CS 839 Probability and Learning in High Dimension Lecture 12 - 03/07/2022

Lecture 12: Lipschitz Concentration and Gaussian Comparison

Lecturer: Yudong Chen *Scribe: Xufeng Cai*

In this lecture,[1] we will briefly discuss the general setup of statistical learning, where one often needs to bound the supremum of random processes. Motivated by this problem, we will present the Lipschitz concentration inequalities and the Gaussian comparison inequality as useful tools for controlling random processes.

1 Statistical Learning

Setup. Observe $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times [0,1]$, where $x_i \overset{i.i.d.}{\sim} \mu$ and $y_i = f^*(x_i)$. Here μ is some unknown distribution, and $f^* : \mathcal{X} \to [0,1]$ is also some unknown regression function. For each function $f \in \mathcal{F}$, we define

1. **empirical risk:** $L_n(f) = \frac{1}{n} \sum_{i=1}^n \left(f\left(x_i\right) - f^*\left(x_i\right)\right)^2$, which can be understood as the training error;
2. **population risk:** $L(f) = \mathbb{E}_{X \sim \mu} \left(f(X) - f^*(X)\right)^2$, which can be understood as the test error.

Ideally, we want to find the minimizer of the population risk, namely

$$f_0 \triangleq \arg\min_{f \in \mathcal{F}} L(f),$$

but $L(f)$ is usually not computable without knowing μ and f^*. Instead, we consider the empirical risk minimization (ERM) approach:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} L_n(f).$$

We use $\hat{f}$ as our estimator for f^*.

Risk decomposition. To bound the population risk of the empirical risk minimizer $\hat{f}$, we may decompose the risk as follows:

$$\begin{aligned}
\underbrace{L(\hat{f})}_{\text{test error}} &= \underbrace{\left[L(\hat{f}) - L_n(\hat{f})\right]}_{\text{generalization gap}} + \underbrace{L_n(\hat{f})}_{\text{training error}} \\
&\le \left[L\left(f^*\right) - L_n(\hat{f})\right] + L_n\left(f_0\right) \\
&= \underbrace{\left[L(\hat{f}) - L_n(\hat{f})\right] + \left[L_n\left(f_0\right) - L\left(f_0\right)\right]}_{\text{estimation error}} + \underbrace{L\left(f_0\right)}_{\text{approximation error}} \\
&\overset{(i)}{\le} 2 \sup_{f \in \mathcal{F}} \left|L_n(f) - L(f)\right| + L\left(f_0\right).
\end{aligned}$$

[1] References:

1. Section 3.1, 5.4, 6.2 in "High-Dimensional Statistics: A Non-Asymptotic Viewpoint", Martin J. Wainwright, Cambridge University Press, 2019,
2. (Additional reading) Sec 3.3 in "Lecture Notes for Statistics 311/Electrical Engineering 377: Information Theory and Statisti cs", John Duchi,
3. (Additional reading) Section 5, 7.2, 7.3 in "High -Dimensional Probability: An Introduction with Applications in Data Science", Roman Vershynin, Cambridge University Press, 2018.

(Note that step (i) may not be tight, especially when $\mathcal{F}$ is be very large.) Equivalently, by rearranging terms, we get the bound on the so-called *excess risk*:

$$
\begin{aligned}
\underbrace{L(\hat{f}) - L\left(f_0\right)}_{\text{excess risk}} &\leq 2 \sup_{f \in \mathcal{F}} \left|L_n(f) - L(f)\right| \\
&= 2 \sup_{f \in \mathcal{F}} \underbrace{\left| \frac{1}{n} \sum_{i=1}^{n} \left(f\left(x_i\right) - f^*\left(x_i\right)\right)^2 - \mathbb{E}_{X \sim \mu} \left(f(X) - f^*(X)\right)^2 \right|}_{=: Z_f}.
\end{aligned}
$$

In the above argument, in order to bound the test error and excess risk, we need to control a quantity of the form $\sup_f Z_f$. It can be done in two steps.

1. Bound the deviation $\sup_f Z_f - \mathbb{E} \sup_f Z_f$ using concentration inequalities.
2. Bound the expectation $\mathbb{E} \sup_f Z_f$, which is the supremum of empirical process.

