Lecture 2: Concentration without Independence

Tianpei Xie

Jan. 6th., 2023

Contents

1 Martingale-based Methods

1.1 Martingale

- **Definition** (***Martingale***) [Resnick, 2013]
Let $\{X_n, n \geq 0\}$ be a stochastic process on $(\Omega, \mathscr{F})$ and $\{\mathscr{F}_n, n \geq 0\}$ be a ***filtration***; that is, $\{\mathscr{F}_n, n \geq 0\}$ is an *increasing sub σ-fields* of $\mathscr{F}$
$$\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}.$$
Then $\{(X_n, \mathscr{F}_n), n \geq 0\}$ is a ***martingale (mg)*** if

  1. X_n is ***adapted*** in the sense that for each n, $X_n \in \mathscr{F}_n$; that is, X_n is $\mathscr{F}_n$-measurable.
  2. $X_n \in L_1$; that is $\mathbb{E}\left[|X_n|\right] < \infty$ for $n \geq 0$.
  3. For $0 \leq m < n$
$$\mathbb{E}\left[X_n \mid \mathscr{F}_m\right] = X_m, \quad \text{a.s.} \tag{1}$$

  If the equality of (1) is replaced by $\geq$; that is, things are getting better on the average:
$$\mathbb{E}\left[X_n \mid \mathscr{F}_m\right] \geq X_m, \quad \text{a.s.} \tag{2}$$
then $\{X_n\}$ is called a ***sub-martingale (submg)*** while if things are getting worse on the average
$$\mathbb{E}\left[X_n \mid \mathscr{F}_m\right] \leq X_m, \quad \text{a.s.} \tag{3}$$
$\{X_n\}$ is called a ***super-martingale (supermg)***.

- **Remark** $\{X_n\}$ is ***martingale*** if it is *both* a ***sub*** and ***supermartingale***. $\{X_n\}$ is a ***supermartingale*** if and only if $\{-X_n\}$ is a ***submartingale***.

- **Remark** If $\{X_n\}$ is a ***martingale***, then $\mathbb{E}\left[X_n\right]$ is *constant*. In the case of a ***submartingale***, *the mean increases* and for a ***supermartingale***, *the mean decreases.*

- **Proposition 1.1** *[Resnick, 2013]*
If $\{(X_n, \mathscr{F}_n), n \geq 0\}$ is a ***(sub, super) martingale****, then*
$$\{(X_n, \sigma(X_0, X_1, \ldots, X_n)), n \geq 0\}$$
is also a ***(sub, super) martingale****.*

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

  1. For $j \geq 0$, $\mathscr{B}_j \subset \mathscr{B}_{j+1}$.
  2. For $j \geq 0$, $d_j \in L_1$, $d_j \in \mathscr{B}_j$; that is, d_j is *absolutely integrable* and *$\mathscr{B}_j$-measurable.*
  3. For $j \geq 0$,
$$\begin{aligned} \mathbb{E}\left[d_{j+1} | \mathscr{B}_j\right] &= 0, && (\textit{martingale difference / fair sequence}); \\ &\geq 0, && (\textit{submartingale difference / subfair sequence}); \\ &\leq 0, && (\textit{supmartingale difference / supfair sequence}) \end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned} \mathbb{E}\left[d_i\, d_j\right] &= \mathbb{E}\left[\mathbb{E}\left[d_i\, d_j \mid \mathscr{B}_i\right]\right] \\ &= \mathbb{E}\left[d_i \mathbb{E}\left[d_j \mid \mathscr{B}_i\right]\right] = 0. \quad \blacksquare \end{aligned}$$

  A consequence is that

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

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

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

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

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

$$d_n := \mathbb{E}\left[X | \mathscr{B}_n\right] - \mathbb{E}\left[X | \mathscr{B}_{n-1}\right], \quad n \geq 1. \tag{4}$$

Proof: See that

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

- **Example** (***Sums of Independent Random Variables***)
  Suppose that $\{Z_n, n \geq 0\}$ is an ***independent*** *sequence of integrable random variables* satisfying for $n \geq 0$, $\mathbb{E}\left[Z_n\right] = 0$. Set

