Modern Topics on Statistical Learning II	Spring 23
Lecture 2 — January 31, 2023	
Prof. Qi Lei	*Scribe: Daiyan Li, Yitao Long, Andrew Zhang*

1 Big-O Notation

Definition 1. If there exists positive constants c, x_0, s.t. $f(x) \leq cg(x)$ for all $x \geq x_0$, then $f(x) = O(g(x))$. In other words,

$$\limsup_{x\to\infty} \frac{|f(x)|}{g(x)} < \infty$$

More generally, one considers

$$\limsup_{x\to a} \frac{|f(x)|}{g(x)} < \infty$$

and says $f(x) = O(g(x))$ as $x \to a$.
If $f(x) = O(g(x))$ and $g(x) = O(f(x))$, then we write $f(x) = \Theta(g(x))$ (this notation is symmetric).

Another useful notation is $f(x) = o(g(x))$ as $x \to a$ if

$$\lim_{x\to a} \frac{f(x)}{g(x)} = 0$$

2 Motivation

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be i.i.d random variables with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. We call the collection $\{f_\theta\}_{\theta\in\Theta\subset\mathbb{R}^k}$ a family of parametric functions where $f_\theta : \mathbb{R}^p \to \mathbb{R}$. Define also a loss function $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$. We are interested in

Definition 2. The mean loss or population risk of a hypothesis f_θ is defined as

$$\mathcal{R}(\theta) \coloneqq \mathop{\mathbb{E}}_{(x,y)\sim P_{X,Y}} l(f_\theta(x), y)$$

However, the joint law μ is unknown, and we only have access to

Definition 3. The empirical risk of a hypothesis f_θ is defined as

$$\hat{\mathcal{R}}(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i)$$

Since ER is intractable, we would like to know how many data size n and what probability to make us believe that $\hat{\mathcal{R}(\theta)} \to \mathcal{R}(\theta)$.
Hence it is very natural to ask for bounds on quantities of the form

$$\mathbb{P}(|Z - \mathbb{E}\, Z| \geq t) \leq f(t)$$

where $Z \in L^1$ (actually, we will need far stronger assumptions on Z) and f is non-decreasing in t. Let us illustrate this idea on the king of all distributions, the standard gaussian. We will need Markov's inequality.

Theorem 4. Let $X \geq 0$ and have finite expectation. Then for all $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}\, X}{t}$$

Proof.

$$\mathbb{E}\, X = \int \mathbb{1}_{[0,t)}(X) X \, d\mathbb{P} + \int \mathbb{1}_{[t,\infty)}(X) X \, d\mathbb{P} \geq t\mathbb{P}(X \geq t) \qquad \square$$

Remark 5. As an immediate corollary, one has Chebyshev's inequality. For $X \in L^2$ with finite expectation and variance, and $t > 0$, one has $P(|X - \mathbb{E}\, X| \geq t) \leq t^{-2}\mathrm{Var}(X)$.

Now let $Z \sim \mathcal{N}(0, 1)$. We will bound the tail probability of Z using the standard technique known as the Chernoff bound. For a fixed $t \geq 0$ and all $\lambda > 0$, one has

$$\mathbb{P}(Z \geq t) = \mathbb{P}(\lambda Z \geq \lambda t) = \mathbb{P}(e^{\lambda Z} \geq e^{\lambda t}) \leq \inf_{\lambda > 0} \frac{\mathbb{E}\, e^{\lambda Z}}{e^{\lambda t}} = \inf_{\lambda > 0} \exp(\lambda^2/2 - \lambda t) = e^{-t^2/2} \tag{1}$$

By symmetry,

$$\mathbb{P}(|Z| \geq t) \leq 2\exp(-t^2/2)$$

This shows the gaussian exhibits tails with quadratic exponential decay, and it is natural to ask if any other distributions have this property.

3 Concentration Inequalities

To get the bound in equation (1), we used that the moment generating function of Z is $\lambda \mapsto \exp(\lambda^2/2)$. Observe that for an arbitrary random variable X, as long as

$$\mathbb{E}\, e^{\lambda X} \leq e^{\lambda^2/2} \quad \forall \lambda \in \mathbb{R}$$

then the proof still goes through and we may conclude X has tails with quadratic exponential decay.