In this lecture, we will introduce some tools for the two tasks above.

2 Lipschitz Concentration

In this section, we introduce several concentration inequalities for Lipschitz functions. We first summarize some terminology for a function $f : \mathbb{R}^n \to \mathbb{R}$. We say f is

1. **separately convex** if the function $x_k \mapsto f\left(x_1, \ldots, x_k, \ldots, x_n\right)$ is convex for each $1 \leq k \leq n$ and each fixed $\{x_j : j \neq k\}$.
2. **Lipschitz** (w.r.t. ℓ_2 norm) if $|f(x) - f(y)| \leq L\|x - y\|_2$ for any $x, y \in \mathbb{R}^n$.

Then we present our main theorem for this section.[2]

Theorem 1. *Let $X_1, \ldots, X_n$ be independent random variables supported on $[a, b]$. Further let $f : \mathbb{R}^n \to \mathbb{R}$ be separately convex and L-Lipschitz. Then for all $t \geq 0$*

$$
\mathbb{P}\left[f\left(X_1, \ldots, X_n\right) \geq \mathbb{E}\left[f\left(X_1, \ldots, X_n\right)\right] + t\right] \leq \exp\left(-\frac{t^2}{4L^2(b-a)^2}\right).
$$

Remark Let $X = (X_1, \ldots, X_n)$. Theorem 1 shows that the upper tail of $f(X)$ behaves like a sub-Gaussian random variable with parameter $\mathcal{O}(L^2(b-a)^2)$.

2.1 Variants and extensions

In this section, we present some variants and extensions of Theorem 1.

The next theorem provides a two-sided bound on $f(X)$ under the stronger *joint convexity* assumption.

Theorem 2 (Two-sided bound)**.** *If $X_1, \ldots, X_n$ are independent random variables, each bounded on $[a, b]$, and $f : \mathbb{R}^n \to \mathbb{R}$ is convex and L-Lipschitz. Then for all $t \geq 0$,*

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

[2] To prove the theorem, we can use the Entropy Method. We first prove that it holds for $n = 1$ dimension, and then generalize it to the n-dimensional case (tensorization step).

Remark Theorem 2 shows that $f(X)$ is sub-Gaussian with parameter $\mathcal{O}(L^2(b-a)^2)$. Note that the joint convexity assumption is necessary for getting a two sided bound.

For Gaussian random variables, we can get rid of the assumption of convexity while still getting a two-sided bound.

Theorem 3 (Gaussian Case). *If $X_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ for $1 \le i \le n$, and $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz. Then for all $t \ge 0$*

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

The above inequalities can be compared with the Bounded Difference Inequality, which can also be used to establish concentration of functions of many independent variables. Below we use the notation $x_{-k} := (x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_n)$.

Theorem 4 (Bounded Difference Inequality). *Suppose $X_1,\ldots,X_n$ are independent random variables, and $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference property*

$$\left|f\left(x_k,x_{-k}\right) - f\left(x_k',x_{-k}\right)\right| \le L_k, \quad \forall k, x_k, x_k', x_{-k}.$$

Then for all $t \ge 0$

$$\mathbb{P}\left[\left|f\left(X_1,\ldots,X_n\right) - \mathbb{E}\left[f\left(X_1,\ldots,X_n\right)\right]\right| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right).$$

Remark Theorem 4 says that if f has bounded difference, then $f(X)$ is a sub-Gaussian random variable with parameter $\mathcal{O}(\sum_{k=1}^n L_k^2)$. In many problems, we have $\sum_{k=1}^n L_k^2 \gg L^2$, where L is the Lipschitz constant of f, in which case Theorem 4 gives weaker results than Theorems 1-3.

2.2 Applications

In this part, we discuss some applications where we can use Lipschitz concentration inequalities.

