# MATH 281C: Mathematical Statistics

## Lecture 3

## 1 Hoeffding's Inequality and Proof of Bounded Differences Inequality

One of the goals of this lecture is to prove the bounded difference inequality. We will prove another standard concentration inequality, called Hoeffding's inequality, and then tweak the proof of Hoeffding's inequality to yield the bounded differences inequality.

**Theorem 1.1** (Hoeffding's inequality)**.** Suppose $\xi_1, \ldots, \xi_n$ are independent random variables. Suppose $a_1, \ldots, a_n, b_1, \ldots, b_n$ are constants such that $a_i \le \xi_i \le b_i$ almost surely for each $i = 1, \ldots, n$. Then, for every $t \ge 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i) \ge t\right\} \le \exp\left\{\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right\} \tag{3.1}$$

and

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

*Proof.* Let $S_n = \sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i)$, and write (for a fixed $\lambda \ge 0$)

$$\mathbb{P}(S_n \ge t) = \mathbb{P}(e^{\lambda S_n} \ge e^{\lambda t}) \le e^{-\lambda t}\mathbb{E}e^{\lambda S_n} = \exp\{-\lambda t + \psi_{S_n}(\lambda)\},$$

where $\psi_{S_n} := \log \mathbb{E}e^{\lambda S_n}$ is the log moment generating function (MGF) of $S_n$. Now by the independence of $\xi_1, \ldots, \xi_n$,

$$\psi_{S_n}(\lambda) = \log \mathbb{E}\exp\left\{\lambda \sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i)\right\} = \sum_{i=1}^{n}\log \mathbb{E}e^{\lambda(\xi_i - \mathbb{E}\xi_i)} = \sum_{i=1}^{n}\psi_{\xi - \mathbb{E}\xi_i}(\lambda),$$

where $\psi_{\xi-\mathbb{E}\xi_i}(\cdot)$ denotes the log MGF of $\xi_i - \mathbb{E}\xi_i$. Fix $1 \le i \le n$, define $U = \xi_i - \mathbb{E}\xi_i$. To bound $\psi_U(\lambda)$, note that $\mathbb{E}U = 0$ and $a_i - \mathbb{E}\xi_i \le U \le b_i - \mathbb{E}\xi_i$ almost surely. By the second order Taylor expansion of $\psi_U(\lambda)$ around 0, we have

$$\psi_U(\lambda) = \psi_U(0) + \lambda\psi_U'(0) + \frac{\lambda^2}{2}\psi_U''(\lambda')$$

for some $0 \le \lambda' \le \lambda$. Note that $\psi_U(0) = \log \mathbb{E}(1) = 0$. Also

$$\psi_U'(\lambda) = \frac{1}{\mathbb{E}e^{\lambda U}}\frac{d}{d\lambda}\mathbb{E}(e^{\lambda U}) = \frac{\mathbb{E}(Ue^{\lambda U})}{\mathbb{E}e^{\lambda U}},$$

so that $\psi'_U(0) = \mathbb{E}U = 0$. And

$$\psi''_U(\lambda) = \mathbb{E}\left(U^2 \frac{e^{\lambda U}}{\mathbb{E}e^{\lambda U}}\right) - \left(\frac{\mathbb{E}Ue^{\lambda U}}{\mathbb{E}e^{\lambda U}}\right)^2.$$

Let $V$ be a random variable whose "density" with respect to that of $U$ is $e^{\lambda U}/(\mathbb{E}e^{\lambda U})$, i.e.,

$$\mathrm{d}P_V = \frac{e^{\lambda U}}{\mathbb{E}e^{\lambda U}} \mathrm{d}P_U.$$

Let $F_U$ be the CDF of $U$. For any given $\lambda$, we can define the function $F_V$ as

$$F_V(u) = \frac{1}{\mathbb{E}e^{\lambda U}} \int_{-\infty}^{u} e^{\lambda u} \mathrm{d}F_U(u), \quad u \in \mathbb{R}.$$

