CUHK STAT6050: Statistical Learning Theory Spring 2026

Lecture 3: Uniform concentration inequality

January 22, 2026

Lecturer: Ben Dai

"*There is Nothing More Practical Than A Good Theory.*" — Kurt Lewin

1 Introduction

As indicated in Lecture 2, we will focus on the asymptotics of the empirical process of the estimation error; specifically, we aim to find a $\delta_n \to 0$ for any small $\varepsilon > 0$,

$$\mathbb{P}\big(\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)| \geq \varepsilon\big) \leq \delta_n$$

Motivated by **concentration**, how a random variable deviates from its expectation, rewrite the probability as:

$$\mathbb{P}\big(\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)| \geq \varepsilon - \mathbb{E}\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)|\big) \leq \delta_n.$$

To investigate this bound, we itemize two aims:

- **A1.** The asymptotics of
$$\mathbb{E}\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)|.$$
- **A2.** The concentration inequality of
$$\mathbb{P}\big(\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)| \geq \varepsilon\big).$$

For **A1**, the minimum requirement is that

$$\mathbb{E}\sup_{f \in \mathscr{F}} |\widehat{R}_n(f) - R(f)| = o(1),$$

to ensure the asymptotic vanishing of (the upper bound of) the estimation error. For **A2**, we can regard it as a uniform version of concentration inequalities.

2 From pointwise to uniform

Before diving into uniform concentration inequalities, let's have a brief review of the fundamental pointwise concentration inequalities.

Theorem 2.1 (Markov's inequality)**.** *Let Z be a non-negative random variable. Then for any $t > 0$,*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}(Z)}{t}.$$

Theorem 2.2 (Chebyshev's inequality)**.** *Let Z be a random variable with mean μ and variance σ^2. Then for any $t > 0$,*

$$\mathbb{P}(|Z-\mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Lemma 2.3 (Hoeffding's Lemma)**.** *Let Z be a random variable such that $a \leq Z \leq b$ almost surely, and $\mathbb{E}(Z) = 0$. Then for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}(e^{\lambda Z}) \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

The above inequalities are fundamental for "pointwise" concentration, where we consider a fixed random variable. In learning theory, we often need "uniform" concentration to bound the error over a class of functions.

2.1 From Hoeffding's inequality to McDiarmid's inequality

Theorem 2.4 (Hoeffding's Inequality (1963))**.** *Suppose $Z_1, \cdots, Z_n$ are independent random variables such that $a_i \leq Z_i \leq b_i$ almost surely, then for any $\varepsilon > 0$:*

$$\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Z_i) \geq \varepsilon\big) \leq \exp\big(\frac{-2n^2\varepsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\big).$$

Note that we cannot use Hoeffding's inequality to bound **A2**, since there is a supremum on the average. McDiarmid's inequality is a generalization of Hoeffding's inequality, which enables us to directly bound the probabilistic bound in **A2**.

Theorem 2.5 (McDiarmid's inequality (1989))**.** *Suppose $Z_1, \cdots, Z_n$ are independent random variables, and there is a function $g : \mathbb{R}^n \to \mathbb{R}$ such that the variation on i-th coordinate is upper bounded, that is, for all $i = 1, \cdots, n$ and all $(z_1, \cdots, z_i, z_i', \cdots z_n)$,*

$$|g(z_1, \cdots, z_i, \cdots, z_n) - g(z_1, \cdots, z_i', \cdots, z_n)| \leq c_i.$$

Then,

$$\mathbb{P}\big(g(Z_1, \cdots, Z_n) - \mathbb{E}g(Z_1, \cdots, Z_n) \geq \varepsilon\big) \leq \exp\big(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\big).$$

McDiarmid's inequality shares the same spirit as Hoeffding's inequality but applies to general functions $g(Z_1,\cdots,Z_n)$ satisfying the bounded difference property. We demonstrate McDiarmid's inequality for our Aim **A2**.