2.2.1 Concentration of norm

Note that the function $x \mapsto \|x\|_2$ is convex, and 1-Lipschitz by the triangular inequality

$$\left|\|X\|_2 - \|Y\|_2\right| \le \|X - Y\|_2.$$

Then for $X = (X_1,\ldots,X_n)$ with independent bounded or Gaussian X_i's, Theorem 2 and 3 give

$$\|X\|_2 - \mathbb{E}\left[\|X\|_2\right] \text{ is } \mathcal{O}(1)\text{-sub-Gaussian.}$$

If X_i's are bounded, then the function $x \mapsto \|x\|_2$ also satisfies the bounded difference property

$$\left|\|x_1,\ldots,x_k,\ldots,x_n\|_2 - \|x_1,\ldots,x_k',\ldots,x_n\|_2\right| \le |x_k - x_k'| \le L_k = \mathcal{O}(1)$$

Theorem 4 gives

$$\|X\|_2 - \mathbb{E}\left[\|X\|_2\right] \text{ is } \mathcal{O}(n)\text{-sub-Gaussian,}$$

which is a much weaker result.

2.2.2 Maximum singular value of random matrices

Consider a random matrix $X \in \mathbb{R}^{n \times m}$ with X_{ij} independently and identically distributed as either Gaussian or bounded. We consider the operator norm (the largest singular value) of X, which can be written as a supremum as follows:

$$\|X\|_{\text{op}} = \text{maximum singular value of } X = \sup_{\|u\|_2 \leq 1, \|v\|_2 \leq 1} u^T X v.$$

Note that the operator norm $\|\cdot\|_{\text{op}}$ is

1. convex, since it is the maximum of affine functions, and
2. 1-Lipschitz with respect to $\|\cdot\|_F$, since $|\|X\|_{\text{op}} - \|Y\|_{\text{op}}| \leq \|X - Y\|_{\text{op}} \leq \|X - Y\|_{\text{F}}$.

By Theorem 2 and 3, $\|X\|_{\text{op}} - \mathbb{E}\left[\|X\|_{\text{op}}\right]$ is $\mathcal{O}(1)$-sub-Gaussian. Note that the sub-Gaussian parameter is independent of the dimension n, m.

2.2.3 Any singular values of a Gaussian matrix

Consider a random matrix $X \in \mathbb{R}^{n \times m}$ with X_{ij} independently and identically distributed as Gaussian $N(0, 1)$, and let $\sigma_k(X) = k$-th largest singular value of X. (Note that $\sigma_1(X) = \|X\|_{\text{op}}$). Observe that $\sigma_k(\cdot)$ is

1. NOT convex for $k \geq 2$, but
2. 1-Lipschitz, since $|\sigma_k(X) - \sigma_k(Y)| \leq \|X - Y\|_{\text{op}} \leq \|X - Y\|_{\text{F}}$ (Weyl's Inequality).

So we restrict our analysis to the Gaussian case, as it does not require convexity. By Theorem 3, we have

$$\sigma_k(X) - \mathbb{E}\left[\sigma_k(X)\right] \text{ is } \mathcal{O}(1)\text{-sub-Gaussian.}$$

2.2.4 Rademacher complexity

Let $A \subset \mathbb{R}^n$. The **Rademacher complexity** of A is

$$R_n(A) \triangleq \mathbb{E}\left[\sup_{a \in A} \sum_{i=1}^{n} a_i \epsilon_i\right],$$

where $\epsilon_i \in \{-1, +1\}$ are i.i.d. Rademacher random variables. Let $\hat{R}_n(A) \triangleq \sup_{a \in A} \sum_{i=1}^{n} a_i \epsilon_i$. Note that $\hat{R}_n(A)$ is

1. a convex function of $\epsilon := (\epsilon_1, \ldots, \epsilon_n)$, and
2. $W(A)$-Lipschitz: $|\sup_{a \in A} \langle a, \epsilon \rangle - \sup_{a \in A} \langle a, \epsilon' \rangle| \leq |\sup_{a \in A} \langle a, \epsilon - \epsilon' \rangle| \leq \underbrace{\sup_{a \in A} \|a\|_2}_{W(A)} \|\epsilon - \epsilon'\|_2$.

Here $W(A) := \sup_{a \in A} \|a\|_2$ is the width/radius of the set A. By Theorem 2, we have

$$\mathbb{P}\left(\left|\hat{R}_n(A) - R_n(A)\right| \geq t\right) \leq 2 \exp\left(\frac{-t^2}{8W(A)^2}\right).$$

2.3 Other Remarks

1. Theorem 2 and 3 imply the usual Hoeffding's inequality, as the function $f(X) = \sum_{i=1}^{n} X_i$ is convex and $\sqrt{n}$-Lipschitz.

