf_{z\in\mathcal{X}} f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n), \\
D_k^+(x) &= \sup_{z\in\mathcal{X}} f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n).
\end{aligned}$$

Let $X_1, \cdots, X_n$ be i.i.d random variables. If $\sum_{k=1}^n |D_k^- f(x)|^2 \le \nu_1^2$, then we have

$$\mathbb{P}\left[f(X_1, \cdots, X_n) \ge \mathbb{E}\left[f(X_1, \cdots, X_n)\right] + t\right] \le \exp\left(-\frac{t^2}{2\nu_1^2}\right)$$

Similarly, if $\sum_{k=1}^n |D_k^+ f(x)|^2 \le \nu_2^2$, we have

$$\mathbb{P}\left[f(X_1, \cdots, X_n) \le \mathbb{E}\left[f(X_1, \cdots, X_n)\right] - t\right] \le \exp\left(-\frac{t^2}{2\nu_2^2}\right).$$

Proof: If we fix $X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n$ and consider $f(X_1, \cdots, X_n)$ as a function of X_k, the application of the MLS yields that

$$\mathrm{Ent}_k\left[e^{\lambda f(X_1,\cdots,X_n)}\right] \le \frac{1}{2}\mathbb{E}_k\left[|\lambda D_k^- f(X_1, \cdots, X_n)|^2 e^{\lambda f(X_1,\cdots,X_n)}\right], \quad \lambda \ge 0.$$

By the tensorization property of entropy, we have

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda f(X_1,\cdots,X_n)}\right] &\le \sum_{k=1}^n \mathbb{E}\left[\mathrm{Ent}_k\left[e^{\lambda f(X_1,\cdots,X_n)}\right]\right] \\
&\le \frac{1}{2}\mathbb{E}\left[\lambda^2\left(\sum_{k=1}^n |D_k^- f(X_1, \cdots, X_n)|^2\right) e^{\lambda f(X_1,\cdots,X_n)}\right] \\
&\le \frac{\lambda^2\nu_1^2}{2}\mathbb{E}\left[e^{\lambda f(X_1,\cdots,X_n)}\right].
\end{aligned}$$

The upper tail bound follows immediately from the Herbst argument and the Chernoff method.

The lower tail bound can be established by considering $-f$. ∎

Remark 2.21 *Let*

$$D_k f(x) = \sup_z f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n) - \inf_z f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n).$$

Note the tail bound obtained by the bounded difference inequality is of the order $\exp\left(-t^2/\sum_{k=1}^n \|D_k f(x)\|_\infty^2\right)$, *while the tail bound established here is of the order* $\exp\left(-t^2/\|\sum_{k=1}^n |D_k f(x)|^2\|_\infty\right)$. *It is trivial that* $\left\|\sum_{k=1}^n |D_k f(x)|^2\right\|_\infty \le \sum_{k=1}^n \|D_k f(x)\|_\infty^2$. *Moreover, there are cases* $\left\|\sum_{k=1}^n |D_k f(x)|^2\right\|_\infty$ *can be sufficiently smaller than* $\sum_{k=1}^n \|D_k f(x)\|_\infty^2$. *Thus, the bounds of Theorem 2.20 are an improvement over that in the bounded difference inequality. This is due to the fact entropy function tensorizes better than the moment generating function.*

Note that the upper and lower tail bounds here are essentially asymmetric: the upper bound is controlled by $\sum_{k=1}^n |D_k^- f(X)|^2$ *while the lower bound is controlled by* $\sum_{k=1}^n |D_k^+ f(X)|^2$. *There are problems where it is may be not clear how to bound one of them. However, if the function* f *satisfies a stronger condition, it is still possible to obtain a two-sided tail bound from the single bound of* $\sum_{k=1}^n |D_k^- f(X)|^2$. *The machinery needed to prove such bounds are discussed in the next lecture.*

