Non-Asymptotic Bounds

패턴인식 및 기계학습 여름학교
(Pattern Recognition and Machine Learning Summer School)
2017. 7. 12 (Wed.)

Yung-Kyun Noh (노영균)
Seoul National University

Many slides are borrowed and modified from Gábor Lugosi's concentration inequalities lecture slides!

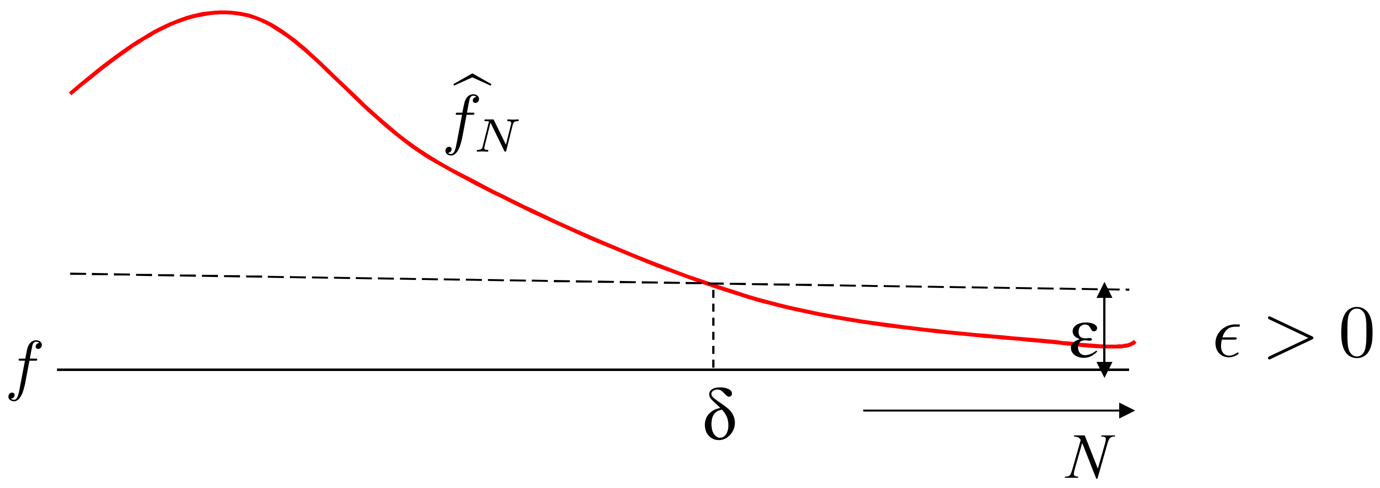
http://machinelearning.snu.ac.kr/PRMLSS2017/nonasymptotic.pdf

Contents

- Asymptotic Convergence
  - Converge in probability
  - Almost sure convergence
  - L2 convergence
  - Maximum likelihood
- Non-Asymptotic Bounds
  - Markov's inequality
  - Chernoff bound
  - Chebyshev's inequality
  - Efron-Stein inequality

Asymptotic / Non-asymptotic Bounds

- Limit, epsilon and delta $\lim_{N\to\infty} \widehat{f}_N = f$



  – For any ε, there exists δ such that

$$|\widehat{f}_N - f| < \epsilon$$

Machine Learning and Consistency

- With increasing number of data
  - Expected loss

$$L = \mathbb{E}_P[l(y, f(\mathbf{x}))]$$

$l(y, y')$: loss function

  - Estimated Error

$$\hat{L} = \sum_n l(y_n, f(\mathbf{x}_n)), \qquad f(\mathbf{x}) \in \mathcal{H}$$

$\mathcal{H}$ satisfies

$$\boxed{\widehat{L} \underset{N \to \infty}{\longrightarrow} L}$$

$$P\{\sup_{f \in \mathcal{H}} (L(f) - \widehat{L}(f)) > \epsilon\} \to 0 \quad \text{for} \quad \epsilon > 0$$

<Uniform convergence>

Interest in This Lecture

- We are interested in bounding random fluctuations of functions of many independent random variables.