Let $Z_i = l(\mathbf{Y}_i, f(\mathbf{X}_i))$, and

$$g(Z_1,\cdots,Z_n) = \sup_{f\in\mathscr{F}} \big(\widehat{R}_n(f) - R(f)\big) = \sup_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\big(Z_i - \mathbb{E}l(\mathbf{Y}, f(\mathbf{X}))\big).$$

Assume that $0 \le l(\mathbf{Y}_i, f(\mathbf{X}_i)) \le U$, we have

$$|g(z_1,\cdots,z_i,\cdots,z_n) - g(z_1,\cdots,z_i',\cdots,z_n)| \le \big|\sup_{f\in\mathscr{F}} \frac{1}{n}\big(z_i - z_i'\big)\big| = U/n.$$

Then, McDiarmid's inequality yields that

$$\mathbb{P}\big(\sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}\sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| \ge \varepsilon\big) \le \exp\big(-\frac{2n\varepsilon^2}{U^2}\big).$$

We summarize the result as the following corollary.

Corollary 2.6. *Suppose the loss function $l(\cdot,\cdot)$ is uniformly bounded by a constant U. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\big(\sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}\sup_{f\in\mathscr{F}} |\widehat{R}_n(f) - R(f)| \ge \varepsilon\big) \le \exp\big(-\frac{2n\varepsilon^2}{U^2}\big).$$

Remark 2.7 (Boundedness). Both Hoeffding's and McDiarmid's inequalities rely solely on the *boundedness* assumption. McDiarmid's inequality can be regarded as the uniform counterpart to Hoeffding's inequality, extending the concentration of sums to general functions, including suprema.

2.2 From Bernstein's inequality to Talagrand's inequality

Hoeffding's inequality is robust but often loose, as it does not exploit the variance information. Bernstein's inequality offers a refinement by incorporating the *variance*, yielding tighter bounds when the variance is small.

Theorem 2.8 (Bernstein's inequality (1920s))**.** *Let $Z_1,\cdots,Z_n$ be independent random variables with $|Z_i| \le U$ almost surely, for all $i = 1,\cdots,n$. Then, for all $\varepsilon > 0$,*

$$\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Z_i) \ge \varepsilon\big) \le \exp\big(-\frac{n\varepsilon^2}{2\sigma^2 + 2U\varepsilon/3}\big),$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(Z_i)$.

Establishing a uniform Bernstein's inequality is a much harder problem, which was solved by *Talagrand* [Talagrand, 1996b, Talagrand, 1996a].

Theorem 2.9 (Talagrand's inequality (1996))**.** *Suppose the loss function* $l(\cdot,\cdot)$ *is uniformly bounded by a constant* U*. Then, for any* $\varepsilon > 0$*,*

$$\mathbb{P}\Big(\sup_{f\in\mathscr{F}}|\widehat{R}_n(f)-R(f)|-\mathbb{E}\sup_{f\in\mathscr{F}}|\widehat{R}_n(f)-R(f)|\geq\varepsilon\Big)\leq K\exp\Big(-\frac{1}{K}\frac{\varepsilon}{U}\log(1+\frac{\varepsilon U}{nV})\Big),$$

where $K>0$ *is a universal constant and* V *is any constant satisfying*

$$V\geq\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\Big(l\big(\mathbf{Y}_i,f(\mathbf{X}_i)\big)-\mathbb{E}l\big(\mathbf{Y}_i,f(\mathbf{X}_i)\big)\Big)^2.$$

Talagrand's inequality can be viewed as the *uniform* counterpart to Bernstein's inequality. The constant V (often referred to as the *wimpy variance*) is analogous to the variance σ^2, capturing the complexity of the function class. Just as Bernstein improves upon Hoeffding, Talagrand improves upon McDiarmid by exploiting the variance structure of the empirical process. However, finding a tight constant V is challenging. Now, given the results of Talagrand's inequality, we slightly modify our aims:

- **A1.** The asymptotics of
$$\mathbb{E}\sup_{f\in\mathscr{F}}|\widehat{R}_n(f)-R(f)|.$$

- **A2′.** Find a tight constant V such that
$$V\geq\mathbb{E}\sup_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\Big(l\big(\mathbf{Y}_i,f(\mathbf{X}_i)\big)-\mathbb{E}l\big(\mathbf{Y}_i,f(\mathbf{X}_i)\big)\Big)^2.$$

In the sequel, we will show that **A1** and **A2′** are closely related.

References

[Talagrand, 1996a] Talagrand, M. (1996a). New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563.

[Talagrand, 1996b] Talagrand, M. (1996b). A new look at independence. *The Annals of probability*, pages 1–34.