Example 2.22 *Let* $\mathcal{A} \subset \mathbb{R}^{n\times n}$ *be a set of symmetric matrices whose entries are* $\{1, -1\}$. *For any* $A \in \mathcal{A}$, *we let* $\lambda_{\max}(A)$ *denote the largest eigenvalue of* A. *By variational formula of the largest eigenvalue of symmetric matrix, we know that*

$$\lambda_{\max}(A) = \sup_{\|v\|_2=1} \langle v, Av\rangle . \tag{2.10}$$

Let $\mathcal{A}' \subset \mathcal{A}$ *be the subset of matrices which can only differ from* A *in the* (i,j)*-th and* (j,i)*-th entries and let* A^- *be a matrix such that*

$$\lambda_{\max}(A^-) = \inf_{A'\in\mathcal{A}',\ A'_{ij}=A'_{ji}\in\{1,-1\}} \lambda_{\max}(A').$$

Letting v_A *be the unit vector where the equality in* (2.10) *is achieved, then we have*

$$\begin{aligned} D_{ij}^- \lambda_{\max}(A) &= \langle v_A, Av_A\rangle - \max_{\|v\|_2} \left\langle v, A^- v\right\rangle \\ &\le \langle v_A, Av_A\rangle - \left\langle v_A, A^- v_A\right\rangle \\ &= \left\langle v_A, (A - A^-)v_A\right\rangle \\ &\le 4|v_A(i)||v_A(j)|. \end{aligned} \tag{2.11}$$

Note the above bound only relies on A. *Thus, it follows that*

$$\begin{aligned} \sum_{1\le i\le j\le n} |D_{ij}^- \lambda_{\max}(A)|^2 &\le 16 \sum_{i,j=1}^n |v_A(i)|^2 |v_A(j)|^2 \\ &= 16 \left(\sum_{i=1}^n |v_A(i)|^2|\right) \left(\sum_{i=1}^n |v_A(j)|^2|\right) \\ &= 16. \end{aligned} \tag{2.12}$$

Thus, if A *is a random symmetric matrix with independent entries (upper triangular part) taking the value from* $\{1, -1\}$ *randomly, then by the general bounded difference inequality we can establish the following upper tail*

$$\mathbb{P}\left[\lambda_{\max}(A) \ge \mathbb{E}\left[\lambda_{\max}(A)\right] + t\right] \le e^{-t^2/32}.$$

Of course we still need to estimate the mean of $\lambda_{\max}(A)$. This will be discussed in the future by studying the mean of the supremum of an empirical process.

It is worth noting that it seems we cannot use the general bounded difference inequality to establish a dimension free lower bound. To see this, first note that we can establish a similar bound to (2.11) *for $D_{ij}^{+}\lambda_{\max}(A)$, but with the principal eigenvector of A being replaced by that of principal eigenvector of a A^{+} which only differs from A in the the (i,j)-th and (j,i)-th entries and in the meantime achieves the maximum principal eigenvalue, i.e.,*

$$D_{ij}^{+}\lambda_{\max}(A) \leq 4|v_{A^{+}}(i)||v_{A^{+}}(j)|.$$

Noting that $v_{A^{+}}$ varies for different (i,j), we cannot proceed in the similar fashion as in (2.12). *As can seen from the next lecture, we can establish the lower tail by a different technique.*

From the general bounded difference inequality we can obtain the following proposition which is relatively easier to manage. Recall that a function $f : \mathcal{X}^n \to \mathbb{R}$ is *separately convex* if for each $i = 1, \cdots, n$, it is a convex function of its i-th coordinate while the rest of the coordinates are fixed.

Proposition 2.23 *Let $X_k, k = 1, \cdots, n$ be independent random variables taking values in an interval $[a,b]$ and let $f : [a,b]^n \to \mathbb{R}$ be a separately convex function which also satisfies the Lipschitz condition $|f(x) - f(y)| \leq L\|x - y\|_2$ for all $x, y \in [a,b]^n$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left[f(X_1, \cdots, X_n) \geq \mathbb{E}\left[f(X_1, \cdots, X_n)\right] + t\right] \leq \exp\left(-\frac{t^2}{2L^2(b-a)^2}\right)$$

