

# Lecture 2: Concentration without Independence

Tianpei Xie

Jan. 6th., 2023

## Contents

<b>1</b>	<b>Martingale-based Methods</b>	<b>2</b>
1.1	Martingale . . . . .	2
1.2	Bernstein Inequality for Martingale Difference Sequence . . . . .	4
1.3	Azuma-Hoeffding Inequality . . . . .	5
1.4	McDiarmid's Inequality . . . . .	5
1.5	Applications . . . . .	7
<b>2</b>	<b>Bounding Variance</b>	<b>9</b>
2.1	The Efron-Stein Inequality . . . . .	9
2.2	Functions with Bounded Differences . . . . .	12
2.3	Self-Bounding Functions . . . . .	13
2.4	Applications . . . . .	14
2.4.1	Kernel Density Estimation . . . . .	14
2.4.2	Convex Poincaré Inequality . . . . .	15
2.4.3	Gaussian Poincaré Inequality . . . . .	16
2.5	A Proof of the Efron-Stein Inequality Based on Duality . . . . .	17
2.6	Exponential Tail Bounds via the Efron-Stein Inequality . . . . .	19

# 1 Martingale-based Methods

## 1.1 Martingale

- **Definition (Martingale)** [Resnick, 2013]

Let  $\{X_n, n \geq 0\}$  be a stochastic process on  $(\Omega, \mathcal{F})$  and  $\{\mathcal{F}_n, n \geq 0\}$  be a **filtration**; that is,  $\{\mathcal{F}_n, n \geq 0\}$  is an *increasing sub  $\sigma$ -fields* of  $\mathcal{F}$

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}.$$

Then  $\{(X_n, \mathcal{F}_n), n \geq 0\}$  is a **martingale (mg)** if

1.  $X_n$  is **adapted** in the sense that for each  $n$ ,  $X_n \in \mathcal{F}_n$ ; that is,  $X_n$  is  $\mathcal{F}_n$ -measurable.
2.  $X_n \in L_1$ ; that is  $\mathbb{E}[|X_n|] < \infty$  for  $n \geq 0$ .
3. For  $0 \leq m < n$

$$\mathbb{E}[X_n | \mathcal{F}_m] = X_m, \quad \text{a.s.} \quad (1)$$

If the equality of (1) is replaced by  $\geq$ ; that is, things are getting better on the average:

$$\mathbb{E}[X_n | \mathcal{F}_m] \geq X_m, \quad \text{a.s.} \quad (2)$$

then  $\{X_n\}$  is called a **sub-martingale (submg)** while if things are getting worse on the average

$$\mathbb{E}[X_n | \mathcal{F}_m] \leq X_m, \quad \text{a.s.} \quad (3)$$

$\{X_n\}$  is called a **super-martingale (supermg)**.

- **Remark**  $\{X_n\}$  is **martingale** if it is *both* a **sub** and **supermartingale**.  $\{X_n\}$  is a **super-martingale** if and only if  $\{-X_n\}$  is a **submartingale**.
- **Remark** If  $\{X_n\}$  is a **martingale**, then  $\mathbb{E}[X_n]$  is *constant*. In the case of a **submartingale**, the mean increases and for a **supermartingale**, the mean decreases.
- **Proposition 1.1** [Resnick, 2013]  
If  $\{(X_n, \mathcal{F}_n), n \geq 0\}$  is a **(sub, super) martingale**, then

$$\{(X_n, \sigma(X_0, X_1, \dots, X_n)), n \geq 0\}$$

is also a **(sub, super) martingale**.

- **Definition (Martingale Differences)**. [Resnick, 2013]  
 $\{(d_j, \mathcal{B}_j), j \geq 0\}$  is a **(sub, super) martingale difference sequence** or a **(sub, super) fair sequence** if

1. For  $j \geq 0$ ,  $\mathcal{B}_j \subset \mathcal{B}_{j+1}$ .
2. For  $j \geq 0$ ,  $d_j \in L_1$ ,  $d_j \in \mathcal{B}_j$ ; that is,  $d_j$  is *absolutely integrable* and  $\mathcal{B}_j$ -measurable.
3. For  $j \geq 0$ ,

$$\begin{aligned} \mathbb{E}[d_{j+1} | \mathcal{B}_j] &= 0, & (\text{martingale difference / fair sequence}); \\ &\geq 0, & (\text{submartingale difference / subfair sequence}); \\ &\leq 0, & (\text{supermartingale difference / supfair sequence}) \end{aligned}$$

- **Proposition 1.2** (*Construction of Martingale From Martingale Difference*) [Resnick, 2013]  
If  $\{(d_j, \mathcal{B}_j), j \geq 0\}$  is *(sub, super) martingale difference sequence*, and

$$X_n = \sum_{j=0}^n d_j,$$

then  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a *(sub, super) martingale*.

- **Proposition 1.3** (*Construction of Martingale Difference From Martingale*) [Resnick, 2013]  
Suppose  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a *(sub, super) martingale*. Define

$$\begin{aligned} d_0 &:= X_0 - \mathbb{E}[X_0] \\ d_j &:= X_j - X_{j-1}, \quad j \geq 1. \end{aligned}$$

Then  $\{(d_j, \mathcal{B}_j), j \geq 0\}$  is a *(sub, super) martingale difference sequence*.

- **Proposition 1.4** (*Orthogonality of Martingale Differences*). [Resnick, 2013]  
If  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a *martingale* where  $X_n$  can be decomposed as

$$X_n = \sum_{j=0}^n d_j,$$

$d_j$  is  $\mathcal{B}_j$ -measurable and  $\mathbb{E}[d_j^2] < \infty$  for  $j \geq 0$ , then  $\{d_j\}$  are *orthogonal*:

$$\mathbb{E}[d_i d_j] = 0 \quad i \neq j.$$

**Proof:** This is an easy verification: If  $j > i$ , then

$$\begin{aligned} \mathbb{E}[d_i d_j] &= \mathbb{E}[\mathbb{E}[d_i d_j | \mathcal{B}_i]] \\ &= \mathbb{E}[d_i \mathbb{E}[d_j | \mathcal{B}_i]] = 0. \quad \blacksquare \end{aligned}$$

A consequence is that

$$\mathbb{E}[X_n^2] = \mathbb{E}\left[\sum_{i=1}^n d_i^2\right] + 2 \sum_{0 \leq i < j \leq n} \mathbb{E}[d_i d_j] = \mathbb{E}\left[\sum_{i=1}^n d_i^2\right],$$

which is *non-decreasing*. From this, it seems likely (and turns out to be true) that  $\{X_n^2\}$  is a *sub-martingale*.

- **Example** (*Smoothing as Martingale*)  
Suppose  $X \in L_1$  and  $\{\mathcal{B}_n, n \geq 0\}$  is an increasing family of sub  $\sigma$ -algebra of  $\mathcal{B}$ . Define for  $n \geq 0$

$$X_n := \mathbb{E}[X | \mathcal{B}_n].$$

Then  $(X_n, \mathcal{B}_n)$  is a *martingale*. From this result, we see that  $\{(d_n, \mathcal{B}_n), n \geq 0\}$  is a *martingale difference sequence* when

$$d_n := \mathbb{E}[X | \mathcal{B}_n] - \mathbb{E}[X | \mathcal{B}_{n-1}], \quad n \geq 1. \quad (4)$$

**Proof:** See that

$$\begin{aligned}\mathbb{E}[X_{n+1}|\mathcal{B}_n] &= \mathbb{E}[\mathbb{E}[X|\mathcal{B}_{n+1}]|\mathcal{B}_n] \\ &= \mathbb{E}[X|\mathcal{B}_n] \quad (\text{Smoothing property of conditional expectation}) \\ &= X_n \quad \blacksquare\end{aligned}$$

- **Example (*Sums of Independent Random Variables*)**

Suppose that  $\{Z_n, n \geq 0\}$  is an *independent* sequence of integrable random variables satisfying for  $n \geq 0$ ,  $\mathbb{E}[Z_n] = 0$ . Set

$$\begin{aligned}X_0 &:= 0, \\ X_n &:= \sum_{i=1}^n Z_i, \quad n \geq 1 \\ \mathcal{B}_n &:= \sigma(Z_0, \dots, Z_n).\end{aligned}$$

Then  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a *martingale* since  $\{(Z_n, \mathcal{B}_n), n \geq 0\}$  is a *martingale difference sequence*.

