# MATH 281C: Mathematical Statistics

Lecture 4

## 1 Bennett's Inequality

Let us recall the Hoeffding inequality from last lecture. It states that

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

for every $t \geq 0$, where $X_1, \ldots, X_n$ are independent random variables with $a_i \leq X_i \leq b_i$ almost surely. We remarked that when $\sum_{i=1}^{n} \mathrm{var}(X_i)$ is much smaller than $\sum_{i=1}^{n}(b_i - a_i)^2/4$, and when the CLT holds, the tail bound given by Hoeffding can be loose. Bennett's inequality attempts to give tail bounds which involve variances.

**Theorem 1.1** (Hoeffding's inequality)**.** Suppose $X_1, \ldots, X_n$ are independent random variables having finite variances. Suppose $X_i \leq B$ almost surely for each $i = 1, \ldots, n$ (here $B$ is deterministic). Let $V = \sum_{i=1}^{n} \mathbb{E}X_i^2$. Then, for every $t \geq 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left\{-\frac{V}{B^2}h\left(\frac{tB}{V}\right)\right\} \tag{4.1}$$

where

$$h(u) := (1+u)\log(1+u) - u, \quad u \geq 0. \tag{4.2}$$

**Remark 1.1.** Bennett's inequality, as stated above, gives only the upper tail bound. To get the lower bound, one needs to impose the assumption $X_i \geq -B$. In this case, we get

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \leq -t\right\} \leq \exp\left\{-\frac{V}{B^2}h\left(\frac{tB}{V}\right)\right\}.$$

**Remark 1.2.** For the function $h$ defined in (4.2), it is easy to see that

$$h(0) = 0, \quad h'(0) = 0 \quad \text{and} \quad h''(0) = 1.$$

Therefore, for $u$ near 0, we have $h(u) \approx u^2/2$. Hence, when $tB/V$ is small, the bound given by Bennett's inequality looks like

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left\{-\frac{V}{B^2}h\left(\frac{tB}{V}\right)\right\} \approx \exp\left(\frac{-t^2}{2V}\right).$$

Thus Bennett's inequality gives Gaussian-like tails with $V = \sum_{i=1}^n \mathbb{E}X_i^2$ in some regimes.

As an example, suppose $\mathbb{E}X_i = 0$, $\mathrm{var}(X_i) = \sigma^2$ and $X_i \le 1$. Then $V = n\sigma^2$ and Bennett's inequality gives

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \ge t\right) \le \exp\left\{-Vh\left(\frac{n^{1/2}t}{V}\right)\right\} = \exp\left\{-n\sigma^2 h\left(\frac{n^{1/2}t}{V}\right)\right\}.$$

When $t$ is small compared to $n^{1/2}\sigma^2$, we get a Gaussian type bound.

Before proving Theorem 1.1, let us first simplify Bennett's inequality by taking a closer look at the function $h$ given in (4.2). It can be shown that (homework) for any $u \ge 0$,

$$h(u) = (1+u)\log(1+u) - u \ge \frac{u^2}{2(1+u/3)}.$$

This leads to the following result, which is known as Bernstein's inequality.

**Theorem 1.2.** Suppose $X_1, \dots, X_n$ are independent random variables with finite variances, and suppose that $\max_{1\le i\le n} |X_i| \le B$ almost surely for some constant $B > 0$. Let $V = \sum_{i=1}^n \mathbb{E}X_i^2$. Then, for every $t \ge 0$,

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \ge t\right\} \le \exp\left\{\frac{-t^2}{2(V + tB/3)}\right\}$$

and

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \le -t\right\} \le \exp\left\{\frac{-t^2}{2(V + tB/3)}\right\}.$$

**Remark 1.3.** There is a version of Bernstein's inequality that replaces the boundedness assumption by weaker moment restrictions. That is, assume moment of any order exists, and the $k$-th moment is subject to certain growth condition for every $k \ge 2$. See Theorem 2.10 in Boucheron, Lugosi and Massart (2013).

*Proof of Theorem 1.1.* Without loss of generality, we assume $B = 1$; otherwise, it suffices to work with $X_1/B, \dots, X_n/B$.

This proof relies on the following observation. Let $\phi : \mathbb{R} \to \mathbb{R}$ denote the function $\phi(u) = e^u - u - 1$. The map $u \mapsto \phi(u)/u^2$ is increasing on $\mathbb{R}$ with $\phi(0)/0^2 = 1/2$. I will leave the verification of this fact as homework.