It is easy to verify that $F_V$ is indeed a CDF. Then we let $V$ be a random variable whose CDF is $F_V$. Based on this construction, it can be shown that $\psi''_U(\lambda) = \mathrm{var}(V) \geq 0$. Also, because $U$ is supported on $[a_i - \mathbb{E}\xi_i, b_i - \mathbb{E}\xi_i]$ (so that its "density" vanishes outside the interval), $V$ is supported on the same interval. Consequently,

$$\psi''_U(\lambda) = \mathrm{var}(V) = \inf_{m \in \mathbb{R}} \mathbb{E}(V - m)^2 \leq \mathbb{E}(V - \eta)^2 \leq \frac{(b_i - a_i)^2}{4},$$

where $\eta$ is the mid-point of $[a_i - \mathbb{E}\xi_i, b_i - \mathbb{E}\xi_i]$. We have thus proved that $\psi''_U(\lambda) \leq (b_i - a_i)^2/4$ for every $\lambda \geq 0$. This, along with $\psi_U(0) = 0$ and $\psi'_U(0) = 0$, implies

$$\psi_U(\lambda) \leq \frac{(b_i - a_i)^2}{8} \lambda^2.$$

As a result,

$$\psi_S(\lambda) = \sum_{i=1}^{n} \psi_{\xi_i - \mathbb{E}\xi_i}(\lambda) \leq \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2,$$

and consequently,

$$\mathbb{P}(S_n \geq t) \leq \exp\left\{-\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^{n} (b_i - a_i)^2\right\}, \quad \text{for every } \lambda \geq 0.$$

We can optimize this bound over $\lambda \geq 0$ by setting

$$\lambda = \frac{4t}{\sum_{i=1}^{n} (b_i - a_i)^2}$$

to prove (3.1). To prove the lower tail inequality, simply apply (3.1) to $-\xi_1, \ldots, -\xi_n$. □

The proof given above bounds the probability $\mathbb{P}(S_n \geq t)$ in terms of the MGF of $S_n$. This technique is known as the *Cramér-Chernoff* method.

## 1.1 Remarks

Consider the following special case of Hoeffding's inequality: Suppose $X_1, \dots, X_n$ are iid with $\mathbb{E}X_i = \mu$, $\text{var}(X_i) = \sigma^2$ and $a \le X_i \le b$ almost surely ($a$ and $b$ are constants). Suppose $\bar{X}_n = (X_1 + \cdots + X_n)/n$. Hoeffding's inequality then gives

$$\mathbb{P}\{n^{1/2}(\bar{X}_n - \mu) \ge t\} \le \exp\left\{\frac{-2t^2}{(b-a)^2}\right\} \quad \text{for all } t \ge 0. \tag{3.2}$$

Is this a good/tight bound? By "good" here, we mean if the probability on the left-hand side above is close to the bound on the right or if the bound is much looser. To answer this question, we of course need a way of approximately computing the probability on the left-hand side. A natural way of doing this is via invoking the Central Limit Theorem (assuming that the CLT is valid). Indeed CLT states that

$$n^{1/2}(\bar{X}_n - \mu) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma^2) \;\; (\text{as } n \to \infty).$$

provided that the distribution of $X_i$ has finite second moment. Thus we may expect

$$\mathbb{P}\{n^{1/2}(\bar{X}_n - \mu) \ge t\} \approx \mathbb{P}\{\mathcal{N}(0, \sigma^2) \ge t\}$$

when $n$ is large and when CLT holds. What is $\mathbb{P}\{\mathcal{N}(0, \sigma^2) \ge t\}$? We can bound this again by the Cramér-Chernoff method: for every $\lambda \ge 0$,

$$\mathbb{P}\{\mathcal{N}(0, \sigma^2) \ge t\} \le \exp\{-\lambda t + \psi_Z(\lambda)\} \quad \text{with } \; Z \sim \mathcal{N}(0, \sigma^2).$$