- **Example (*Likelihood Ratios*).**

Suppose  $\{Y_n, n \geq 0\}$  are *independent identically distributed* random variables and suppose the *true density* of  $Y_n$  is  $f_0$ . (The word “density” can be understood with respect to some fixed reference measure  $\mu$ .) Let  $f_1$  be some other probability density. For simplicity suppose  $f_0(y) > 0$ , for all  $y$ . For  $n \geq 0$ , define the likelihood ratio

$$\begin{aligned}X_n &:= \frac{\prod_{i=0}^n f_1(Y_i)}{\prod_{i=0}^n f_0(Y_i)} \\ \mathcal{B}_n &:= \sigma(Y_0, \dots, Y_n)\end{aligned}$$

Then  $(X_n, \mathcal{B}_n)$  is a *martingale*.

**Proof:** See that

$$\begin{aligned}\mathbb{E}[X_{n+1}|\mathcal{B}_n] &= \mathbb{E}\left[\left(\frac{\prod_{i=0}^n f_1(Y_i)}{\prod_{i=0}^n f_0(Y_i)}\right) \frac{f_1(Y_{n+1})}{f_0(Y_{n+1})} \mid Y_0, \dots, Y_n\right] \\ &= X_n \mathbb{E}\left[\frac{f_1(Y_{n+1})}{f_0(Y_{n+1})} \mid Y_0, \dots, Y_n\right] \\ &= X_n \mathbb{E}\left[\frac{f_1(Y_{n+1})}{f_0(Y_{n+1})}\right] \quad (\text{by independence}) \\ &:= X_n \int \frac{f_1(y_{n+1})}{f_0(y_{n+1})} f_0(y_{n+1}) d\mu(y_{n+1}) = X_n. \quad \blacksquare\end{aligned}$$

## 1.2 Bernstein Inequality for Martingale Difference Sequence

- **Proposition 1.5 (*Bernstein Inequality, Martingale Difference Sequence Version*)**

[Wainwright, 2019]

Let  $\{(D_k, \mathcal{B}_k), k \geq 1\}$  be a *martingale difference sequence*, and suppose that

$$\mathbb{E}[\exp(\lambda D_k) | \mathcal{B}_{k-1}] \leq \exp\left(\frac{\lambda^2 \nu_k^2}{2}\right)$$

almost surely for any  $|\lambda| < 1/\alpha_k$ . Then the following hold:

1. The sum  $\sum_{k=1}^n D_k$  is **sub-exponential** with **parameters**  $(\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha_*)$  where  $\alpha_* := \max_{k=1, \dots, n} \alpha_k$ . That is, for any  $|\lambda| < 1/\alpha_*$ ,

$$\mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^n D_k \right) \right\} \right] \leq \exp \left( \frac{\lambda^2 \sum_{k=1}^n \nu_k^2}{2} \right)$$

2. The sum satisfies **the concentration inequality**

$$\mathbb{P} \left\{ \left| \sum_{k=1}^n D_k \right| \geq t \right\} \leq \begin{cases} 2 \exp \left( -\frac{t^2}{2 \sum_{k=1}^n \nu_k^2} \right) & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*} \\ 2 \exp \left( -\frac{t}{\alpha_*} \right) & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*}. \end{cases} \quad (5)$$

**Proof:** We follow the standard approach of controlling the moment generating function of  $\sum_{k=1}^n D_k$ , and then applying *the Chernoff bound*. For any scalar  $\lambda$  such that  $|\lambda| < 1/\alpha_*$ , conditioning on  $\mathcal{B}_{n-1}$  and applying iterated expectation yields

$$\begin{aligned} \mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^n D_k \right) \right\} \right] &= \mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^{n-1} D_k \right) \right\} \mathbb{E} \left[ \exp \{ \lambda D_n \} \mid \mathcal{B}_{n-1} \right] \right] \\ &\leq \mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^{n-1} D_k \right) \right\} \right] \exp \left( \frac{\lambda^2 \nu_n^2}{2} \right), \end{aligned}$$

where the inequality follows from the stated assumption on  $D_n$ . Iterating this procedure yields the bound  $\mathbb{E} [\exp \{ \lambda (\sum_{k=1}^n D_k) \}] \leq \exp \left( \frac{\lambda^2 \sum_{k=1}^n \nu_k^2}{2} \right)$ , valid for all  $|\lambda| < 1/\alpha_*$ . By definition, we conclude that  $\sum_{k=1}^n D_k$  is *sub-exponential* with *parameters*  $(\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha_*)$ , as claimed. The tail bound (5) follows by properties of sub-exponential distribution.  $\blacksquare$

- **Remark** This result is a **generalization** of the *Bernstein's inequality* when  $\{D_k\}$  are **independent sub-exponential distributed** random variables.

The proof used the property of conditional expectation

$$\mathbb{E} [\mathbb{E} [X | \mathcal{B}_n]] = \mathbb{E} [X], \quad \mathbb{E} [h(X)g(Y) | Y] \stackrel{a.s.}{=} h(X) \mathbb{E} [g(Y) | Y]$$

### 1.3 Azuma-Hoeffding Inequality

- **Corollary 1.6 (Azuma-Hoeffding Inequality, Martingale Difference)** [Wainwright, 2019] Let  $\{(D_k, \mathcal{B}_k), k \geq 1\}$  be a **martingale difference sequence** for which there are constants  $\{(a_k, b_k)\}_{k=1}^n$  such that  $D_k \in [a_k, b_k]$  almost surely for all  $k = 1, \dots, n$ . Then, for all  $t \geq 0$ ,

$$\mathbb{P} \left\{ \left| \sum_{k=1}^n D_k \right| \geq t \right\} \leq 2 \exp \left( -\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2} \right) \quad (6)$$

### 1.4 McDiarmid's Inequality

- An important application of *Azuma-Hoeffding Inequality* concerns functions that satisfy a *bounded difference property*.

**Definition (*Functions with Bounded Difference Property*)**

Given vectors  $x, x' \in \mathcal{X}^n$  and an index  $k \in \{1, 2, \dots, n\}$ , we define a new vector  $x^{(-k)} \in \mathcal{X}^n$  via

$$x_j^{(-k)} = \begin{cases} x_j & j \neq k \\ x'_k & j = k \end{cases}$$

With this notation, we say that  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  satisfies **the bounded difference inequality** with parameters  $(L_1, \dots, L_n)$  if, for each index  $k = 1, 2, \dots, n$ ,

$$\left| f(x) - f(x^{(-k)}) \right| \leq L_k, \quad \text{for all } x, x' \in \mathcal{X}^n. \quad (7)$$

- **Corollary 1.7 (*McDiarmid's Inequality / Bounded Differences Inequality*)** [Wainwright, 2019]

Suppose that  $f$  satisfies **the bounded difference property** (7) with parameters  $(L_1, \dots, L_n)$  and that the random vector  $X = (X_1, X_2, \dots, X_n)$  has **independent** components. Then

$$\mathbb{P} \{ |f(X) - \mathbb{E}[f(X)]| \geq t \} \leq 2 \exp \left( - \frac{2t^2}{\sum_{k=1}^n L_k^2} \right). \quad (8)$$

**Proof:** Consider the associated *martingale difference sequence*

$$D_k := \mathbb{E}[f(X) | X_1, \dots, X_k] - \mathbb{E}[f(X) | X_1, \dots, X_{k-1}].$$

We claim that  $D_k$  lies in an interval of length at most  $L_k$  almost surely. In order to prove this claim, define the random variables

$$\begin{aligned} A_k &:= \inf_x \{ \mathbb{E}[f(X) | X_1, \dots, X_{k-1}, x] \} - \mathbb{E}[f(X) | X_1, \dots, X_{k-1}] \\ B_k &:= \sup_x \{ \mathbb{E}[f(X) | X_1, \dots, X_{k-1}, x] \} - \mathbb{E}[f(X) | X_1, \dots, X_{k-1}]. \end{aligned}$$