Let $S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$. Then, for every $\lambda \ge 0$ as before,

$$\mathbb{P}(S \ge t) \le e^{-\lambda t}\mathbb{E}e^{\lambda \sum_{i=1}^n (X_i - \mathbb{E}X_i)} = e^{-\lambda t}e^{-\lambda \sum_{i=1}^n \mathbb{E}X_i}\prod_{i=1}^n \mathbb{E}e^{\lambda X_i}, \tag{4.3}$$

where we used the independence of $X_1, \dots, X_n$. Now because $X_i \le 1$, we have $\lambda X_i \le \lambda$. Using the monotonicity of $u \mapsto \phi(u)/u^2$, we deduce that

$$\frac{\phi(\lambda X_i)}{(\lambda X_i)^2} \le \frac{\phi(\lambda)}{\lambda^2},$$

which implies that

$$e^{\lambda X_i} \leq \lambda X_i + 1 + X_i^2 \phi(\lambda).$$

Using this bound in the right-hand side of (4.3), we obtain

$$\mathbb{P}(S \geq t) \leq e^{-\lambda t - \lambda \sum_{i=1}^n \mathbb{E}X_i} \prod_{i=1}^n \{1 + \lambda \mathbb{E}X_i + \phi(\lambda)\mathbb{E}X_i^2\}.$$

By the trivial inequality $1 + x \leq e^x$,

$$1 + \lambda \mathbb{E}X_i + \phi(\lambda)\mathbb{E}X_i^2 \leq e^{\lambda \mathbb{E}X_i + \phi(\lambda)\mathbb{E}X_i^2},$$

implying

$$\mathbb{P}(S \geq t) \leq e^{-\lambda t + \phi(\lambda)V},$$

valid for every $\lambda \geq 0$. We now optimize the above bound by taking the derivative with respect to $\lambda$ and setting it equal to zero to obtain:

$$-t + V(e^\lambda - 1) = 0 \;\Rightarrow\; \lambda = \log(1 + t/V).$$

For this value of $\lambda$, it is straightforward to obtain (4.1). □

The two bounds in Bernstein's inequality can be combined to write

$$\mathbb{P}\left\{\left|\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right| \geq t\right\} \leq 2\exp\left\{\frac{-t^2}{2(V + tB/3)}\right\}.$$

We can now attempt to find the value of $t$ which makes the bound on the right-hand side above exactly equal to $\alpha$, i.e., we want to solve the equation

$$2\exp\left\{\frac{-t^2}{2(V + tB/3)}\right\} = \alpha.$$

This leads to the quadratic equation

$$t^2 - \frac{2B\log(2/\alpha)}{3}\, t - 2V\log(2/\alpha) = 0$$

whose nonnegative solution is given by

$$t = \frac{B\log(2/\alpha)}{3} + \sqrt{\frac{B^2\log^2(2/\alpha)}{9} + 2V\log(2/\alpha)} \leq \sqrt{2V\log(2/\alpha)} + \frac{2B\log(2/\alpha)}{3}.$$

Thus Bernstein's inequality implies that, with probability at least $1 - \alpha$,

$$\left|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right| \leq \frac{\sqrt{2V\log(2/\alpha)}}{n} + \frac{2B\log(2\alpha)}{3n}.$$

Now if $X_1, \ldots, X_n$ are iid with mean zero, variance $\sigma^2$ and bounded in absolute value by $B$, then $V = n\sigma^2$ so that the inequality

$$|\bar{X}_n| \leq \sigma\sqrt{\frac{2\log(2/\alpha)}{n}} + \frac{2B\log(2/\alpha)}{3n} \tag{4.4}$$

holds with probability at least $1 - \alpha$. Note that if $X_1, \ldots, X_n$ are iid normal, then $\bar{X}_n$ is normal and $|\bar{X}_n|$ will be bounded by the first term in the right-hand side above with probability at least $1 - \alpha$. Therefore, the deviation bound (4.4) agrees with the normal approximation bound except for the smaller order term (which if of order $1/n$; the leading term being of order $1/\sqrt{n}$).

## 2 Concentration of Supremum of Empirical Process

Let us now study the concentration behavior of

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right|. \tag{4.5}$$

We first introduce some notation. We denote the empirical (probability) measure of $X_1, \ldots, X_n$ by $P_n$. The probability measure $P_n$ has the CDF $F_n(x) = (1/n)\sum_{i=1}^{n} I(X_i \le x)$. The common distribution of the iid random observations $X_1, \ldots, X_n$ will be denoted by $P$. We also write

$$Pf = \mathbb{E}f(X_1) \quad \text{and} \quad P_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

The quantity in (4.5) can therefore be written as

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \quad \text{or} \quad \sup_{f\in\mathcal{F}} |(P_n - P)f|.$$

The concentration inequality that we proved via the bounded differences inequality is the following. Suppose that $\mathcal{F}$ consists of functions that are uniformly bounded by $B$, then

$$\sup_{f\in\mathcal{F}} |P_n f - Pf| \le \mathbb{E}\left(\sup_{f\in\mathcal{F}} |P_n f - Pf|\right) + B\sqrt{\frac{2\log(1/\alpha)}{n}} \tag{4.6}$$

with probability $1 - \alpha$.