Remark 6. The observant reader will notice that Z is centered, but we never said X was centered. Actually, one can show that if X exhibits the specified condition, then X is centered (exercise).

This motivates the following definition.

Definition 7. A random variable X is σ sub-gaussian if

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

Remark 8. (2) requires the moments of X to be all finite, and the reason is

$$\mathbb{E}[e^{\lambda x}] = \mathbb{E}[\sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!}] = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[x^k]$$

where we assume $\mathbb{E}\, X = 0$ and use tylor expansion.

Using the Chernoff strategy, we immediately have tail bound for sub-Gaussian random variables.

Theorem 9. If X is σ sub-gaussian, then for all $t > 0$, we have

$$\mathbb{P}(|X - \mathbb{E}\,X| \geq t) \leq 2\exp(-t^2/(2\sigma^2)) =: \psi(t)$$

Remark 10. It is easy to show that the standard Gaussian tail satisfies

$$\mathbb{P}(Z \geq t) \underset{t\to\infty}{\sim} \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/(2\sigma^2)} =: \varphi(t)$$

Hence it is natural to ask whether the tail bound for sub-gaussians given in Theorem 9 is too loose as $\varphi(t) = o(\psi(t))$. Surprisingly, the answer is no in the following sense. One can show that if $\mathbb{E}\,e^{\lambda X} \leq \exp(\lambda^2\sigma^2/2)$ for all real λ, then for some $\tilde{\sigma}^2$ and $c > 0$, one has

$$\mathbb{P}(|X| \geq t) \leq c\mathbb{P}(|Z| \geq t), \quad Z \sim \mathcal{N}(0, \tilde{\sigma}^2)$$

Remark 11. The sub-gaussian quality is much stronger than the existence of all moments (the latter hypothesis is surprisingly rather useless. See Bilingsley's Probability and Measure section on the method of moments.) In fact, being sub-gaussian with parameter unity is equivalent to requiring that the moments decay no slower than a standard Gaussian:

$$\mathbb{E}\,|X|^p = O(p^{p/2})$$

Note that if X_1 and X_2 are independent and mean zero with sub-gaussian parameters σ_1, σ_2, then it is immediate that the convolution satisfies

$$\mathbb{E}\,e^{\lambda(X_1+X_2)} \leq \exp(\lambda^2(\sigma_1^2 + \sigma_2^2)/2)$$

hence $X_1 + X_2$ follows $\sqrt{\sigma_1^2 + \sigma_2^2}$ sub-gaussian. (In fact the sub-gaussian property is preserved even under dependent sums, but we leave this as an exercise to the reader. Hint: conditional Cauchy-Schwartz and Jensen will be helpful.) Using this observation, we are immediately led to Hoeffding's inequality.

Theorem 12. Let $X_1, \ldots, X_n$ be independent with sub-gaussian parameters $\sigma_1, \ldots, \sigma_n$. Then

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

In particular, if the X_i are compactly supported on $[a_i, b_i]$, then one has

$$\mathbb{P}\left[\left|\sum X_i - \mathbb{E}\,X_i\right| \geq t\right] \leq 2\exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

That X_i compactly supported on $[a_i, b_i]$ has sub-gaussian paramter $(b-a)/2$ is homework.

Remark 13. As noted before, the sub-gaussian property is preserved even under dependent sums. Hence once may state a Hoeffding-type result without the random sample assumption.

We also conclude the concentration inequality in functions of random variables. To set the stage, we need a definition.

Definition 14. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy the bounded differences inequality with parameters $c_1, \ldots, c_n$ if

$$\left| f(x) - f(x^{\setminus i}) \right| \leq c_i$$

where $x^{\setminus i}$ has arbitrary ith coordinate and jth coordinate given by x_j for $i \neq j$.

The last result is known as McDiarmid's Inequality or also the Azuma-Hoeffding Inequality.

Theorem 15. Let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the bounded differences inequality with parameters $c_1, \ldots, c_n$ and $X_1, \ldots, X_n$ be i.i.d random variables. Then

$$\mathbb{P}(|f(X_1, \ldots, X_n) - \mathbb{E}\, f(X_1, \ldots, X_n)| \geq t) \leq 2 \exp\left(\frac{-t^2}{2 \sum_{i=1}^n c_i^2} \right)$$

McDiarmid's Inequality looks very similar to Hoeffding's Inequality, and Hoeffding's Inequality for compactly supported convolutions can in fact be proved as a corollary to McDiarmid's inequality.