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

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

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

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

  Proof: See that

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

1.2 Bernstein Inequality for Martingale Difference Sequence

- **Proposition 1.5** ***(Bernstein Inequality, Martingale Difference Sequence Version)*** *[Wainwright, 2019]*
  Let $\{(D_k, \mathscr{B}_k), k \geq 1\}$ *be a* ***martingale difference sequence****, and suppose that*

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

  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*** $\left(\sqrt{\sum_{k=1}^{n}\nu_k^2}\;,\;\alpha_*\right)$ *where* $\alpha_* := \max_{k=1,\dots,n}\alpha_k$. *That is, for any* $|\lambda| < 1/\alpha_*$,

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

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

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

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 λ such that $|\lambda| < 1/\alpha_*$, conditioning on $\mathscr{B}_{n-1}$ and applying iterated expectation yields

$$\begin{aligned}\mathbb{E}\left[\exp\left\{\lambda\left(\sum_{k=1}^{n} D_k\right)\right\}\right] &= \mathbb{E}\left[\exp\left\{\lambda\left(\sum_{k=1}^{n-1} D_k\right)\right\}\mathbb{E}\left[\exp\left\{\lambda D_n\right\}\;\middle|\;\mathscr{B}_{n-1}\right]\right] \\ &\le \mathbb{E}\left[\exp\left\{\lambda\left(\sum_{k=1}^{n-1} D_k\right)\right\}\right]\exp\left(\frac{\lambda^2\nu_k^2}{2}\right),\end{aligned}$$

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

- **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

$$\mathbb{E}\left[\mathbb{E}\left[X|\mathscr{B}_n\right]\right] = \mathbb{E}\left[X\right], \quad \mathbb{E}\left[h(X)g(Y)|Y\right] \overset{a.s.}{=} h(X)\mathbb{E}\left[g(Y)|Y\right]$$

1.3 Azuma-Hoeffding Inequality

- **Corollary 1.6** ***(Azuma-Hoeffding Inequality, Martingale Difference)****[Wainwright, 2019]* *Let* $\{(D_k, \mathscr{B}_k), k \ge 1\}$ *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* $t \ge 0$,

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

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 $k \in \{1, 2, \ldots, n\}$, 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 $f : \mathcal{X}^n \to \mathbb{R}$ satisfies ***the bounded difference inequality*** with parameters $(L_1, \ldots, L_n)$ if, for each index $k = 1, 2, \ldots, n$,

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

- **Corollary 1.7** *(**McDiarmid's Inequality / Bounded Differences Inequality**)[Wainwright, 2019]*
*Suppose that f satisfies **the bounded difference property** (7) with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, X_2, \ldots, X_n)$ has **independent** components. Then*

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

Proof: Consider the associated *martingale difference sequence*

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

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

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

On one hand, we have

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

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

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

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

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

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 Lipschitz Functions of Gaussian Variables

1.6 Applications

2 Bounding Variance

2.1 The Efron-Stein Inequality

- **Remark** (***Variance of Independence Random Variables***)
  Let $X_n = \sum_{i=1}^{n} Z_i$ be the sum of ***independent*** real-valued random variables $Z_1, \ldots, Z_n$. Then we have

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

- **Remark** (***Variance of Smoothing Martingale Difference Sequence***)
  Suppose $X \in L_1$ and $\{\mathscr{B}_n, n \geq 0\}$ is an increasing family of sub σ-algebra of $\mathscr{B}$ formed by

$$\mathscr{B}_n := \sigma(Z_1, \ldots, Z_n).$$

For $n \geq 1$, define

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

From (4) we see that $(d_n, \mathscr{B}_n)$ 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

$$\begin{aligned} \mathrm{Var}(X) = \mathbb{E}\left[\left(\sum_{i=1}^{n} d_i\right)^2\right] &= \sum_{i=1}^{n} \mathbb{E}\left[d_i^2\right] + 2\sum_{i>j} \mathbb{E}[d_i\, d_j] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[d_i^2\right]. \end{aligned} \tag{9}$$

- **Remark** (***Variance of General Functions of Independent Random Variables***)
  Then above formula (9) holds when $X = f(Z_1, \ldots, Z_n)$ for general function $f : \mathbb{R}^n \to \mathbb{R}$ with n independent random variables $(Z_1, \ldots, Z_n)$. By *Fubini's theorem*,

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

  where μ_j is the probability distribution of Z_j for $j \geq 1$. Define the conditional expectation of X given all random variables $(Z_1, \ldots, Z_n)$ ***except for*** Z_i as

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

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

$$\mathbb{E}\left[\mathbb{E}_{(-i)}\left[X\right] | Z_1, \ldots, Z_i\right] = \mathbb{E}\left[X | Z_1, \ldots, Z_{i-1}\right] \tag{10}$$

  Denote $Z_{(-i)} := (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$.