- Let $f : \mathcal{X}^n \to \mathbb{R}$ and

$$Z = f(\mathbf{x}_1, \ldots, \mathbf{x}_N)$$

with independent random variables $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

Interest in This Lecture

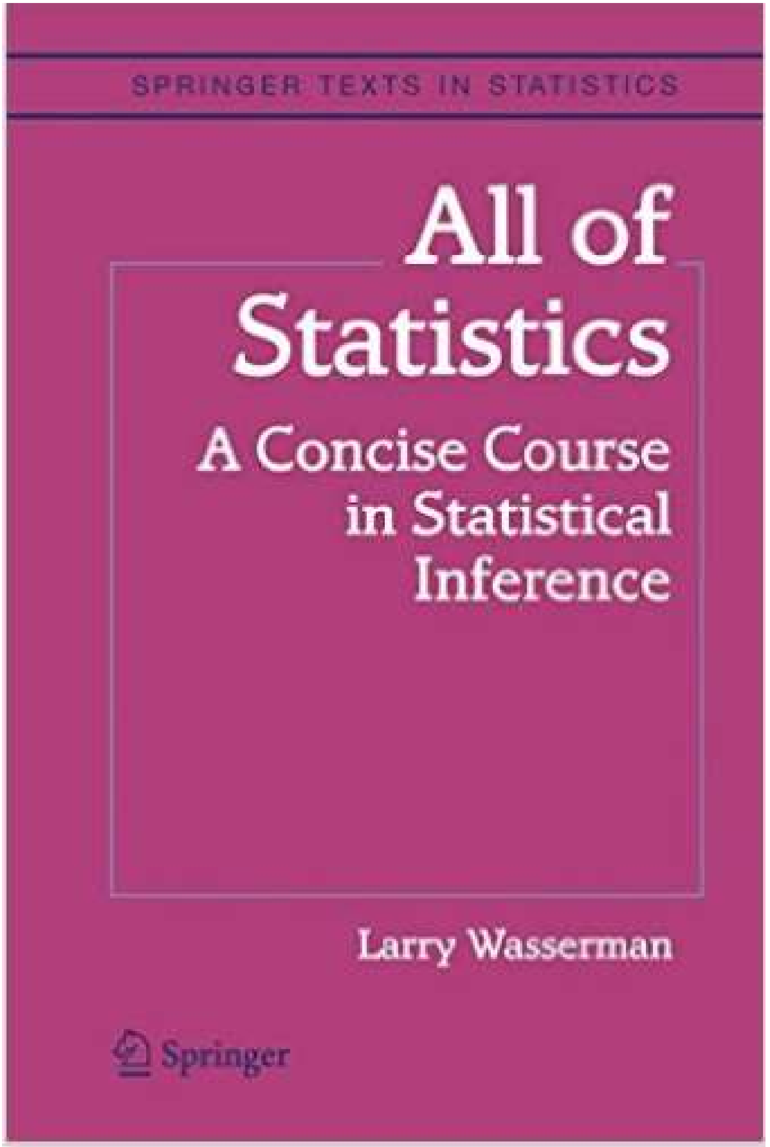
- The estimator will be close enough to the true value (or expectation of the estimator).
  - ***Asymptotically***