On one hand, we have

$$D_k - A_k = \mathbb{E}[f(X) | X_1, \dots, X_k] - \inf_x \{ \mathbb{E}[f(X) | X_1, \dots, X_{k-1}, x] \},$$

so that  $D_k \geq A_k$  almost surely. A similar argument shows that  $D_k \leq B_k$  almost surely. We now need to show that  $B_k - A_k \leq L_k$  almost surely. Observe that by the independence of  $\{X_k\}_{k=1}^n$ , we have

$$\mathbb{E}[f(X) | x_1, \dots, x_k] = \mathbb{E}_{(k+1)}[f(x_1, \dots, x_k, X_{k+1}, \dots, X_n)], \text{ for any } (x_1, \dots, x_k),$$

where  $\mathbb{E}_{(k+1)}[\cdot]$  denote the expectation over  $(X_{k+1}, \dots, X_n)$ . Consequently, we have

$$\begin{aligned} B_k - A_k &= \sup_x \mathbb{E}_{(k+1)}[f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)] \\ &\quad - \inf_x \mathbb{E}_{(k+1)}[f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)] \\ &\leq \sup_{x, y} \{ \mathbb{E}_{(k+1)}[f(X_{1:k-1}, x, X_{k+1:n})] - \mathbb{E}_{(k+1)}[f(X_{1:k-1}, y, X_{k+1:n})] \} \\ &\leq L_k, \end{aligned}$$

using *the bounded differences assumption*. Thus, the variable  $D_k$  lies within an interval of length  $L_k$  at most surely, so that the claim follows as a corollary of *the Azuma-Hoeffding inequality*. ■

## 1.5 Applications

- **Example (*U-Statistics*)** [Wainwright, 2019]

Let  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  be a **symmetric function** of its arguments. Given an i.i.d. sequence  $\{X_k, k \geq 1\}$ , of random variables, the quantity

$$U := \frac{1}{\binom{n}{2}} \sum_{j < k} g(X_j, X_k) \quad (9)$$

is known as a **pairwise U-statistic**. For instance, if  $g(s, t) = |s - t|$ , then  $U$  is an *unbiased estimator of the mean absolute pairwise deviation*  $\mathbb{E}[|X_1 - X_2|]$ . Note that, while  $U$  is **not a sum of independent random variables**, the *dependence is relatively weak*, and this fact can be revealed by a martingale analysis.

If  $g$  is bounded (say  $\|g\|_\infty \leq b$ ), then the *Bounded Difference Inequality* can be used to establish the concentration of  $U$  around its mean. Viewing  $U$  as a function  $f(x) = f(x_1, \dots, x_n)$ , for any given coordinate  $k$ , we have

$$\begin{aligned} \left| f(x) - f(x^{(-k)}) \right| &\leq \frac{1}{\binom{n}{2}} \sum_{j \neq k} |g(x_j, x_k) - g(x_j, x'_k)| \\ &\leq \frac{(n-1)2b}{\binom{n}{2}} = \frac{4b}{n}, \end{aligned}$$

so that the bounded differences property holds with parameter  $L_k = \frac{4b}{n}$  in each coordinate. Thus, we conclude that

$$\mathbb{P}\{|U - \mathbb{E}[U]| \geq t\} \leq 2 \exp\left(-\frac{nt^2}{8b^2}\right),$$

This tail inequality implies that  $U$  is a consistent estimate of  $\mathbb{E}[U]$ , and also yields *finite sample bounds* on its quality as an estimator. Similar techniques can be used to obtain *tail bounds on U-statistics of higher order*, involving sums over  $k$ -tuples of variables. ■

- **Example (*Clique Number in Erdős-Rényi Random Graphs*)** [Wainwright, 2019]

Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  be an *undirected graph*, where  $\mathcal{V} = \{1, \dots, d\}$  is the vertex set and  $\mathcal{E} = \{(i, j), i, j \in \mathcal{V}\}$  is the undirected edge set. A **graph clique**  $C$  is a subset of vertices such that  $(i, j) \in \mathcal{E}$  for all  $i, j \in C$ . **The clique number**  $C(\mathcal{G})$  of the graph is **the cardinality of the largest clique**. Note that  $C(\mathcal{G}) \in [1, d]$ . When the edges  $\mathcal{E}$  of the graph are drawn according to some random process, then the *clique number*  $C(\mathcal{G})$  is a *random variable*, and we can study its concentration around its mean  $\mathbb{E}[C(\mathcal{G})]$ .

**The Erdős-Rényi ensemble of random graphs** is one of the most well-studied models: it is defined by a parameter  $p \in (0, 1)$  that specifies the probability with which each edge  $(i, j)$  is *included* in the graph, **independently across all**  $\binom{d}{2}$  edges. More formally, for each  $i < j$ , let us introduce a **Bernoulli edge-indicator variable**  $X_{i,j}$  with parameter  $p$ , where  $X_{i,j} = 1$  means that edge  $(i, j)$  is *included* in the graph, and  $X_{i,j} = 0$  means that it is *not included*.

Note that the  $\binom{d}{2}$ -dimensional random vector  $Z := \{X_{i,j}\}_{i < j}$  specifies the edge set; thus, we may view the *clique number*  $C(\mathcal{G})$  as a function  $Z \rightarrow f(Z)$ . Based on definition in Section 2.3, we see that  $f(Z)$  is a **configuration function** with property of “being in a clique”.

Let  $Z'$  denote a vector in which a *single coordinate* of  $Z$  has been changed, and let  $\mathcal{G}'$  and  $\mathcal{G}$  be the associated graphs. It is easy to see that  $C(\mathcal{G}')$  can differ from  $C(\mathcal{G})$  by at most 1, so that

$$|f(Z) - f(Z')| \leq 1,$$

Thus, the function  $C(\mathcal{G}) = f(Z)$  satisfies *the bounded difference property* in each coordinate with parameter  $L = 1$ , so that

$$\mathbb{P} \left\{ \frac{1}{n} |C(\mathcal{G}) - \mathbb{E}[C(\mathcal{G})]| \geq \delta \right\} \leq 2 \exp(-2n\delta^2).$$

Consequently, we see that the clique number of an *Erdős-Rényi random graph* is *very sharply concentrated around its expectation*. ■

• **Example (*Rademacher Complexity*)** [Wainwright, 2019]

Let  $\{\epsilon_k\}_{k=1}^n$  be an i.i.d. sequence of *Rademacher variables* (i.e., taking the values  $\{-1, +1\}$  *equiprobably*). Given a collection of vectors  $\mathcal{A} \subset \mathbb{R}^n$ , define the random variable

$$Z := \sup_{a \in \mathcal{A}} \sum_{k=1}^n \epsilon_k a_k = \sup_{a \in \mathcal{A}} \langle a, \epsilon \rangle. \quad (10)$$

The random variable  $Z$  measures the **size** of  $\mathcal{A}$  in a certain sense, and its expectation

$$\mathfrak{R}(\mathcal{A}) := \mathbb{E}[Z(\mathcal{A})] \quad (11)$$

is known as **the Rademacher complexity** of the set  $\mathcal{A}$ .

Let us now show how the bounded difference inequality can be used to establish that  $Z(\mathcal{A})$  is **sub-Gaussian**. Viewing  $Z(\mathcal{A})$  as a function  $(\epsilon_1, \dots, \epsilon_n) \rightarrow f(\epsilon_1, \dots, \epsilon_n)$ , we need to *bound the maximum change* when coordinate  $k$  is changed. Given two Rademacher vectors  $\epsilon, \epsilon' \in \{-1, +1\}^n$ , recall our definition of the modified vector  $\epsilon^{(-k)}$ . Since

$$f(\epsilon^{(-k)}) \geq \langle a, \epsilon^{(-k)} \rangle, \quad \text{for any } a \in \mathcal{A},$$

we have

$$\langle a, \epsilon \rangle - f(\epsilon^{(-k)}) \leq \langle a, \epsilon - \epsilon^{(-k)} \rangle = a_k(\epsilon_k - \epsilon'_k) \leq 2|a_k|.$$

Taking *the supremum over  $\mathcal{A}$  on both sides*, we obtain the inequality

$$f(\epsilon) - f(\epsilon^{(-k)}) \leq 2 \sup_{a \in \mathcal{A}} |a_k|.$$

Since the same argument applies with the roles of  $\epsilon$  and  $\epsilon^{(-k)}$  *reversed*, we conclude that  $f$  satisfies *the bounded difference inequality* in coordinate  $k$  with parameter  $L_k := 2 \sup_{a \in \mathcal{A}} |a_k|$ .