We remarked previously that when $\mathrm{var}(f(X_1))$ is small compared to $B$ for every $f \in \mathcal{F}$, this inequality is not sharp. In such situations, it is much more helpful to use *Talagrand's concentration inequality* for the suprema of empirical processes, which is stronger than (4.6) and also deeper and harder to prove. We shall give the statement of this inequality but not the proof (for a proof, you can refer to Section 12.4 in Boucheron, Lugosi and Massart (2013)). Before stating Talagrand's inequality, let us look at a statistical application where it becomes necessary to deal with function classes $\mathcal{F}$ where the variances are small compared to the uniform (upper) bound. This application concerns the regression problem (it also applies similarly to the classification problem).

**Example 2.1** (Bounded Regression)**.** We have two random objects $X$ and $Y$ taking values in spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume that $\mathcal{Y}$ is a bounded subinterval of the real line. The problem is to predict $Y \in \mathcal{Y}$ on the basis of $X \in \mathcal{X}$. A predictor (or estimator) is any function $g$ which maps $\mathcal{X}$ to $\mathbb{R}$. The (test) error of an estimator $g$ is defined by

$$L(g) := \mathbb{E}\{Y - g(X)\}^2.$$

The goal of regression is to construct an estimator with small error based on $n$ iid observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ having the same distribution as $(X, Y)$. For an estimator $g$, its empirical error is given by

$$L_n(g) := \frac{1}{n}\sum_{i=1}^{n}\{Y_i - g(X_i)\}^2.$$

A natural strategy is to select a class of predictors $\mathcal{G}$, and then to choose the predictor in $\mathcal{G}$ which has the smallest empirical error, i.e.,

$$\widehat{g}_n := \operatorname*{argmin}_{g \in \mathcal{G}} L_n(g).$$

The key question now is how good the predictor $\widehat{g}_n$ is in terms of test error, i.e., how small is its error

$$L(\widehat{g}_n) := \mathbb{E}[\{Y - \widehat{g}_n(X)\}^2 | X_1, Y_1, \ldots, X_n, Y_n].$$

In particular, we are interested in how small $L(\widehat{g}_n)$ is compared to $\inf_{g \in \mathcal{G}} L(g)$. Suppose that this infimum is achieved at some $g^* \in \mathcal{G}$ (oracle). To bound $L(\widehat{g}_n) - L(g^*)$ ($\geq 0$), it is natural to write

$$\begin{aligned} L(\widehat{g}_n) - L(g^*) &= L(\widehat{g}_n) - L_n(\widehat{g}_n) + \underbrace{L_n(\widehat{g}_n) - L_n(g^*)}_{\leq 0} + L_n(g^*) - L(g^*) \\ &\leq L(\widehat{g}_n) - L_n(\widehat{g}_n) + L_n(g^*) - L(g^*). \end{aligned}$$

We can now use the Empirical Process Notion. Let $P$ denote the joint distribution of $(X, Y)$, and $P_n$ denote the empirical distribution of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $\mathcal{F}$ denote the class of all functions $(x, y) \mapsto \{y - g(x)\}^2$ as $g$ varies over $\mathcal{G}$.

With this notation, the above inequality becomes

$$P(\widehat{f}_n - f^*) \leq (P - P_n)(\widehat{f}_n - f^*), \tag{4.7}$$

where $\widehat{f}_n(x, y) := \{y - \widehat{g}_n(x)\}^2$ and $f^*(x, y) := \{y - g^*(x)\}^2$. In order to proceed further, we need to bound the right-hand side above. A crude bound is

$$(P - P_n)(\widehat{f}_n - f^*) \leq 2 \sup_{f \in \mathcal{F}} |P_n f - Pf|. \tag{4.8}$$

If we now assume that the class of functions $\mathcal{F}$ is uniformly bounded by $B$, we can use the concentration inequality (4.6). This will give some bound on $L(\widehat{g}_n) - L(g^*)$ provided one can control the expectation (we will discuss this topic later). It is important now to note that this method will never give a bound better than $n^{-1/2}$ for $L(\widehat{g}_n) - L(g^*)$. This is because there is already a term of $n^{-1/2}$ in the right-hand side of (4.6). But in regression, at least for small classes $\mathcal{G}$ (such as finite dimensional function class), we would expect the test error to decay much faster than $n^{-1/2}$ (such as at the $n^{-1}$ rate). Such fast rates cannot be proved by this method.