Proof: For ease of presentation, we assume the partial derivatives of f exist (Otherwise we can adopt a standard approximation argument). Letting x_k' be the random variable at which

$$\inf_z f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n)$$

is achieved. Then it follows from the separately convex property of f that

$$\begin{aligned} D_k^{-} f(x) &= f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - \inf_z f(x_1, \cdots, x_{k-1}, z, x_{k+1}, \cdots, x_n) \\ &= f(x_1, \cdots, x_{k-1}, x_k, x_{k+1}, \cdots, x_n) - f(x_1, \cdots, x_{k-1}, x_k', x_{k+1}, \cdots, x_n) \\ &\leq \partial_k f(x)(x_k - x_k'). \qquad \text{(use the separately convexity here)} \end{aligned}$$

It follows that

$$\begin{aligned} \sum_{k=1}^{n} |D_k^{-} f(x)|^2 &\leq \sum_{k=1}^{n} \partial_k^2 f(x)(x_k - x_k')^2 \\ &\leq \sum_{k=1}^{n} \partial_k^2 f(x)(b-a)^2 \\ &= \|\nabla f(x)\|_2^2 (b-a)^2 \leq L^2(b-a)^2, \end{aligned}$$

where $\|\nabla f(x)\|_2 \leq L$ follows from the Lipschitz condition of f (**check this!**). The upper tail bound then follows immediately from Theorem 2.20. ∎

Remark 2.24 *The lower tail cannot be established by considering $-f$ since it is a concave function.*

Example 2.25 (Sharper upper bounds on Rademacher complexity) *Let us revisit Example 1.43 of Lecture 1, which is about establishing an upper tail bound for*

$$Z = \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^{n} a_k \varepsilon_k \right],$$

where $\varepsilon_k, k = 1, \cdots, n$ are i.i.d Rademacher variables. Let

$$f(x_1, \cdots, x_n) = \sup_{a \in \mathcal{A}} \left[\sum_{k=1}^{n} a_k x_k \right], \quad x_k \in \{1, -1\}.$$

*Since f is a supremum of a collection of linear function, it is a convex function and hence separately convex. Moreover, it is not hard to show that (**check this!**)*

$$|f(x_1, \cdots, x_n) - f(x_1', \cdots, x_n')| \leq \sup_{a \in \mathcal{A}} \|a\|_2 \|x - x'\|_2,$$

where $x = (x_1, \cdots, x_n)$ and $x' = (x_1', \cdots, x_n')$ That is, f is Lipschitz with parameter $\sup_{a \in \mathcal{A}} \|a\|_2$. Thus, it follows from Proposition 2.23 that

$$\mathbb{P}\left[f(\varepsilon_1, \cdots, \varepsilon_n) \geq \mathbb{E}\left[f(\varepsilon_1, \cdots, \varepsilon_n) \right] + t \right] \leq \exp\left(-\frac{t^2}{8 \sup_{a \in \mathcal{A}} \|a\|_2^2} \right).$$

Note that the quantity $\sup_{a \in \mathcal{A}} \|a\|_2^2$ (the squared Euclidean width of the set) used in the upper bound here may be substantially than $\sum_{k=1}^{n} \sup_{a \in \mathcal{A}} a_k^2$ established in Lecture 1.

2.3 Gaussian concentration

Lastly, we present a classical concentration inequality of standard Gaussian random variables. The proof of the inequality requires a type of Gaussian log-Sobolev inequality which is listed below without proof. Interested readers are referred to Chapter 3.4 of [2] or Chapter 5.3 of [3] for a proof.