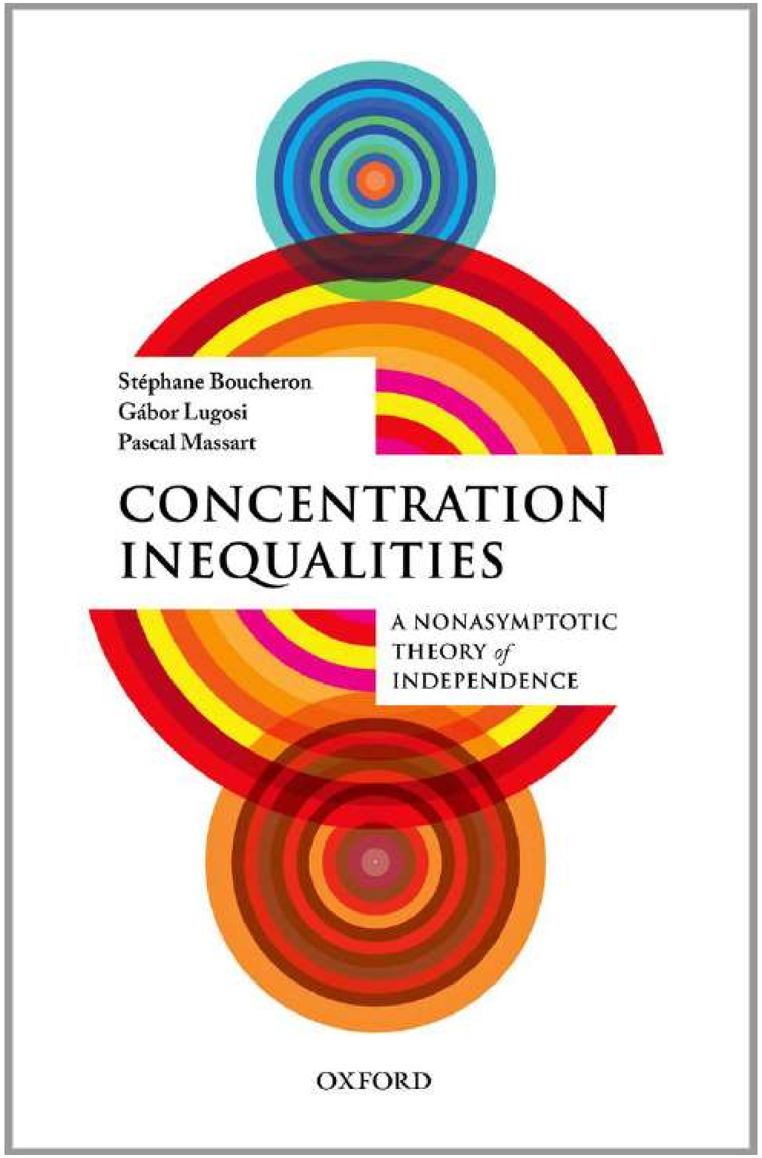
$$\widehat{f}_N \underset{N \to \infty}{\longrightarrow} f$$

  - ***Non-asymptotically (with fixed N)***

$$P\{Z > \mathbb{E}Z + t\} \text{ and } P\{Z < \mathbb{E}Z - t\}$$

$$\text{for } \ t > 0$$


SPRINGER TEXTS IN STATISTICS
All of Statistics
A Concise Course in Statistical Inference
Larry Wasserman
Springer


Stéphane Boucheron
Gábor Lugosi
Pascal Massart
CONCENTRATION INEQUALITIES
A NONASYMPTOTIC THEORY of INDEPENDENCE
OXFORD

ASYMPTOTIC CONVERGENCE

Convergence

- Convergence in probability: $Z_N \overset{P}{\to} Z$

$$P\left\{|Z_N - Z| > \epsilon\right\} \to 0 \quad (N \to \infty, \epsilon > 0)$$

- Almost sure convergence: $Z_N \overset{as}{\to} Z$

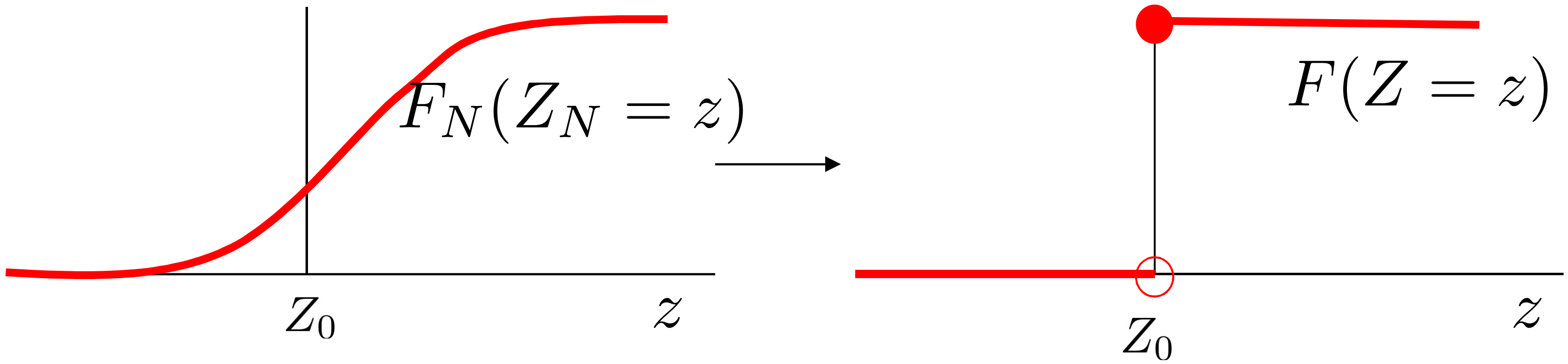
$$P\left(\lim_{N\to\infty} Z_N = Z\right) = 1$$