4 Appendix: Proof of McDiarmid's Inequality

To prove McDiarmid's Inequality, we will need conditional expectation and martingales.

Definition 16. Let $(\Omega, \mathcal{F}_0, \mathbb{P})$ be a probability space and $X \in \mathcal{F}_0$ satisfy $\mathbb{E}\,|X| < \infty$. Given a sub σ-field $\mathcal{F} \subset \mathcal{F}_0$, the conditional expectation of X given $\mathcal{F}$ is any random variable Y satisfying

1. $Y \in \mathcal{F}$.
2. For all $B \in \mathcal{F}$,
$$\int_B X \, d\mathbb{P} = \int_B Y \, d\mathbb{P}$$

Such a random variable is denoted by $\mathbb{E}(X \mid \mathcal{F})$. Existence of conditional expectation is given by the Radon-Nikodym theorem, but not uniqueness. In general, there are many random variables satisfying the above two properties, say Y and Z, but we do have $Y = Z$ almost surely. Hence to be formal, one would say $\mathbb{E}(X \mid \mathcal{F})$ is a version of the conditional expectation.

Some very useful properties are (note all (in)equalities here hold almost surely):

1. $\mathbb{E}\,\mathbb{E}(X \mid \mathcal{F}) = \mathbb{E}\, X$. This is called law of iterated expectation.
2. If $Y \in \mathcal{F}$, then $\mathbb{E}(XY \mid \mathcal{F}) = Y\, \mathbb{E}(X \mid \mathcal{F})$. This is called the take out what you know property.
3. If $\mathcal{F} \subset \mathcal{G}$, then $\mathbb{E}(\mathbb{E}(X \mid \mathcal{G}) \mid \mathcal{F}) = \mathbb{E}(X \mid \mathcal{F})$, as well as $\mathbb{E}(\mathbb{E}(X \mid \mathcal{F}) \mid \mathcal{G}) = \mathbb{E}(X \mid \mathcal{F})$. This is called the smaller σ-algebra always wins property, or sometimes the tower property.
4. Let $\mathbb{E}\,|f(X, Y| < \infty$. Suppose $Y \in \mathcal{F}$. Then $\mathbb{E}(f(X, Y) \mid \mathcal{F}) = \varphi(Y)$ where $\varphi : y \mapsto \mathbb{E}(f(X, y))$. This has no name, but is nevertheless standard.

5. Let $X, Y \in L^1$ and $a, b \in \mathbb{R}$. Then

$$\mathbb{E}(aX + bY \mid \mathcal{F}) = a\,\mathbb{E}(X \mid \mathcal{F}) + b\,\mathbb{E}(Y \mid \mathcal{F})$$

   This is called linearity of conditional expectation.

6. Let $X, Y \in L^1$ with $X \leq Y$ almost surely. Then $\mathbb{E}(X \mid \mathcal{F}) \leq \mathbb{E}(Y \mid \mathcal{F})$. This is called monotonicity of conditional expectation.

7. Let φ be convex with $\mathbb{E}\,|\varphi(X)| < \infty$. Then $\mathbb{E}(\varphi(X) \mid \mathcal{F}) \geq \varphi(\mathbb{E}(X \mid \mathcal{F}))$. This is called conditional Jensen's inequality.

We will use all of these in what follows.

Definition 17. Let $(\mathcal{F}_n)_{n\geq 0}$ be a sequence of σ-algebras. $(\mathcal{F}_n)_n$ is a filtration if $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots$. We say the sequence $(X_n)_n$ is adapted to the filtration $\mathcal{F}_n$ if $X_n \in \mathcal{F}_n$ for all n. If in addition

1. $\mathbb{E}\,|X_n| < \infty$ for all n.
2. $\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) = X_n$ for all n.

then we say $(X_n)_n$ is a $(\mathcal{F}_n)_n$-martingale.