Lemma 2.26 (Gaussian log-Sobolev inequality) *Let $X_1, \cdots, X_n$ be a collection of n independent standard Gaussian random variables, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Then*

$$\text{Ent}\left[f^2(X_1, \cdots, X_n) \right] \leq 2\mathbb{E}\left[\|\nabla f(X_1, \cdots, X_n)\|_2^2 \right].$$

To see why the above inequality is referred to as a type of log-Sobolev inequality, assume for simplicity f is single variable function. Then by the chain rule we have

$$\text{Ent}\left[e^{\lambda f(X)} \right] \leq \frac{1}{2}\mathbb{E}\left[|\lambda f'(X)|^2 e^{\lambda f(X)} \right], \quad \text{for all } \lambda \in \mathbb{R}$$

which is analogous to the one give in Lemma 2.19, but with the discrete gradient replaced by the calculus gradient. Similarly, when f is a multivariable function, we have (**check this!**)

$$\text{Ent}\left[e^{\lambda f(X_1, \cdots, X_n)} \right] \leq \frac{1}{2}\mathbb{E}\left[\|\lambda \nabla f(X_1, \cdots, X_n)\|_2^2 \, e^{\lambda f(X_1, \cdots, X_n)} \right], \quad \text{for all } \lambda \in \mathbb{R} \tag{2.13}$$

Theorem 2.27 *Let $X_1, \cdots, X_n$ be a collection of n independent standard Gaussian random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function with parameter $L > 0$. That is, for any $x, y \in \mathbb{R}^n$,*

$$|f(x) - f(y)| \leq L\|x - y\|_2.$$

Then $f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]$ is sub-Gaussian with parameter $\nu^2 = L^2$, and hence

$$\mathbb{P}[|f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)]| \geq t] \leq 2\exp\left(-\frac{t^2}{2L^2}\right).$$

Proof: We may assume f is differentiable (otherwise we can use an approximation argument). Then $\|\nabla f(X_1, \cdots, X_n)\|_2$ is bounded by L (**check this!**). It follows from (2.13) that

$$\text{Ent}\left[e^{\lambda f(X_1, \cdots, X_n)}\right] \leq \frac{\lambda^2 L^2}{2}\mathbb{E}\left[e^{\lambda f(X_1, \cdots, X_n)}\right].$$

Then claim follows immediately by the Herbst argument. ■

Example 2.28 (Gaussian complexity) *Let $X_1, \cdots, X_n$ be an i.i.d sequence of $\mathcal{N}(0, 1)$ variables. Given a set $\mathcal{A} \in \mathbb{R}^n$, define the random variable*

$$Z = \sup_{a \in \mathcal{A}}\left[\sum_{k=1}^n a_k X_k\right] = \sup_{a \in \mathcal{A}}\langle a, X\rangle,$$

where $X = (X_1, \cdots, X_n)$. The Gaussian complexity, denoted $\mathcal{G}_n(\mathcal{A})$, is defined as the expectation of Z,

$$\mathcal{G}_n(\mathcal{A}) = \mathbb{E}[Z],$$

which is another way to measure the complexity of a set (cf. the Rademacher complexity).

*Define $f(x_1, \cdots, x_n) = \sup_{a \in \mathcal{A}}[\sum_{k=1}^n a_k x_k]$. Since f is a Lipschitz function with parameter $\sup_{a \in \mathcal{A}}\|a\|_2$ (**check this!**), by the Gaussian concentration inequality we know that $Z = \sup_{a \in \mathcal{A}}\langle a, X\rangle$ is sub-Gaussian with parameter $\nu^2 = \sup_{a \in \mathcal{A}}\|a\|_2^2$.*

Example 2.29 (Singular values of Gaussian random matrices) *Let $A \in \mathbb{R}^{n \times n}$ be a random Gaussian matrix whose entries obey the i.i.d standard Gaussian distribution. Let $\sigma_k(A)$ be the k-th largest singular value of A. By Weyl's theorem (this can be found in any standard linear algebra textbook), we have*

$$|\sigma_k(A) - \sigma_k(A')| \leq \|A - A'\|_F.$$

That is, $\sigma_k(A)$ is Lipschitz with parameter 1. Therefore, we can conclude that $\sigma_k(A)$ is sub-Gaussian with parameter $\nu^2 = 1$.

Reading Materials

[1] Martin Wainwright, *High-dimensional statistics – A non-asymptotic viewpoint*, Chapter 3.1.

[2] Ramon van Handel, *Probability in High Dimension*, Chapters 3.3, 3.4.

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Chapters 5.3, 5.4, 6.1, 6.3, 6.4, 6.6.