- L2-convergence: $Z_N \overset{qm}{\to} Z$

$$\mathbb{E}(Z_N - Z)^2 \to 0 \quad (N \to \infty)$$

Convergence

- Convergence in distribution $Z_N \rightsquigarrow Z$



$$\lim_{N \to \infty} F_N(z) = F(z)$$

For this distribution, convergence in distribution implies

$$Z_N \xrightarrow{P} Z \ (= Z_0)$$

Relationship Between Types of Convergence

$$
\begin{tikzcd}
\text{quadratic mean} \arrow[dr] & & \\
& \text{probability} \arrow[r] & \text{distribution} \arrow[l, dashed, bend right, "\text{point-mass distribution}"'] \\
\text{almost sure} \arrow[ur] & &
\end{tikzcd}
$$

Relationship Between Types of Convergence

$$
\begin{tikzcd}
\text{quadratic mean} \arrow[dr] & & \\
& \text{probability} \arrow[r] & \text{distribution} \arrow[l, dashed, bend right, "\text{point-mass distribution}"'] \\
\text{almost sure} \arrow[ur] & &
\end{tikzcd}
$$

- Ex) $Z_N \rightsquigarrow Z$ & $P(Z = Z_0) = 1$

$$\longrightarrow \quad Z_N \xrightarrow{P} Z \ (= Z_0)$$

$$
\begin{aligned}
P(|Z_N - Z_0| > \epsilon) &= P(Z_N < Z_0 - \epsilon) + P(Z_N > Z_0 - \epsilon) \\
&\le P(Z_N \le Z_0 - \epsilon) + P(Z_N > Z_0 - \epsilon) \\
&= F_N(Z_0 - \epsilon) + 1 - F_N(Z_0 + \epsilon) \\
&\to F(Z_0 - \epsilon) + 1 - F(Z_0 + \epsilon) \\
&= 0 + 1 - 1 = 0
\end{aligned}
$$

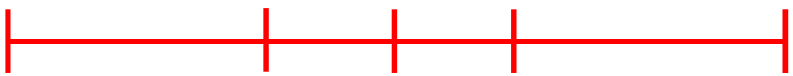
Convergence in Probability

$$X(s) = s \qquad\qquad s \sim \mathrm{Unif}\,[0, 1]$$

A sequence $X_1, X_2, \ldots$ is defined as follows:

$$X_1(s) = s + I_{[0,1]}(s) \qquad X_2(s) = s + I_{[0,\frac{1}{2}]}(s)$$
$$X_3(s) = s + I_{[\frac{1}{2},1]}(s) \qquad X_4(s) = s + I_{[0,\frac{1}{3}]}(s)$$
$$X_5(s) = s + I_{[\frac{1}{3},\frac{2}{3}]}(s) \qquad X_6(s) = s + I_{[\frac{2}{3},1]}(s)$$

$\ldots$



$$P(|X_N - X| \geq \epsilon) = 0 \quad (N \to \infty)$$

Convergence in Probability

$$X_1(s) = s + I_{[0,1]}(s) \qquad X_2(s) = s + I_{[0,\frac{1}{2}]}(s)$$
$$X_3(s) = s + I_{[\frac{1}{2},1]}(s) \qquad X_4(s) = s + I_{[0,\frac{1}{3}]}(s)$$
$$X_5(s) = s + I_{[\frac{1}{3},\frac{2}{3}]}(s) \qquad X_6(s) = s + I_{[\frac{2}{3},1]}(s)$$