It is known that $\mathbb{E}e^{\lambda Z} = e^{\lambda^2\sigma^2/2}$, and hence $\psi_Z(\lambda) = \lambda^2\sigma^2/2$. It follows that

$$\mathbb{P}\{\mathcal{N}(0, \sigma^2) \ge t\} \le \inf_{\lambda \ge 0} \exp\left(-\lambda t + \frac{1}{2}\lambda^2\sigma^2\right) = \exp\left(\frac{-t^2}{2\sigma^2}\right), \quad \text{for every } t \ge 0. \tag{3.3}$$

Is this bound sharp? For standard normal random variable $Z_0$, it can be shown that (exercise) for any $t > 0$,

$$\frac{t}{1+t^2}\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \le \mathbb{P}(Z_0 \ge t) \le \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

So $e^{-t^2/(2\sigma^2)}$ is the correct exponential term controlling the behavior of $\mathbb{P}\{\mathcal{N}(0, \sigma^2) \ge t\}$. Now let us compare Hoeffding's result with the bound (3.3). Hoeffding gives the bound

$$\exp\left\{\frac{-2t^2}{(b-a)^2}\right\},$$

while normal approximation suggests

$$\exp\left(\frac{-t^2}{2\sigma^2}\right).$$

Note that, because $a \le X_1 \le b$ almost surely,

$$\sigma^2 = \text{var}(X_1) \le \frac{(b-a)^2}{4}.$$

Thus in the regime where CLT holds, Hoeffding is a looser inequality where the variance $\sigma^2$ is replaced by the upper bound $(b-a)^2/4$. This looseness can be quite pronounced when $X_1$ puts less mass near the end points $a$ and $b$. Here is a potential statistical implication of this looseness.

**Example 1.1.** Suppose $X_1, \ldots, X_n$ are iid with $\mathbb{E}X_i = \mu$, $\mathrm{var}(X_i) = \sigma^2$ and $a \le X_i \le b$ almost surely. Suppose $\sigma^2$, $a$ and $b$ are known while $\mu$ is unknown, and that we seek a confidence interval for $\mu$. There are two ways of solving this problem.

The first method uses the CLT (normal approximation). Indeed, by CLT:

$$\mathbb{P}\left\{\left|\frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma}\right| \le t\right\} \to \mathbb{P}(|Z_0| \le t)$$

as $n \to \infty$ for each $t$, where $Z_0 \sim \mathcal{N}(0, 1)$. Thus

$$\mathbb{P}\left\{\left|\frac{n^{1/2}(\bar{X}_n - \mu)}{\sigma}\right| \le z_{\alpha/2}\right\} \to \mathbb{P}(|Z_0| \le a_{\alpha/2}) = 1 - \alpha,$$

where $z_{\alpha/2}$ is defined so that the last equality above holds. This leads to the following confidence interval (CI) for $\mu$:

$$\left[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right]. \tag{3.4}$$

Note that this is an "asymptotically valid" $100(1-\alpha)\%$ CI for $\mu$. Its finite sample coverage, on the other hand, may not be $100(1-\alpha)\%$.

The second method for constructing a CI for $\mu$ uses Hoeffding's inequality which states that

$$\mathbb{P}\{|n^{1/2}(\bar{X}_n - \mu)| \ge t\} \le 2\exp\left\{\frac{-2t^2}{(b-a)^2}\right\} \quad \text{for every } t \ge 0.$$

Thus, by taking

$$t = (b-a)\sqrt{\frac{\log(2/\alpha)}{2}},$$

one gets the following CI for $\mu$:

$$\left[\bar{X}_n - \frac{b-a}{\sqrt{n}}\sqrt{\frac{\log(2/\alpha)}{2}},\ \bar{X}_n + \frac{b-a}{\sqrt{n}}\sqrt{\frac{\log(2/\alpha)}{2}}\right]. \tag{3.5}$$

This inequality has guaranteed finite sample coverage $100(1-\alpha)\%$. But this interval might be much too wide compared to (3.4). Which of the two intervals (3.4) and (3.5) would you prefer?