Consequently, the bounded difference inequality implies that the random variable  $Z(\mathcal{A})$  is *sub-Gaussian* with parameter at most  $2 \sqrt{\sum_{k=1}^n \sup_{a \in \mathcal{A}} a_k^2}$ . This sub-Gaussian parameter can be reduced to the (potentially much) smaller quantity  $\sqrt{\sup_{a \in \mathcal{A}} \sum_{k=1}^n a_k^2}$  using alternative techniques. ■



## 2 Bounding Variance

### 2.1 The Efron-Stein Inequality

- **Remark** Let  $X$  be a random variable with finite variance  $\text{Var}(X)$ . By *Chebyshev's Inequality*, for any  $t > 0$ , we have

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq \frac{\text{Var}(X)}{t^2}.$$

Thus, we can obtain the tail probability by bounding the variance  $\text{Var}(X)$ .

- **Remark (Variance of Independence Random Variables)**

Let  $X_n = \sum_{i=1}^n Z_i$  be the sum of *independent* real-valued random variables  $Z_1, \dots, Z_n$ . Then we have

$$\begin{aligned} \mathbb{E}\left[(X_n - \mathbb{E}[X_n])^2\right] &= \sum_{i=1}^n \mathbb{E}\left[(Z_i - \mathbb{E}[Z_i])^2\right] \\ \Rightarrow \text{Var}(X_n) &= \sum_{i=1}^n \text{Var}(Z_i). \end{aligned}$$

- **Remark (Variance of Smoothing Martingale Difference Sequence)**

Suppose  $X \in L_1$  and  $\{\mathcal{B}_n, n \geq 0\}$  is an increasing family of sub  $\sigma$ -algebra of  $\mathcal{B}$  formed by

$$\mathcal{B}_n := \sigma(Z_1, \dots, Z_n).$$

For  $n \geq 1$ , define

$$\begin{aligned} d_0 &:= \mathbb{E}[X] \\ d_n &:= \mathbb{E}[X|\mathcal{B}_n] - \mathbb{E}[X|\mathcal{B}_{n-1}] \\ &= \mathbb{E}[X|Z_1, \dots, Z_n] - \mathbb{E}[X|Z_1, \dots, Z_{n-1}]. \end{aligned}$$

From (4) we see that  $(d_n, \mathcal{B}_n)$  is a martingale difference sequence. By *orthogonality of martingale difference*, we see that

$$\mathbb{E}[d_i d_j] = 0 \quad i \neq j.$$

Therefore, based on the decomposition

$$X - EX = \sum_{i=1}^n d_i$$

we have

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}\left[\left(\sum_{i=1}^n d_i\right)^2\right] = \sum_{i=1}^n \mathbb{E}[d_i^2] + 2 \sum_{i>j} \mathbb{E}[d_i d_j] \\ &= \sum_{i=1}^n \mathbb{E}[d_i^2]. \end{aligned} \tag{12}$$

- **Remark (Variance of General Functions of Independent Random Variables)**

Then above formula (12) holds when  $X = f(Z_1, \dots, Z_n)$  for general function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  with  $n$  independent random variables  $(Z_1, \dots, Z_n)$ . By *Fubini's theorem*,

$$\mathbb{E}[X|Z_1, \dots, Z_i] = \int_{\mathcal{Z}^{n-i}} f(Z_1, \dots, Z_i, z_{i+1}, \dots, z_n) d\mu_{i+1}(z_{i+1}) \dots d\mu_n(z_n)$$

where  $\mu_j$  is the probability distribution of  $Z_j$  for  $j \geq 1$ .

Let  $Z_{(-i)} := (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$  be all random variables  $(Z_1, \dots, Z_n)$  **except for**  $Z_i$ . Denote  $\mathbb{E}_{(-i)}[\cdot]$  as the conditional expectation of  $X$  given  $Z_{(-i)}$

$$\begin{aligned} \mathbb{E}_{(-i)}[X] &:= \mathbb{E}[X|Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n] \\ &= \int_{\mathcal{Z}} f(Z_1, \dots, Z_{i-1}, z_i, Z_{i+1}, \dots, Z_n) d\mu_i(z_i). \end{aligned}$$

Then, again by *Fubini's theorem (smoothing properties of conditional expectation)*,

$$\mathbb{E}[\mathbb{E}_{(-i)}[X]|Z_1, \dots, Z_i] = \mathbb{E}[X|Z_1, \dots, Z_{i-1}] \quad (13)$$

- **Proposition 2.1 (Efron-Stein Inequality)** [Boucheron et al., 2013]

Let  $Z_1, \dots, Z_n$  be **independent random variables** and let  $X = f(Z)$  be a square-integrable function of  $Z = (Z_1, \dots, Z_n)$ . Then

$$\text{Var}(X) \leq \sum_{i=1}^n \mathbb{E}[(X - \mathbb{E}_{(-i)}[X])^2] := \nu. \quad (14)$$

Moreover, if  $Z'_1, \dots, Z'_n$  are **independent** copies of  $Z_1, \dots, Z_n$  and if we define, for every  $i = 1, \dots, n$ ,

$$X'_i := f(Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_n),$$

then

$$\nu = \frac{1}{2} \sum_{i=1}^n \mathbb{E}[(X - X'_i)^2] = \sum_{i=1}^n \mathbb{E}[(X - X'_i)_+^2] = \sum_{i=1}^n \mathbb{E}[(X - X'_i)_-^2]$$

where  $x_+ = \max\{x, 0\}$  and  $x_- = \max\{-x, 0\}$  denote the **positive** and **negative** parts of a real number  $x$ . Also,

$$\nu = \inf_{X_i} \sum_{i=1}^n \mathbb{E}[(X - X_i)^2],$$

where the infimum is taken over the class of all  $Z_{(-i)}$ -measurable and square-integrable variables  $X_i$ ,  $i = 1, \dots, n$ .

**Proof:** We begin with the proof of the first statement. Note that, using (13), we may write

$$\begin{aligned} d_i &:= \mathbb{E}[X|Z_1, \dots, Z_i] - \mathbb{E}[X|Z_1, \dots, Z_{i-1}] \\ &= \mathbb{E}[X|Z_1, \dots, Z_i] - \mathbb{E}[\mathbb{E}_{(-i)}[X]|Z_1, \dots, Z_i] \\ &= \mathbb{E}[X - \mathbb{E}_{(-i)}[X]|Z_1, \dots, Z_i]. \end{aligned}$$

By *Jensen's inequality* used conditionally,

$$d_i^2 \leq \mathbb{E} \left[ (X - \mathbb{E}_{(-i)}[X])^2 \mid Z_1, \dots, Z_i \right]$$

Using (12)  $\text{Var}(X) = \sum_{i=1}^n \mathbb{E} [d_i^2]$ , we have

$$\text{Var}(X) \leq \sum_{i=1}^n \mathbb{E} \left[ \mathbb{E} \left[ (X - \mathbb{E}_{(-i)}[X])^2 \mid Z_1, \dots, Z_i \right] \right] = \sum_{i=1}^n \mathbb{E} \left[ (X - \mathbb{E}_{(-i)}[X])^2 \right],$$

we obtain the desired inequality.

To prove the identities for  $\nu$ , denote by  $\text{Var}_{(-i)}$  the *conditional variance operator* conditioned on  $Z_{(-i)} := (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$ . Then we may write  $\nu$  as

$$\nu = \sum_{i=1}^n \mathbb{E} [\text{Var}_{(-i)}(X)].$$

Now note that one may simply use (conditionally) the elementary fact that if  $X$  and  $Y$  are *independent and identically distributed* real-valued random variables, then

$$\text{Var}(X) = \frac{1}{2} \mathbb{E} [(X - Y)^2].$$

Since conditionally on  $Z_{(-i)}$ ,  $X'_i$  is an independent copy of  $X$ , we may write

$$\text{Var}_{(i)}(X) = \frac{1}{2} \mathbb{E}_{(-i)} [(X - X'_i)^2] = \sum_{i=1}^n \mathbb{E}_{(-i)} [(X - X'_i)^2_+] = \sum_{i=1}^n \mathbb{E}_{(-i)} [(X - X'_i)^2_-],$$

where we used the fact that the conditional distributions of  $X$  and  $X'_i$  are *identical*.