$\cdots$

- If $s = 3/8$,

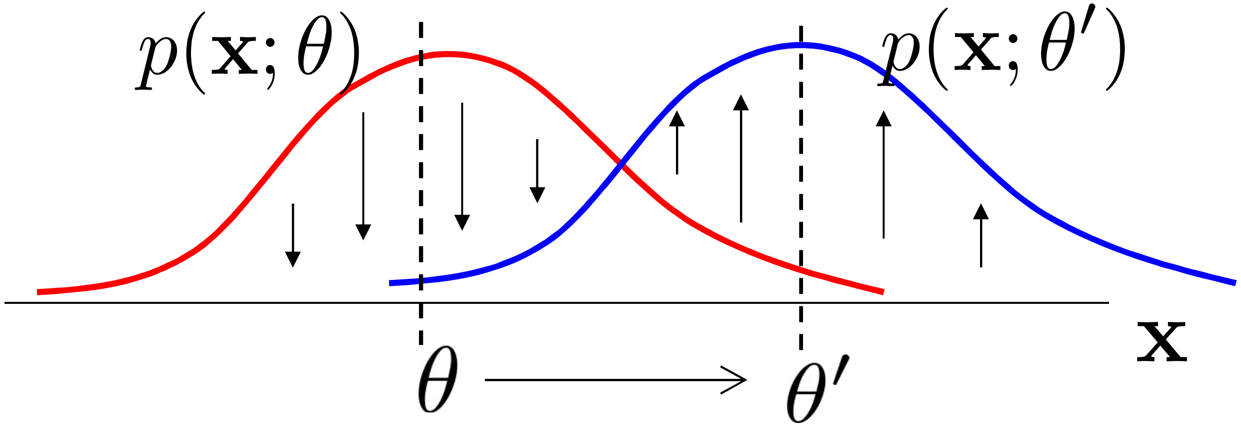
$$X_1(s) = 11/8 \quad X_2(s) = 11/8 \quad X_3(s) = 3/8 \quad X_4(s) = 3/8$$
$$X_5(s) = 11/8 \quad X_6(s) = 3/8 \quad \cdots$$

For every *s*, $X_N(s)$ alternates between *s* and *s+1* infinitely often.

$$\lim_{N\to\infty} X_N \neq X$$ Not a.s. convergence

Convergence of Maximum Likelihood Estimation

- Sensitivity of probability density function with respect to the parameter θ: $\dfrac{\partial \log p(\mathbf{x};\theta)}{\partial \theta} \ (= \nabla_\theta \log p(\mathbf{x};\theta))$



We note

$$\mathbb{E}_{\mathbf{x}\sim p}\left[\nabla_\theta \log p(\mathbf{x};\theta)\right] = \int p(\mathbf{x};\theta)\nabla_\theta \log p(\mathbf{x};\theta) d\mathbf{x}$$

$$= \int p(\mathbf{x};\theta)\frac{\nabla_\theta p(\mathbf{x};\theta)}{p(\mathbf{x};\theta)} d\mathbf{x} = \nabla_\theta \int p(\mathbf{x};\theta) d\mathbf{x} = \nabla_\theta 1 = 0$$

Convergence of MLE

- Fisher information

$$I(\theta) = \mathbb{E}_{\mathbf{x}\sim p}\left[(\nabla_\theta \log p(\mathbf{x};\theta))^2\right]$$
$$= Var\left(\nabla_\theta \log p(\mathbf{x};\theta)\right)$$
$$= -\int \left(\nabla_\theta^2 \log p(\mathbf{x};\theta)\right) p(\mathbf{x};\theta) d\mathbf{x}$$

- Asymptotic Normality of MLE

$$\frac{(\widehat{\theta}_N - \theta)}{\sqrt{1/(N\cdot I(\theta))}} \rightsquigarrow \mathcal{N}(0,1) \quad \left(\text{or} \quad \frac{(\widehat{\theta}_N - \theta)}{\sqrt{1/(N\cdot I(\widehat{\theta}_N))}} \rightsquigarrow \mathcal{N}(0,1)\right)$$

$$\text{se}(\widehat{\theta}_N) = \sqrt{1/(N\cdot I(\theta))} \quad \left(\text{or} \quad \widehat{\text{se}} = \sqrt{1/(N\cdot I(\widehat{\theta}_N))}\right)$$