We are now ready to begin proving McDiarmid's inequality. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a map satisfying $\mathbb{E}\,|f(X_1, \dots, X_n)| < \infty$. Put $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$. Define the sequence $Y_k \coloneqq \mathbb{E}(f(X_1, \dots, X_n) \mid \mathcal{F}_k)$ and $Y_0 = \mathbb{E}(f(X_1, \dots, X_n))$. Then

$$f(X_1, \dots, X_n) - \mathbb{E}\,f(X_1, \dots, X_n) = \sum_{i=1}^{n} \underbrace{Y_n - Y_{n-1}}_{=:D_n} = \sum_{i=1}^{n} D_n \tag{3}$$

We claim Y_k is a $\mathcal{F}_k$ martingale.

Proof. We first need to check $Y_k \in \mathcal{F}_k$. Of course $Y_k \in \mathcal{F}_k$. Now we check finite expectation

$$\mathbb{E}\,|Y_k| = \mathbb{E}\,|\mathbb{E}(f(X_1, \dots, X_n) \mid \mathcal{F}_k)| \leq \mathbb{E}\,\mathbb{E}(|f(X_1, \dots, X_n)| \mid \mathcal{F}_k) = \mathbb{E}\,|f(X_1, \dots, X_n)| < \infty$$

Finally check the so-called fair game property.

$$\mathbb{E}(Y_{k+1} \mid \mathcal{F}_k) = \mathbb{E}(\mathbb{E}(f(X_1, \dots, X_n) \mid \mathcal{F}_{k+1}) \mid \mathcal{F}_k) = \mathbb{E}(f(X_1, \dots, X_n) \mid \mathcal{F}_k) = Y_k \qquad \square$$

The sequence $D_n = Y_n - Y_{n-1}$ in equation (3) is known as a martingale difference sequence and $Y_n = Y_0 + \sum_{i=1}^{n} D_i$. Three important facts about the decomposition D_n are

1. D_k is adapted to $\mathcal{F}_k$ or D_k is $\mathcal{F}_k$ measurable for each k.
2. $\mathbb{E}\,|D_k| < \infty$ for all k.
3. $\mathbb{E}(D_{k+1} \mid \mathcal{F}_k) = 0$.

Proof. To prove $D_k \in \mathcal{F}_k$, recall $D_k = Y_k - Y_{k-1}$ and $Y_k \in \mathcal{F}_k, Y_{k-1} \in \mathcal{F}_{k-1} \subset \mathcal{F}_k$.

To prove $D_k \in L^1$,

$$\mathbb{E}\,|D_k| = \mathbb{E}\,|Y_k - Y_{k-1}| \leq \mathbb{E}\,|Y_k| + \mathbb{E}\,|Y_{k-1}| < \infty$$

Finally,

$$\mathbb{E}(D_{k+1} \mid \mathcal{F}_k) = \mathbb{E}(Y_{k+1} - Y_k \mid \mathcal{F}_k) = \mathbb{E}(Y_{k+1} \mid \mathcal{F}_k) - \mathbb{E}(Y_k \mid \mathcal{F}_k) = Y_k - Y_k = 0 \qquad \square$$

The following result is called Azuma-Hoeffding's inequality.

Theorem 18. Let Y_n be a $\mathcal{F}_n$ martingale. Suppose the martingale differences $D_n = Y_n - Y_{n-1}$ are bounded almost surely, that is, there exists a sequence c_n such that $\mathbb{P}(|D_n| \leq c_n) = 1$. Then

$$\mathbb{P}(|Y_n - Y_0| \geq t) \leq 2\exp\left(\frac{-t^2}{2\sum_{j=1}^n c_j^2}\right)$$

Proof. Just about every result comes from the Chernoff strategy, including this one.

$$\mathbb{P}(Y_n - Y_0 \geq t) \leq \inf_{\lambda > 0} e^{-\lambda t}\,\mathbb{E}\,e^{\lambda(Y_n - Y_0)}$$

So we need to bound $\mathbb{E}\exp(\lambda(Y_n - Y_0))$. We have

$$\mathbb{E}(\exp(\lambda(Y_n - Y_0)) \mid \mathcal{F}_{n-1}) = e^{\lambda(Y_{n-1} - Y_0)}\,\mathbb{E}(e^{\lambda D_n} \mid \mathcal{F}_{n-1})$$