The last identity is obtained by recalling that, for any real-valued random variable  $X$ ,

$$\text{Var}(X) = \inf_{a \in \mathbb{R}} \mathbb{E} [(X - a)^2].$$

Using this fact conditionally, we have, for every  $i = 1, \dots, n$ ,

$$\text{Var}_{(-i)}(X) = \inf_{X_i} \mathbb{E}_{(-i)} [(X - X_i)^2].$$

Note that this infimum is achieved whenever  $X_i = \mathbb{E}_{(-i)}[X]$ . ■

- **Example (*The Jackknife Estimate*)**

We should note here that the Efron-Stein inequality was first motivated by the study of the so-called *jackknife estimate of statistics*.

To describe this estimate, assume that  $Z_1, \dots, Z_n$  are i.i.d. random variables and one wishes to *estimate a functional  $\theta$  of the distribution* of the  $Z_i$  by a function  $X = f(Z_1, \dots, Z_n)$  of the data. The quality of the estimate is often measured by its bias  $\mathbb{E}[X] - \theta$  and its variance  $\text{Var}(X)$ . Since the distribution of the  $Z_i$ 's is unknown, one needs to *estimate* the bias and variance ***from the same sample***. *The jackknife estimate of the bias* is defined by

$$(n-1) \left( \frac{1}{n} \sum_{i=1}^n X_i - X \right) \tag{15}$$

where  $X_i$  is an appropriately defined function of  $Z_{(-i)} := (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$ .  $Z_{(-i)}$  is often called *the  $i$ -th jackknife sample* while  $X_i$  is the so-called *jackknife replication* of  $X$ . In an analogous way, the jackknife estimate of the variance is defined by

$$\sum_{i=1}^n (X - X_i)^2 \quad (16)$$

Using this language, *the Efron-Stein inequality* simply states that *the jackknife estimate of the variance is always positively biased*. In fact, this is how Efron and Stein originally formulated their inequality.

- **Remark** Observe that in the case when  $X = \sum_{i=1}^n Z_i$  is a sum of *independent* random variables (with *finite variance*), then *the Efron-Stein inequality* becomes an *equality*. Thus, *the bound in the Efron-Stein inequality is, in a sense, not improvable*.

## 2.2 Functions with Bounded Differences

- **Remark** Recall that a function  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  satisfies *the bounded difference inequality* with parameters  $(L_1, \dots, L_n)$  if, for each index  $k = 1, 2, \dots, n$ ,

$$\left| f(z) - f(z^{(-k)}) \right| \leq L_k, \quad \text{for all } z, z' \in \mathcal{X}^n.$$

where

$$z_j^{(-k)} = \begin{cases} z_j & j \neq k \\ z'_k & j = k \end{cases}$$

- **Corollary 2.2** [Boucheron et al., 2013]

If  $f$  has the *bounded differences property* with parameters  $(L_1, \dots, L_n)$ , then

$$\text{Var}(f(Z)) \leq \frac{1}{4} \sum_{i=1}^n L_i^2.$$

**Proof:** From the Efron-Stein inequality,

$$\begin{aligned} \text{Var}(f(Z)) &\leq \sum_{i=1}^n \mathbb{E} [\text{Var}_{(-i)}(f(Z))] \\ &= \inf_{X_i} \sum_{i=1}^n \mathbb{E}_{(-i)} [(f(Z) - X_i)^2] \end{aligned}$$

where the infimum is taken over the class of all  $Z_{(-i)}$ -measurable and square-integrable variables  $X_i$ . Here we choose

$$X_i = \frac{1}{2} \left( \sup_{z'_i} f(Z_{1:i-1}, z'_i, Z_{i+1:n}) - \inf_{z'_i} f(Z_{1:i-1}, z'_i, Z_{i+1:n}) \right)$$

Hence

$$(f(Z) - X_i)^2 \leq \frac{1}{4} L_i^2,$$

and the proposition follows. ■

## 2.3 Self-Bounding Functions

- Another simple property which is satisfied for many important examples is the so-called *self-bounding property*.

**Definition (*Self-Bounding Property*)**

A **nonnegative** function  $f : \mathcal{X}^n \rightarrow [0, \infty)$  has the **self-bounding property** if there exist functions  $f_i : \mathcal{X}^{n-1} \rightarrow \mathbb{R}$  such that for all  $z_1, \dots, z_n \in \mathcal{X}$  and all  $i = 1, \dots, n$ ,

$$0 \leq f(z_1, \dots, z_n) - f_i(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n) \leq 1 \quad (17)$$

and also

$$\sum_{i=1}^n (f(z_1, \dots, z_n) - f_i(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n)) \leq f(z_1, \dots, z_n). \quad (18)$$

- **Remark** Clearly if  $f$  has the *self-bounding property*,

$$\sum_{i=1}^n (f(z_1, \dots, z_n) - f_i(z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n))^2 \leq f(z_1, \dots, z_n) \quad (19)$$

Taking expectation on both sides, we have the following inequality

- **Corollary 2.3** [Boucheron et al., 2013]  
If  $f$  has the *self-bounding property*, then

$$\text{Var}(f(Z)) \leq \mathbb{E}[f(Z)].$$

- **Remark (*Relative Stability*)** [Boucheron et al., 2013]

A sequence of nonnegative random variables  $(Z_n)_{n \in \mathbb{N}}$  is said to be **relatively stable** if

$$\frac{Z_n}{\mathbb{E}[Z_n]} \xrightarrow{\mathbb{P}} 1.$$

This property guarantees that *the random fluctuations* of  $Z_n$  around its *expectation* are of *negligible size* when compared to the expectation, and therefore *most information about the size of  $Z_n$  is given by  $\mathbb{E}[Z_n]$* .

*Bounding the variance of  $Z_n$  by its expected value implies, in many cases, the relative stability* of  $(Z_n)_{n \in \mathbb{N}}$ . If  $Z_n$  has the *self-bounding property*, then, by *Chebyshev's inequality*, for all  $\epsilon > 0$ ,

$$\mathbb{P} \left\{ \left| \frac{Z_n}{\mathbb{E}[Z_n]} - 1 \right| > \epsilon \right\} \leq \frac{\text{Var}(Z_n)}{\epsilon^2 (\mathbb{E}[Z_n])^2} \leq \frac{1}{\epsilon^2 \mathbb{E}[Z_n]}.$$

Thus, for relative stability, it suffices to have  $\mathbb{E}[Z_n] \rightarrow \infty$ .

- An important class of functions satisfying the *self-bounding property* consists of the so-called *configuration functions*.

**Definition (*Configuration Function*)**

Assume that we have a property  $\Pi$  *defined over the union of finite products* of a set  $\mathcal{X}$ , that is, a sequence of sets

$$\Pi_1 \subset \mathcal{X}, \Pi_2 \subset \mathcal{X} \times \mathcal{X}, \dots, \Pi_n \subset \mathcal{X}^n.$$

We say that  $(z_1, \dots, z_m) \in \mathcal{X}^m$  *satisfies the property*  $\Pi$  if  $(z_1, \dots, z_m) \in \Pi_m$ .

We assume that  $\Pi$  is **hereditary** in the sense that if  $(z_1, \dots, z_m)$  satisfies  $\Pi$  then so does **any sub-sequence**  $\{z_{i_1}, \dots, z_{i_k}\}$  of  $(z_1, \dots, z_m)$ .

The function  $f$  that maps any vector  $z = (z_1, \dots, z_n)$  to **the size of a largest sub-sequence satisfying**  $\Pi$  is **the configuration function** associated with property  $\Pi$ .

- **Corollary 2.4** [Boucheron et al., 2013]

Let  $f$  be a **configuration function**, and let  $X = f(Z_1, \dots, Z_n)$ , where  $Z_1, \dots, Z_n$  are **independent** random variables. Then

$$\text{Var}(f(Z)) \leq \mathbb{E}[f(Z)].$$

**Proof:** It suffices to show that **any configuration function is self-bounding**. Let  $X_i := f(Z_{(-i)}) = f(Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$ . By definition of configuration function, the condition  $0 \leq X - X_i \leq 1$  is trivially satisfied.