2. There are Bernstein versions of these inequalities that account for the variance.

3. This type of inequalities are often used to bound $f(x) = \sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} g(x_i)$ (supremum of empirical process). In particular, we have the "Functional Hoeffding Inequality":

   Theorem 5 (Functional Hoeffding Inequality). *If $X_i \in \mathcal{X}_i, i = 1, \ldots, n$ are independent RV's, and for all $g \in G$,*

   $$g(x_i) \in [a_{i,g}, b_{i,g}], \quad \forall x_i \in \mathcal{X}_i,$$

   then

   $$\mathbb{P}(f(X_1, \ldots, X_n) - \mathbb{E}[f(X_1, \ldots, X_n)] \geq t) \leq \exp\left(-\frac{nt^2}{4L^2}\right),$$

   where $L^2 \triangleq \sup_{g \in G} \left\{ \frac{1}{n} \sum_{i=1}^{n} (b_{i,g} - a_{i,g})^2 \right\}$.

   Note that if we instead apply Theorem 4 (bounded difference inequality) to the above setting, it would involve a quantity of the form $L^2 = \frac{1}{n} \sum_{i=1}^{n} \sup_{g \in G} (b_{i,g} - a_{i,g})^2$, which usually leads to weaker result than the functional Hoeffding bound.

3 Gaussian Comparison Inequality

In this section, we present Gaussian Comparison Inequalities, which can be used to bound the expectation $\mathbb{E} \sup_f Z_f$.

Theorem 6 (Slepian's Inequality). *Let $Z, Y \in \mathbb{R}^N$ be zero-mean Gaussian vectors such that*

$$\mathbb{E}\left[Z_i^2\right] = \mathbb{E}\left[Y_i^2\right], \forall i$$
$$\mathbb{E}\left[Z_i Z_j\right] \geq \mathbb{E}\left[Y_i Y_j\right], \forall i, j.$$

Then

$$\mathbb{E}\left[\max_i Z_i\right] \leq \mathbb{E}\left[\max_i Y_i\right].$$

Remark

1. Note that the condition $\mathbb{E}\left[Z_i^2\right] = \mathbb{E}\left[Y_i^2\right]$ shows the variances of Z_i and Y_i are equal for all i, as they are zero-mean. Also, the condition $\mathbb{E}\left[Z_i Z_j\right] \geq \mathbb{E}\left[Y_i Y_j\right], \forall i, j$ means that Z_i's are more correlated than Y_i's. So Theorem 6 basically tells us that for zero-mean Gaussian vectors under the condition that variances are equal, high correlations reduce the expectation of maximum. We can think of the extreme case in which Z_i's are all identical to each other.

2. Slepian's Inequality holds for any N, so it can be used to compare the expectation of the supremum over infinite sets.

There is a more general Gaussian comparison inequality that does not require equal variance.

Theorem 7 (Sudakov-Fernique's Inequality). *Let $Z, Y \in \mathbb{R}^N$ be zero-mean Gaussian vectors such that*

$$\mathbb{E}(Z_i - Z_j)^2 \leq \mathbb{E}(Y_i - Y_j)^2, \forall i, j.$$

Then

$$\mathbb{E}\left[\max_i Z_i\right] \leq \mathbb{E}\left[\max_i Y_i\right].$$

It is clear that Theorem 7 strictly generalizes Theorem 6.