## 1.2 Hoeffding's Inequality for Martingale Differences

**Theorem 1.2** (Hoeffding's inequality for martingale differences)**.** Suppose $\mathcal{F}_1, \ldots, \mathcal{F}_n$ are increasing $\sigma$-fields, and suppose $\xi_1, \ldots, \xi_n$ are random variables with $\xi_i$ being $\mathcal{F}_i$-measurable. Assume that

$$\mathbb{E}(\xi_i - \mathbb{E}\xi_i | \mathcal{F}_{i-1}) = 0 \quad \text{almost surely} \tag{3.6}$$

for all $i = 1, \ldots, n$. Also assume that, for each $1 \le i \le n$, the conditional distribution of $\xi_i$ given $\mathcal{F}_{i-1}$ is supported on an interval whose length is bounded from above by $R_i$ (deterministic quantity). Then, for every $t \ge 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} (\xi_i - \mathbb{E}\xi_i) \ge t\right\} \le \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} R_i^2}\right) \tag{3.7}$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i) \leq -t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} R_i^2}\right).$$

**Remark 1.1.** Assumption (3.6) means that $\{(S_j, \mathcal{F}_j)\}_{j=1}^n$ is a martingale, where $S_j = \sum_{i=1}^{j}(\xi_i - \mathbb{E}\xi_i)$. Therefore, the sequence $\{\xi_i - \mathbb{E}\xi_i\}_{i=1}^n$ is a martingale difference sequence.

*Proof.* Let $S_n = \sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i)$. As before, for every $t \geq 0$ and $\lambda \geq 0$,

$$\mathbb{P}(S_n \geq t) \leq \exp\{-\lambda t + \psi_{S_n}(\lambda)\}$$

with

$$\psi_{S_n}(\lambda) = \log \mathbb{E} e^{\lambda S_n} = \log \mathbb{E} \exp\left\{\lambda \sum_{i=1}^{n}(\xi_i - \mathbb{E}\xi_i)\right\}.$$

Observe that

$$\mathbb{E}(e^{\lambda S_n} | \mathcal{F}_{n-1}) = \exp\left\{\lambda \sum_{i=1}^{n-1}(\xi_i - \mathbb{E}\xi_i)\right\} \times \mathbb{E}\{e^{\lambda(\xi_n - \mathbb{E}\xi_n)} | \mathcal{F}_{n-1}\}.$$

Now because $\mathbb{E}\xi_n = \mathbb{E}(\xi_n | \mathcal{F}_{n-1})$, we can use exactly the same argument as in the proof of Hoeffding's inequality in the independent case (via second order Taylor expansion of the log MGF) (exercise) to deduce that

$$\mathbb{E}\{e^{\lambda(\xi_n - \mathbb{E}\xi_n)} | \mathcal{F}_{n-1}\} \leq e^{\lambda^2 R_n^2/8},$$

and this gives

$$\mathbb{E} e^{\lambda S_n} \leq e^{\lambda^2 R_n^2/8} \times \mathbb{E} \exp\left\{\lambda \sum_{i=1}^{n-1}(\xi_i - \mathbb{E}\xi_i)\right\}.$$

Now repeat the above argument (by conditioning on $\mathcal{F}_{n-2}$, then $\mathcal{F}_{n-3}$ and so on) to deduce that

$$\mathbb{E} e^{\lambda S_n} \leq \exp\left(\frac{\lambda^2}{8} \sum_{i=1}^{n} R_i^2\right).$$

This gives

$$\mathbb{P}(S_n \geq t) \leq \exp\left(-\lambda t + \frac{\lambda^2}{8} \sum_{i=1}^{n} R_i^2\right).$$

Optimize the right-hand side over $\lambda \geq 0$ gives (3.7). For the proof of the lower tail inequality, argue with $-\xi_i$ in place of $\xi_i$. ∎

## 1.3 Proof of the Bounded Differences Inequality

We now prove the bounded differences inequality as a simple consequence of Theorem 1.2.