On the other hand, assume that  $X = k$  and let  $\{Z_{i_1}, \dots, Z_{i_k}\} \subset \{Z_1, \dots, Z_n\}$  be a sub-sequence of cardinality  $k$  such that  $f_k(Z_{i_1}, \dots, Z_{i_k}) = k$ . (Note that by *the definition of a configuration function* such a sub-sequence exists.) Clearly, if the index  $i$  is such that  $i \notin \{i_1, \dots, i_k\}$  then  $X = X_i$ , and therefore

$$\sum_{i=1}^n (X - X_i) \leq X$$

is also satisfied, which concludes the proof. ■

- **Example (VC Dimension)**

Let  $\mathcal{H}$  be an arbitrary collection of subsets of  $\mathcal{X}$ , and let  $x = (x_1, \dots, x_n)$  be a vector of  $n$  points of  $\mathcal{X}$ . Define the **trace** of  $\mathcal{H}$  on  $x$  by

$$\text{tr}(x) = \{A \cap \{x_1, \dots, x_n\} : A \in \mathcal{H}\}.$$

**The shatter coefficient**, (or *Vapnik-Chervonenkis growth function*) of  $\mathcal{H}$  in  $x$  is  $\tau_{\mathcal{H}}(x) = |\text{tr}(x)|$ , the size of the trace.  $\tau_{\mathcal{H}}(x)$  is the number of different subsets of the  $n$ -point set  $\{x_1, \dots, x_n\}$  generated by intersecting it with elements of  $\mathcal{H}$ . A subset  $\{x_{i_1}, \dots, x_{i_k}\}$  of  $\{x_1, \dots, x_n\}$  is said to be **shattered** if  $2^k = T(x_{i_1}, \dots, x_{i_k})$ .

**The VC dimension**  $D(x)$  of  $\mathcal{H}$  (with respect to  $x$ ) is the cardinality  $k$  of the largest shattered subset of  $x$ . From the definition it is obvious that  $f(x) = D(x)$  is a **configuration function** (associated with the property of “**shatteredness**”) and therefore if  $X_1, \dots, X_n$  are *independent random variables*, then

$$\text{Var}(D(X)) \leq \mathbb{E}[D(X)].$$

## 2.4 Applications

### 2.4.1 Kernel Density Estimation

- **Example (Kernel Density Estimation)**

Let  $Z_1, \dots, Z_n$  be i.i.d. samples drawn according to some (unknown) density  $\phi$  on the real

line. The density is estimated by the kernel estimate

$$\phi_n(z) = \frac{1}{n h_n} \sum_{i=1}^n K\left(\frac{z - Z_i}{h_n}\right),$$

where  $h_n > 0$  is a *smoothing parameter*, and  $K$  is a nonnegative function with  $\int K(z) = 1$ . The performance of the estimate is typically measured by **the  $L_1$  error**:

$$X(n) := f(Z_1, \dots, Z_n) = \int |\phi(z) - \phi_n(z)| dz.$$

It is easy to see that

$$\begin{aligned} |f(z_1, \dots, z_n) - f_i(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n)| &\leq \frac{1}{n h_n} \int \left| K\left(\frac{z - z_i}{h_n}\right) - K\left(\frac{z - z'_i}{h_n}\right) \right| dz \\ &\leq \frac{2}{n}, \end{aligned}$$

so without further work we obtain

$$\text{Var}(X(n)) \leq \frac{1}{n}$$

It is known that for every  $\phi$ ,  $\sqrt{n} \mathbb{E}[X(n)] \rightarrow \infty$ , which implies, by *Chebyshev's inequality*, that for every  $\epsilon > 0$

$$\mathbb{P}\left\{\left|\frac{X(n)}{\mathbb{E}[X(n)]} - 1\right| > \epsilon\right\} = \mathbb{P}\{|X(n) - \mathbb{E}[X(n)]| > \epsilon \mathbb{E}[X(n)]\} \leq \frac{\text{Var}(X(n))}{\epsilon^2 (\mathbb{E}[X(n)])^2} \rightarrow 0$$

as  $n \rightarrow \infty$ . That is,  $\frac{X(n)}{\mathbb{E}[X(n)]} \rightarrow 1$  **in probability**, or in other words,  $X(n)$  is **relatively stable**. This means that **the random  $L_1$ -error** essentially *behaves like its expected value*.

By bounded difference inequality, we have

$$\mathbb{P}\{|X(n) - \mathbb{E}[X(n)]| \geq t\} \leq 2 \exp\left(-\frac{nt^2}{2}\right) \quad \blacksquare$$

## 2.4.2 Convex Poincaré Inequality

- **Theorem 2.5 (Convex Poincaré Inequality)** [Boucheron et al., 2013]

Let  $Z_1, \dots, Z_n$  be **independent** random variables taking values in the interval  $[0, 1]$  and let  $f : [0, 1]^n \rightarrow \mathbb{R}$  be a **separately convex function** whose partial derivatives exist; that is, for every  $i = 1, \dots, n$  and fixed  $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ ,  $f$  is a convex function of its  $i$ -th variable. Then  $f(Z) = f(Z_1, \dots, Z_n)$  satisfies

$$\text{Var}(f(Z)) \leq \mathbb{E} \left[ \|\nabla f(Z)\|_2^2 \right]. \quad (20)$$

**Proof:** The proof is an easy consequence of the Efron-Stein inequality, because it suffices to bound the random variable  $\sum_{i=1}^n (X - X_i)^2$  where  $X_i := \inf_{z'_i} f(Z_1, \dots, Z_{i-1}, z_i, Z_{i+1}, \dots, Z_n)$ . Denote by  $Z'_i$  the value of  $z'_i$  for which the minimum is achieved. This is guaranteed by **continuity** and the **compactness** of the domain of  $f$ . Then, writing  $\bar{Z}_{(i)} = (Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_n)$ ,

we have

$$\begin{aligned}
\sum_{i=1}^n (X - X_i)^2 &= \sum_{i=1}^n (f(Z) - f(\bar{Z}_{(i)}))^2 \\
&\leq \sum_{i=1}^n \left( \frac{\partial f}{\partial z_i}(Z) \right)^2 (Z - \bar{Z}_{(i)})^2 \quad (\text{by separate convexity}) \\
&\leq \sum_{i=1}^n \left( \frac{\partial f}{\partial z_i}(Z) \right)^2 = \|\nabla f(Z)\|_2^2. \quad \blacksquare
\end{aligned}$$

- **Remark (*Dimension-Free Concentration*)**

Note that the *convex Poincaré inequality* provides a **dimension-free concentration** for an **arbitrary sub-Gaussian distribution** given that  $f$  is separately *convex*.

### 2.4.3 Gaussian Poincaré Inequality

- **Theorem 2.6 (*Gaussian Poincaré Inequality*)** [Boucheron et al., 2013]

Let  $Z = (Z_1, \dots, Z_n)$  be a vector of **i.i.d. standard Gaussian** random variables (i.e.  $Z$  is a Gaussian vector with **zero mean** vector and **identity covariance matrix**). Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be any **continuously differentiable** function. Then

$$\text{Var}(f(Z)) \leq \mathbb{E} \left[ \|\nabla f(Z)\|_2^2 \right]. \quad (21)$$

**Proof:** We may assume that  $\mathbb{E} \left[ \|\nabla f(Z)\|_2^2 \right] < \infty$ , since otherwise the inequality is trivial.

The proof is based on a *double use* of the *Efron-Stein inequality*. A first straightforward use of it reveals that it suffices to prove the theorem when the dimension  $n$  equals 1. Thus, the problem reduces to show that

$$\text{Var}(f(Z)) \leq \mathbb{E} \left[ (f'(Z))^2 \right], \quad (22)$$

where  $f : \mathbb{R} \rightarrow \mathbb{R}$  is any continuously differentiable function on the real line and  $Z$  is a standard normal random variable.

1. First, notice that it suffices to prove this inequality when  $f$  has a **compact support** and is **twice continuously differentiable**.
2. Now let  $\epsilon_1, \dots, \epsilon_n$  be *independent Rademacher random variables* and introduce

$$S_n := \frac{1}{\sqrt{n}} \sum_{j=1}^n \epsilon_j.$$

Note that  $\epsilon_i \in \{-1, +1\}$  with equal probability, thus

$$\begin{aligned}
\mathbb{E}_{(-i)}[S_n] &= \frac{1}{2} \left[ \left( \frac{1}{\sqrt{n}} \sum_{j \neq i} \epsilon_j + \frac{1}{\sqrt{n}} \right) + \left( \frac{1}{\sqrt{n}} \sum_{j \neq i} \epsilon_j - \frac{1}{\sqrt{n}} \right) \right] \\
&= \frac{1}{2} \left[ \left( S_n + \frac{1 - \epsilon_i}{\sqrt{n}} \right) + \left( S_n - \frac{1 + \epsilon_i}{\sqrt{n}} \right) \right]
\end{aligned}$$



Since for every  $i$

$$\text{Var}_{(-i)}(f(S_n)) = \frac{1}{4} \left( f\left(S_n + \frac{1 - \epsilon_i}{\sqrt{n}}\right) - f\left(S_n - \frac{1 + \epsilon_i}{\sqrt{n}}\right) \right)^2,$$

applying the *Efron-Stein inequality* again, we obtain

$$\text{Var}(f(S_n)) \leq \frac{1}{4} \sum_{i=1}^n \mathbb{E} \left[ \left( f\left(S_n + \frac{1 - \epsilon_i}{\sqrt{n}}\right) - f\left(S_n - \frac{1 + \epsilon_i}{\sqrt{n}}\right) \right)^2 \right] \quad (23)$$

**The central limit theorem** implies that  $S_n$  converges in distribution to  $Z$ , where  $Z$  has the standard normal law. Hence  $\text{Var}(f(S_n))$  converges to  $\text{Var}(f(Z))$ .