3.1 Application to Gaussian matrix

Suppose $X \in \mathbb{R}^{n\times n}$, where $X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. We want to bound $\mathbb{E}\left[\|X\|_{\text{op}}\right] = \mathbb{E}\left[\sup_{u,v\in\mathbb{S}^{n-1}} u^T X v\right]$, where $\mathbb{S}^{n-1}$ denotes the unit sphere in $\mathbb{R}^n$. To do so, we compare two processes

$$\begin{aligned}
&Z_{uv} := u^T X v = \left\langle X, uv^T \right\rangle, \\
&Y_{uv} := g^T u + h^T v, \qquad \text{where } g, h \sim N\left(0, I_n\right) \text{ and } g \perp h.
\end{aligned}$$

For all $u, v, \tilde{u}, \tilde{v} \in \mathbb{S}^{n-1}$, we have

$$\begin{aligned}
\mathbb{E}\left[\left(Z_{uv} - Z_{\tilde{u}\tilde{v}}\right)^2\right] &= \mathbb{E}\left[\left\langle X, uv^T - \tilde{u}\tilde{v}^T \right\rangle^2\right] \\
&= \left\|uv^T - \tilde{u}\tilde{v}^T\right\|_{\text{F}}^2 \\
&= \|\tilde{v}\|_2^2\|u-\tilde{u}\|_2^2 + \|u\|_2^2\|v-\tilde{v}\|_2^2 + 2\underbrace{\left(\|u\|_2^2 - \langle u,\tilde{u}\rangle\right)}_{\geq 0}\underbrace{\left(\langle v,\tilde{v}\rangle - \|\tilde{v}\|_2^2\right)}_{\leq 0} \\
&\leq \|u-\tilde{u}\|_2^2 + \|v-\tilde{v}\|_2^2.
\end{aligned}$$

On the other hand, we get

$$\begin{aligned}
\mathbb{E}\left[\left(Y_{uv} - Y_{\tilde{u}\tilde{v}}\right)^2\right] &= \mathbb{E}\left[\left(g^T(u-\tilde{u}) + h^T(v-\tilde{v})\right)^2\right] \\
&= \|u-\tilde{u}\|_2^2 + \|v-\tilde{v}\|_2^2.
\end{aligned}$$

So $\mathbb{E}\left[\left(Z_{uv} - Z_{\tilde{u}\tilde{v}}\right)^2\right] \leq \mathbb{E}\left[\left(Y_{uv} - Y_{\tilde{u}\tilde{v}}\right)^2\right]$. By Sudakov-Fernique (Theorem 7), we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{u,v\in\mathbb{S}^{n-1}} u^T X v\right] &\leq \mathbb{E}\left[\sup_{u,v\in\mathbb{S}^{n-1}} g^T u + h^T v\right] \\
&= \mathbb{E}\left[\|g\|_2 + \|h\|_2\right] \\
&\overset{(i)}{\leq} \sqrt{\mathbb{E}\left[\|g\|_2^2\right]} + \sqrt{\mathbb{E}\left[\|h\|_2^2\right]} \\
&\overset{(ii)}{=} 2\sqrt{n},
\end{aligned}$$

where for (i) we use Jensen's inequality as $\sqrt{\cdot}$ is concave. Note that here in (ii), the constant 2 is asymptotically tight.

By results of Lipschitz concentration in Section 2.2, we have $\|X\|_{\text{op}} - \mathbb{E}\left[\|X\|_{\text{op}}\right]$ is $\mathcal{O}(1)$-sub-Gaussian, so

$$\mathbb{P}\left[\left|\|X\|_{\text{op}} - \mathbb{E}\left[\|X\|_{\text{op}}\right]\right| \geq t\right] \leq 2e^{-t^2/4}.$$

Combining this concentration result with the bound on $\mathbb{E}\left[\|X\|_{\text{op}}\right]$, we have

$$\|X\|_{\text{op}} \leq (2+\epsilon)\sqrt{n}, \qquad \text{with probability } \geq 1 - 2e^{-\epsilon^2 n/4}.$$

In particular, taking $\epsilon = 1$ gives

$$\|X\|_{\text{op}} \leq 3\sqrt{n}, \qquad \text{with probability } \geq 1 - 2e^{-n/4}.$$

Remark

1. The bound we obtained here is better than the bound by matrix Bernstein inequality by a $C\sqrt{\log n}$ multiplicative factor.
2. If $X \in \mathbb{R}^{n\times m}$, we have $\mathbb{E}\left[\|X\|_{\text{op}}\right] \leq \sqrt{n} + \sqrt{m}$. The proof is left as an exercise. Furthermore, we have $\mathbb{E}\left[\lambda_{\min}(X)\right] \approx \sqrt{n} - \sqrt{m}$ assuming $n \geq m$.