**Theorem 1.3** (Bounded Differences Inequality)**.** Suppose $X_1, \dots, X_n$ are independent random variables taking values in a set $\mathcal{X}$. Suppose $g : \mathcal{X} \times \cdots \times \mathcal{X} \to \mathbb{R}$ is a function satisfying the following "bounded differences" assumption:

$$\left|g(x_1, \dots, x_n) - g(z_1, \dots, z_n)\right| \le \sum_{i=1}^{n} c_i I(x_i \ne z_i) \tag{3.8}$$

for some constants $c_1, \dots, c_n$. Then, for every $t \ge 0$,

$$\mathbb{P}\{g(X_1, \dots, X_n) \ge \mathbb{E}g(X_1, \dots, X_n) + t\} \le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right) \tag{3.9}$$

and

$$\mathbb{P}\{g(X_1, \dots, X_n) \le \mathbb{E}g(X_1, \dots, X_n) - t\} \le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2}\right).$$

*Proof.* We will apply the martingale Hoeffding inequality to

$$\xi_i = \mathbb{E}\{g(X_1, \dots, X_n)|X_1, \dots, X_i\} - \mathbb{E}\{g(X_1, \dots, X_n)|X_1, \dots, X_{i-1}\}, \quad i = 1, \dots, n,$$

and $\mathcal{F}_i$ taken to be the sigma field generated by $X_1, \dots, X_i$ for $i = 1, \dots, n$. Clearly, $\xi_i$, which is a function of $X_1, \dots, X_i$, is $\mathcal{F}_i$ measurable and satisfies $\mathbb{E}\xi_i = 0$. Also, check that

$$\mathbb{E}(\xi_i|\mathcal{F}_{i-1}) = 0.$$

Thus $\{(\xi_i, \mathcal{F}_i)\}$ is a martingale difference sequence. We now argue that the conditional distribution of $\xi_i$ given $\mathcal{F}_{i-1}$ is supported on an interval of length bounded from above by $c_i$. For this, let us look at the conditional distribution of $\xi_i$ given $X_1, \dots, X_{i-1}$. Fix $X_1, \dots, X_{i-1}$ at $x_1, \dots, x_{i-1}$, so that $\xi_i$ is a function solely on $X_i$ and we need to look at the range of $\xi_i$ as $X_i = x$ varies. We therefore need to look at the values

$$x \mapsto \mathbb{E}\{g(X_1, \dots, X_n)|X_1 = x_1, \dots, X_{i-1} = x_{i-1}, X_i = x\} - \mathbb{E}\{g(X_1, \dots, X_n)|X_1 = x_1, \dots, X_{i-1} = x_{i-1}\}$$

as $x$ varies and $x_1, \dots, x_{i-1}$ are fixed. Now, by independence of $X_1, \dots, X_n$, the right-hand side above equals

$$\mathbb{E}\{g(x_1, \dots, x_{i-1}, x, X_{i+1}, \dots, X_n)\} - \text{constant},$$

where the "constant" term depends on $x_1, \dots, x_{i-1}$. Thus we can take $R_i$ to be

$$\begin{aligned} R_i &:= \sup_{x,x' \in \mathcal{X}} \left|\mathbb{E}g(x_1, \dots, x_{i-1}, x, X_{i+1}, \dots, X_n) - \mathbb{E}g(x_1, \dots, x_{i-1}, x', X_{i+1}, \dots, X_n)\right| \\ &\le \sup_{x,x' \in \mathcal{X}} \mathbb{E}\left|g(x_1, \dots, x_{i-1}, x, X_{i+1}, \dots, X_n) - g(x_1, \dots, x_{i-1}, x', X_{i+1}, \dots, X_n)\right|. \end{aligned}$$

It is clear now that $R_i \le c_i$ by the bounded differences assumption (1.1). We can therefore apply Theorem 1.2 with $R_i = c_i$, which finishes the proof of Theorem 1.3. □

# References


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