3. Let  $K$  denote the supremum of the absolute value of the second derivative of  $f$ . Taylor's theorem implies that, for every  $i$ ,

$$\left| f\left(S_n + \frac{1 - \epsilon_i}{\sqrt{n}}\right) - f\left(S_n - \frac{1 + \epsilon_i}{\sqrt{n}}\right) \right| \leq \frac{2}{\sqrt{n}} |f'(S_n)| + \frac{2K}{n}$$

and therefore

$$\frac{n}{4} \left( f\left(S_n + \frac{1 - \epsilon_i}{\sqrt{n}}\right) - f\left(S_n - \frac{1 + \epsilon_i}{\sqrt{n}}\right) \right)^2 \leq (f'(S_n))^2 + \frac{2K}{\sqrt{n}} |f'(S_n)| + \frac{K^2}{n}$$

This and the central limit theorem imply that

$$\limsup_{n \rightarrow \infty} \frac{1}{4} \sum_{i=1}^n \mathbb{E} \left[ \left( f\left(S_n + \frac{1 - \epsilon_i}{\sqrt{n}}\right) - f\left(S_n - \frac{1 + \epsilon_i}{\sqrt{n}}\right) \right)^2 \right] = \mathbb{E} [(f'(Z))^2],$$

which means that (23) leads to (22) by letting  $n$  go to infinity. ■

• **Remark (*Lipschitz Function of Gaussian Vector*)**

A straightforward consequence of the *Gaussian Poincaré inequality* is that, whenever  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is *L-Lipschitz*, that is, for all  $x, y \in \mathbb{R}^n$ ,

$$|f(x) - f(y)| \leq L \|x - y\|$$

and  $Z$  is a **standard Gaussian random vector**, then

$$\text{Var}(f(Z)) \leq L. \quad (24)$$

Note that this is independent from the dimensionality of function domain.

## 2.5 A Proof of the Efron-Stein Inequality Based on Duality

• **Proposition 2.7 (*Variational Principle for Variance Estimator*)** [Boucheron et al., 2013]

If  $Y$  is a real-valued square-integrable random variable ( $Y \in L^2$  in short), then

$$\text{Var}(Y) = \sup_{T \in L^2} (2\text{Cov}(Y, T) - \text{Var}(T)). \quad (25)$$

**Proof:** since  $\text{Var}(Y-T) \geq 0$ , and

$$\text{Var}(Y) \geq 2\text{Cov}(Y, T) - \text{Var}(T)$$

and since this inequality becomes an equality whenever  $T = Y$ , the duality formula follows. ■

- **Remark** Let  $X = f(Z)$  for  $Z = (Z_1, \dots, Z_n)$  and define  $\mathbb{E}_i[X] := \mathbb{E}[f(Z)|Z_1, \dots, Z_i]$ . Then consider the telescoping sum

$$(f(Z))^2 - (\mathbb{E}[f(Z)])^2 = \sum_{i=1}^n ((\mathbb{E}_i[f(Z)])^2 - (\mathbb{E}_{i-1}[f(Z)])^2),$$

which leads to

$$\text{Var}(f(Z)) = \mathbb{E}[(f(Z))^2] - (\mathbb{E}[f(Z)])^2 = \sum_{i=1}^n \mathbb{E}[(\mathbb{E}_i[f(Z)])^2 - (\mathbb{E}_{i-1}[f(Z)])^2]$$

Note that  $f(Z) - \mathbb{E}[f(Z)] = \sum_{i=1}^n (\mathbb{E}_i[f(Z)] - \mathbb{E}_{i-1}[f(Z)])$

$$\text{Var}(f(Z)) = \mathbb{E}[(f(Z) - \mathbb{E}[f(Z)])^2] = \sum_{i=1}^n \mathbb{E}[(\mathbb{E}_i[f(Z)] - \mathbb{E}_{i-1}[f(Z)])^2]$$

Thus

$$\sum_{i=1}^n \mathbb{E}[(\mathbb{E}_i[f(Z)] - \mathbb{E}_{i-1}[f(Z)])^2] = \sum_{i=1}^n \mathbb{E}[(\mathbb{E}_i[f(Z)])^2 - (\mathbb{E}_{i-1}[f(Z)])^2]$$

Similarly to our first proof of the Efron-Stein inequality, the independence of the variables  $X_1, \dots, X_n$  is used by noting that

$$\mathbb{E}_{(-i)}[\mathbb{E}_i[f(Z)]] = \mathbb{E}_{i-1}[f(Z)]$$

and therefore

$$\sum_{i=1}^n \mathbb{E}[(\mathbb{E}_i[f(Z)])^2 - (\mathbb{E}_{i-1}[f(Z)])^2] = \mathbb{E}[\text{Var}_{(-i)}(\mathbb{E}_i[f(Z)])]$$

In other words, we have proven the following alternative formulation (using *independence* but *without using the orthogonality structure of the martingale differences*):

$$\text{Var}(f(Z)) = \sum_{i=1}^n \mathbb{E}[\text{Var}_{(-i)}(\mathbb{E}_i[f(Z)])]. \quad (26)$$

It remains to commute the  $\text{Var}_{(-i)}$  and  $\mathbb{E}_i$  operators and this is precisely the step where we use a *duality argument*.

- **Lemma 2.8** [Boucheron et al., 2013]  
For every  $i = 1, \dots, n$ ,

$$\mathbb{E}[\text{Var}_{(-i)}(\mathbb{E}_i[f(Z)])] \leq \mathbb{E}[\text{Var}_{(-i)}(f(Z))].$$

**Proof:** Applying the duality formula of (25) conditionally on  $Z_{(-i)}$ , we show that for any square-integrable variable  $T$ ,

$$2\text{Cov}_{(-i)}(f(Z), T) - \text{Var}_{(-i)}(T) \leq \text{Var}_{(-i)}(f(Z)) \quad (27)$$

But if we take  $T$  to be  $(Z_1, \dots, Z_i)$ -measurable, then

$$\begin{aligned} \mathbb{E} [\text{Cov}_{(-i)}(f(Z), T)] &= \mathbb{E} [f(Z) (T - \mathbb{E}_i [T])] \\ &= \mathbb{E} [\mathbb{E}_i [f(Z)] (T - \mathbb{E}_i [T])] \\ &= \mathbb{E} [\text{Cov}_{(-i)}(\mathbb{E}_i [f(Z)], T)] . \end{aligned}$$

Hence, choosing  $T = \mathbb{E}_i [f(Z)]$  leads to

$$\mathbb{E} [\text{Cov}_{(-i)}(f(Z), \mathbb{E}_i [f(Z)])] = \mathbb{E} [\text{Var}_{(-i)}(\mathbb{E}_i [f(Z)])]$$

and therefore, by (27),

$$\mathbb{E} [\text{Var}_{(-i)}(\mathbb{E}_i [f(Z)])] \leq \mathbb{E} [\text{Var}_{(-i)}(f(Z))] \quad \blacksquare$$

- **Remark** Combining Lemma above with the decomposition (26) leads to

$$\text{Var}(f(Z)) \leq \sum_{i=1}^n \mathbb{E} [\text{Var}_{(-i)}(f(Z))] \quad (28)$$

which is equivalent to the Efron–Stein inequality.  $\blacksquare$

## 2.6 Exponential Tail Bounds via the Efron-Stein Inequality

- **Remark (*Assumption*)** [Boucheron et al., 2013]  
Suppose  $Z := (Z_1, \dots, Z_n)$  are independent random variables and  $X := f(Z)$ .  $X_i$  is *the  $i$ -th Jackknife replication* of  $X$ , which is a function of  $Z_{(-i)} := (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$ , *the  $i$ -th Jackknife sample*.