Now, $|D_n| \leq c_n$ almost surely, and thus $\mathbb{E}(e^{\lambda D_n} \mid \mathcal{F}_{n-1}) \leq e^{\lambda^2 c_n^2/2}$. This can be proven using regular conditional distributions. Hence

$$\mathbb{E}\exp(\lambda(Y_n - Y_0)) \leq e^{\lambda^2 c_n^2/2}\,\mathbb{E}\,e^{\lambda(Y_{n-1} - Y_0)}$$

Induction shows

$$\mathbb{E}\exp(\lambda(Y_n - Y_0)) \leq e^{\lambda^2\left(\sum_{i=1}^n c_i^2\right)/2}$$

Hence

$$\mathbb{P}(Y_n - Y_0 \geq t) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\right)$$

Symmetry and union bound shows

$$\mathbb{P}(|Y_n - Y_0| \geq t) \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\right) \qquad \square$$

Finally, we have McDiarmid's Inequality.

Theorem 19. Let $X = (X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ independent. Suppose $\mathbb{E}\,|f(X)| < \infty$ and that f satisfies the bounded differences inequality with parameters $c_1, \ldots, c_n$. Then

$$\mathbb{P}(|f(X) - \mathbb{E}\,f(X)| \geq t) \leq 2\exp\left(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\right)$$

Proof. The strategy is clear. Put $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$. Define $Y_k = \mathbb{E}(f(X) \mid \mathcal{F}_k)$ for $1 \leq k \leq n$ and $Y_0 = \mathbb{E}\, f(X)$. Set $D_k = Y_k - Y_{k-1}$ for $1 \leq k \leq n$. We need to show $|D_k| \leq c_k$ almost surely, and then apply Azuma-Hoeffding to conclude.

We have

$$\begin{aligned} D_k &= Y_k - Y_{k-1} \\ &= \mathbb{E}(f(X) \mid \mathcal{F}_k) - \mathbb{E}(f(X) \mid \mathcal{F}_{k-1}) \\ &= \underbrace{\mathbb{E}(f(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) \mid \mathcal{F}_k)}_{=:\varphi(X_1, \ldots, X_k)} - \underbrace{\mathbb{E}(f(X_1, \ldots, X_{k-1}, X_k, X_{k+1}, \ldots, X_n) \mid \mathcal{F}_{k-1})}_{=:\psi(X_1, \ldots, X_{k-1})} \end{aligned}$$

where $\varphi(x_1, \ldots, x_k) = \mathbb{E}\, f(x_1, \ldots, x_k, X_{k+1}, \ldots, X_n)$ and $\psi(x_1, \ldots, x_{k-1}) = \mathbb{E}\, f(x_1, \ldots, x_{k-1}, X_k, \ldots, X_n)$. Now,

$$\begin{aligned} |\varphi - \psi| &\leq \mathbb{E}\, |f(x_1, \ldots, x_{k-1}, x_k, X_{k+1}, \ldots, X_n) - f(x_1, \ldots, x_{k-1}, X_k, X_{k+1}, \ldots, X_n)| \\ &\leq c_k \end{aligned}$$

as desired. □

Remark 20. The notation $|\varphi - \psi|$ is sloppy, but the meaning should be clear in the context of what we are trying to do, i.e. bound D_k.

Remark 21. Note the following is **not** correct.

$$\begin{aligned} \sup |\varphi - \psi| &\leq \mathbb{E} \sup |f(x_1, \ldots, x_{k-1}, x_k, X_{k+1}, \ldots, X_n) - f(x_1, \ldots, x_{k-1}, X_k, X_{k+1}, \ldots, X_n)| \\ &\leq c_k \end{aligned}$$

It is **not** clear that the quantity inside the expectation is even measurable.

References

[1] Billingsley, P. (1995). Probability and measure. Wiley.

[2] Durrett, R. (2019). Probability: Theory and Examples (5th ed., Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press. doi:10.1017/9781108591034

[3] Vershynin, R. (2018). High-Dimensional Probability: An Introduction with Applications in Data Science (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press. doi:10.1017/9781108231596

[4] Wainwright, M. (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge: Cambridge University Press. doi:10.1017/9781108627771