- **Proposition 2.1** *(**Efron-Stein Inequality**) [Boucheron et al., 2013]*
  *Let $Z_1, \ldots, Z_n$ be **independent random variables** and let $X = f(Z)$ be a square-integrable function of $Z = (Z_1, \ldots, Z_n)$. Then*

$$Var(X) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(X - \mathbb{E}_{(-i)}\left[X\right]\right)^2\right] := \nu. \tag{11}$$

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

$$X_i' := f\left(Z_1, \ldots, Z_{i-1}, Z_i', Z_{i+1}, \ldots, Z_n\right),$$

  then

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

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

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

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

  Proof: We begin with the proof of the first statement. Note that, using (10), we may write

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

By *Jensen's inequality* used conditionally,

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

Using (9) $\mathrm{Var}(X) = \sum_{i=1}^n \mathbb{E}\left[d_i^2\right]$, we have

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

we obtain the desired inequality.

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

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

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

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

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

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

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,

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

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

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

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

2.2 Functions with Bounded Differences

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

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

where

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

- **Corollary 2.2** *[Boucheron et al., 2013]*
*If f has the **bounded differences property** with parameters $(L_1, \dots, L_n)$, then*

$$\mathit{Var}(f(X)) \le \frac{1}{4}\sum_{i=1}^{n} L_i^2.$$

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* $f : \mathcal{X}^n \to [0, \infty)$ has the <u>***self-bounding property***</u> if *there exist* functions $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ such that for all $x_1, \dots, x_n \in \mathcal{X}$ and all $i = 1, \dots, n$,

$$0 \le f(x_1, \dots, x_n) - f_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \le 1 \tag{12}$$

and also

$$\sum_{i=1}^{n} \left(f(x_1, \dots, x_n) - f_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)\right) \le f(x_1, \dots, x_n). \tag{13}$$

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

$$\sum_{i=1}^{n} \left(f(x_1, \dots, x_n) - f_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)\right)^2 \le f(x_1, \dots, x_n) \tag{14}$$

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

$$\mathit{Var}(f(X)) \le \mathbb{E}\left[f(X)\right].$$

- **Remark** (***Relative Stability***) [Boucheron et al., 2013]
A sequence of nonnegative random variables $(Z_n)_{n\in\mathbb{N}}$ is said to be <u>***relatively stable***</u> if

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

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*** $\mathbb{E}\left[Z_n\right]$.

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 $\epsilon > 0$,

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

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

- 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 Π ***defined over the union of finite products*** of a set $\mathcal{X}$, that is, a sequence of sets

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

  We say that $(x_1, \ldots, x_m) \in \mathcal{X}^m$ ***satisfies the property*** Π if $(x_1, \ldots, x_m) \in \Pi_m$.

  We assume that Π is ***hereditary*** in the sense that if $(x_1, \ldots, x_m)$ satisfies Π then so does ***any sub-sequence*** $\{x_{i_1}, \ldots, x_{i_k}\}$ of $(x_1, \ldots, x_m)$.

  The function f that maps any vector $x = (x_1, \ldots, x_n)$ to ***the size*** *of a* ***largest sub-sequence*** *satisfying* Π is ***the configuration function*** *associated with property* Π.

- **Corollary 2.4** *[Boucheron et al., 2013]*
  Let f be a ***configuration function****, and let $Z = f(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are* ***independent*** *random variables. Then*

  $$Var(Z) \le \mathbb{E}\left[Z\right].$$

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

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

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

  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 $X_1, \ldots, X_n$ are *independent random variables*, then

  $$\mathrm{Var}(D(X)) \le \mathbb{E}\left[D(X)\right].$$

2.4 Applications

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

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.