Assume that there exists a positive constant  $\nu$  such that

$$\sum_{i=1}^n (X - X_i)_+^2 \leq \nu \quad (29)$$

holds almost surely. Define, for any  $\alpha \in (0, 1)$ , *the  $\alpha$ -quantile* of  $X := f(Z)$  by

$$Q_\alpha = \inf \{ \alpha : \mathbb{P} \{X \leq x\} \geq \alpha \} .$$

In particular, we denote *the median* of  $Z$  by  $\text{Med}(Z) = Q_{1/2}$ .

- **Remark (*Clipping of Function*)**  
The trick is to use the Efron-Stein inequality for the random variable  $g_{a,b}(Z) = g_{a,b}(Z_1, \dots, Z_n)$  where  $b \geq a$  and the function  $g_{a,b} : \mathcal{X}^n \rightarrow \mathbb{R}$  is defined as

$$g_{a,b}(z) = \begin{cases} b & \text{if } f(z) \geq b \\ f(z) & \text{if } a < f(z) < b \\ a & \text{if } f(z) \leq a \end{cases}$$

It is the clipping of function  $f(z)$  taking values within  $[a, b]$ , i.e.  $g_{a,b}(z) = \max \{a, \min \{f(z), b\}\}$ .

- **Remark (*Bounding Variance of  $g_{a,b}(Z)$* )**

First observe that if  $a \geq \text{Med}(X)$ , then  $\mathbb{E}[g_{a,b}(Z)] \leq (a+b)/2$  and therefore the lower bound of the variance is

$$\text{Var}(g_{a,b}(Z)) \geq \frac{(b-a)^2}{4} \mathbb{P}\{g_{a,b}(Z) = b\} = \frac{(b-a)^2}{4} \mathbb{P}\{f(Z) \geq b\}.$$

On the other hand, we may use the Efron-Stein inequality to obtain *an upper bound* for the variance of  $g_{a,b}(Z)$ . To this end, observe that if  $f(z) \leq a$  then

$$g_{a,b}(\bar{z}^{(i)}) \geq g_{a,b}(z),$$

for

$$\bar{z}^{(i)} := (z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n)$$

and so

$$\begin{aligned} \sum_{i=1}^n \mathbb{E} \left[ \left( g_{a,b}(Z) - g_{a,b}(\bar{Z}^{(i)}) \right)^2 \right] &= 2 \sum_{i=1}^n \mathbb{E} \left[ \left( g_{a,b}(Z) - g_{a,b}(\bar{Z}^{(i)}) \right)_+^2 \right] \\ &\leq 2 \sum_{i=1}^n \mathbb{E} \left[ \mathbf{1}_{\{f(Z) > a\}} \left( g_{a,b}(Z) - g_{a,b}(\bar{Z}^{(i)}) \right)_+^2 \right] \\ &\leq 2\nu \mathbb{P}\{f(Z) > a\}, \end{aligned}$$

where, in the last step, we used the fact that condition (29) implies that

$$\sum_{i=1}^n \left( g_{a,b}(Z) - g_{a,b}(\bar{Z}^{(i)}) \right)_+^2 \leq \sum_{i=1}^n \left( f(Z) - f(\bar{Z}^{(i)}) \right)_+^2 \leq \nu.$$

Comparing the obtained upper and lower bounds for  $\text{Var}(g_{a,b}(Z))$ , we get

$$b - a \leq \sqrt{8\nu \frac{\mathbb{P}\{f(Z) > a\}}{\mathbb{P}\{f(Z) \geq b\}}}.$$

- **Remark (*Bounding the Distance between Quantiles of  $f(Z)$* )**

To this end, let  $0 < \delta < \gamma \leq 1/2$  and choose  $a = Q_{1-\gamma}$  and  $b = Q_{1-\delta}$ . Then  $\mathbb{P}\{f(Z) > a\} \leq \gamma$  and  $\mathbb{P}\{f(Z) \geq b\} \geq \delta$  and therefore the distance between any two quantiles of  $X = f(Z)$  (to the *right* of the median) can be *bounded* as

$$b - a = Q_{1-\delta} - Q_{1-\gamma} \leq \sqrt{\frac{8\nu\gamma}{\delta}}$$

It is instructive to choose  $\gamma = 2^{-k}$  and  $\delta = 2^{-(k+1)}$  for some integer  $k \geq 1$ . Then, denoting  $a_k = Q_{1-2^{-k}}$ , we get

$$a_{k+1} - a_k \leq 4\sqrt{\nu},$$

so the difference between *consecutive quantiles* corresponding to *exponentially decreasing tail probabilities* is *bounded by a constant*. In particular, by summing this inequality for  $k = 1, \dots, m$ , we have

$$a_{m+1} \leq \text{Med}(f(Z)) + 4m\sqrt{\nu}$$

which implies that for all  $t > 0$ ,

$$\mathbb{P}\{f(Z) > \text{Med}(f(Z)) + t\} \leq 2^{-\frac{t}{4\sqrt{\nu}}}. \quad (30)$$

• **Remark** (*Bounding the Deviations from Mean instead of Median of  $f(Z)$* )

An alternative route to obtain exponential bounds is by applying the *Efron-Stein inequality* to  $\exp(\lambda X/2)$  with  $\lambda > 0$ . Then, by the *mean-value theorem*,

$$\begin{aligned} \mathbb{E}[\exp(\lambda X)] - \left( \mathbb{E} \left[ \exp \left( \frac{\lambda X}{2} \right) \right] \right)^2 &\leq \mathbb{E} \left[ \sum_{i=1}^n \left( \exp \left( \frac{\lambda X}{2} \right) - \exp \left( \frac{\lambda X'}{2} \right) \right)_+^2 \right] \quad (\text{Efron-Stein inequality}) \\ &\leq \frac{\lambda^2}{4} \mathbb{E} \left[ \exp(\lambda X) \sum_{i=1}^n (X - X')_+^2 \right] \quad (\text{mean-value theorem}) \end{aligned}$$

Now we may use our condition (29) to derive

$$\mathbb{E}[\exp(\lambda X)] - \left( \mathbb{E} \left[ \exp \left( \frac{\lambda X}{2} \right) \right] \right)^2 \leq \frac{\nu \lambda^2}{4} \mathbb{E}[\exp(\lambda X)]$$

or equivalently

$$\left( 1 - \frac{\nu \lambda^2}{4} \right) \Phi(\lambda) \leq \left( \Phi \left( \frac{\lambda}{2} \right) \right)^2,$$

where  $\Phi(\lambda) := \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))]$  is the *moment generating function* of  $X - \mathbb{E}[X]$ .

**Lemma 2.9** *Let  $g : (0, 1) \rightarrow (0, \infty)$  be a function such that  $\lim_{x \rightarrow 0} (g(x) - 1)/x = 0$ . If for every  $x \in (0, 1)$*

$$(1 - x^2)g(x) \leq g(x/2)^2,$$

*then*

$$g(x) \leq (1 - x^2)^{-2}.$$

Since  $\Phi(0) = 1$  and  $\Phi'(0) = 0$ , we may apply above Lemma to the function  $x \rightarrow \Phi(2x\nu^{-1/2})$  and get, for every  $\lambda \in (0, 2\nu^{-1/2})$ ,

$$\Phi(\lambda) \leq \left( 1 - \frac{\nu \lambda^2}{4} \right)^{-2}. \quad (31)$$

Thus, the *Efron-Stein inequality* may be used to prove *exponential integrability* of  $X$ .

Moreover, since by (31)  $\Phi(\nu^{-1/2}) \leq 2$ , by *Markov's inequality*, for every  $t > 0$ ,

$$\mathbb{P}\{f(Z) - \mathbb{E}[f(Z)] \geq t\} \leq 2 \exp \left( -\frac{t}{\sqrt{\nu}} \right). \quad (32)$$

This inequality has the same form as the one derived using the first method of this section but now we *bound deviations* from the *mean* instead of the *median* and the constants are somewhat better. ■

## References

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Sidney I Resnick. *A probability path*. Springer Science & Business Media, 2013.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.