To prove faster rates, one needs to use a technique called "localization" instead of the crude bound (4.8). Let $\widehat{\delta}$ denote the left-hand side of (4.7), and the goal is to derive upper bounds for $\widehat{\delta}$. Inequality (4.8) implies that

$$\widehat{\delta} \leq \sup_{f \in \mathcal{F} : P(f - f^*) \leq \widehat{\delta}} (P - P_n)(f - f^*).$$

This is a bit complicated because the class of functions in the supremum depends on $\widehat{\delta}$ and hence is random. To simplify the problem, let us ignore this for the moment and focus on

$$\sup_{f \in \mathcal{F} : P(f - f^*) \leq \delta} (P - P_n)(f - f^*) \tag{4.9}$$

for a deterministic but small $\delta$. The key is to realize that the functions involved here have small variances (at least in the well specified case where $g^*(x) = \mathbb{E}(Y|X = x)$). Indeed, in the well specified case, we have

$$P(f - f^*) = \mathbb{E}[\{Y - g(X)\}^2 - \{Y - g^*(X)\}^2] = \mathbb{E}\{g(X) - g^*(X)\}^2.$$

Hence when $P(f - f^*) \le \delta$, we have

$$\begin{aligned}
\mathrm{var}(f - f^*) &\le \mathbb{E}\{f(X,Y) - f^*(X,Y)\}^2 \\
&= \mathbb{E}[\{Y - g(X)\}^2 - \{Y - g^*(X)\}^2]^2 \\
&= \mathbb{E}[\{2Y - g(X) - g^*(X)\}\{g(X) - g^*(X)\}]^2 \\
&\le C_B \mathbb{E}\{g(X) - g^*(X)\}^2 \le C_B \delta.
\end{aligned}$$

If we use the concentration inequality (4.6) to control (4.9), the resulting bound will be at least $C_B n^{-1/2}$ that is independent of $\delta$. This will not lead to any faster rates. However, Talagrand's inequality will give a better bound under small variances. Together with suitable bounds on the expectation, one will obtain faster rates for regression under appropriate assumptions on $\mathcal{G}$.

Similar analysis can be done for classification under certain assumptions.

Let us now state Talagrand's concentration inequality for empirical processes. As before, assume that $\mathcal{F}$ is uniformly bounded by a constant $B$. Then, letting $Z := \sup_{f\in\mathcal{F}} |P_n f - Pf|$, we have

$$Z \le C_1 \mathbb{E}(Z) + C_2 \sqrt{\sup_{f\in\mathcal{F}} \mathrm{var}(f(X_1)) \frac{\log(1/\alpha)}{n}} + C_3 \frac{B \log(1/\alpha)}{n}$$

with probability at least $1 - \alpha$. Here $C_1$–$C_3$ are universal constants which can be made explicit. Note that the leading terms are $\mathbb{E}(Z)$ and the second term which only involves the variances. The last term is of order $1/n$. Bousquet (2003) proved the following version of Talagrand's inequality with optimal constants; see Theorem 7.3 therein.

**Theorem 2.1.** Assume for each $f \in \mathcal{F}$ that $\mathbb{E}f(X_i) = 0$ and $\sup_{x\in\mathcal{X}} |f(x)| \le B$. Let $\sigma > 0$ be such that $n\sigma^2 \ge \sum_{i=1}^n \sup_{f\in\mathcal{F}} \mathbb{E}f^2(X_i)$. Write

$$Z = \sup_{f\in\mathcal{F}} \sum_{i=1}^n f(X_i) \quad \text{and} \quad v = n\sigma^2 + 2\mathbb{E}(Z).$$

Then, for any $t \ge 0$,

$$\mathbb{P}\left\{ Z \ge \mathbb{E}(Z) + \sqrt{2vt} + \frac{Bt}{3} \right\} \le e^{-t}.$$

After learning how to control $\mathbb{E}(Z)$, we will come back to regression and classification to provide explicit error bounds on the test error for various classes $\mathcal{G}$. We will use Talagrand's inequality (or Bousquet's version of it) together with localization.

We will start with the topic of controlling the expectation in the next class.

## References


Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, Oxford.

Bousquet, O. (2003). Concentration inequalities for sub-additive functions using the entropy method. *In Stochastic Inequalities and Applications. Progress in Probability* **56** 213–247. Birkhäuser, Basel.