Convergence of MLE

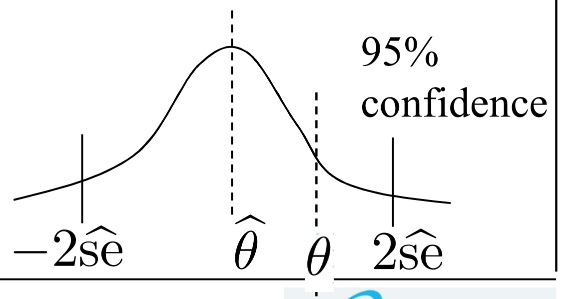
- Ex) $$X_1, \ldots, X_N \sim \text{Bernoulli}(\theta)$$

$$p(x;\theta) = \theta^x (1-\theta)^{1-x} \qquad \widehat{\theta}_N = \sum_{i=1}^{N} \frac{X_i}{N}$$

$$-\frac{\partial^2 \log p(x;\theta)}{\partial \theta^2} = \frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

$$I(\theta) = \mathbb{E}_{x\sim p}\left[-\frac{\partial^2 \log p(x;\theta)}{\partial \theta^2}\right] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$$

$$\widehat{\text{se}}(\widehat{\theta}) = \frac{1}{\sqrt{N \cdot I(\widehat{\theta})}} = \left(\frac{\widehat{\theta}(1-\widehat{\theta})}{N}\right)^{\frac{1}{2}}$$



NON-ASYMPTOTIC CONVERGENCE

Markov's Inequality

- If $Z \geq 0$

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

Markov's Inequality

- If $Z \geq 0$

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

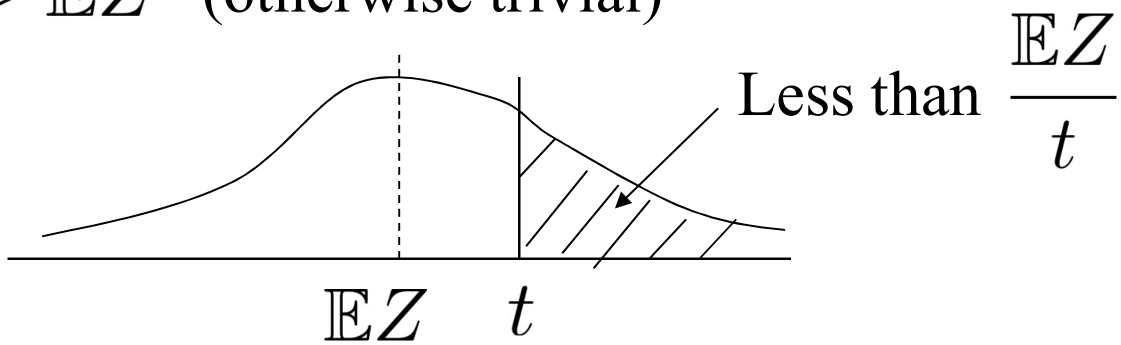
$$\begin{aligned}
\mathbb{E}Z &= \int_0^\infty ZP(Z)dZ \\
&= \int_0^t ZP(Z)dZ + \int_t^\infty ZP(Z)dZ \\
&\geq \int_t^\infty ZP(Z)dZ \quad \geq \int_t^\infty tP(Z)dZ \\
&= t\int_t^\infty P(Z)dZ \;=\; tP(Z > t)
\end{aligned}$$

Markov's Inequality

- If $Z \geq 0$

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

$t > \mathbb{E}Z$ (otherwise trivial)

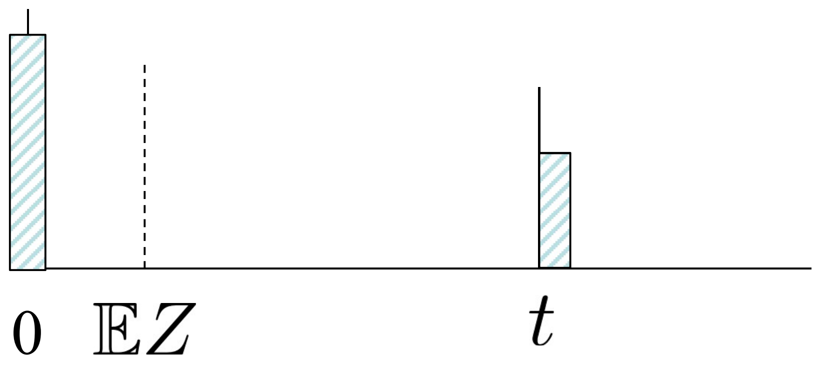


Markov's Inequality

- If $Z \geq 0$

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

- The bound is tight when



Chebyshev's Inequality

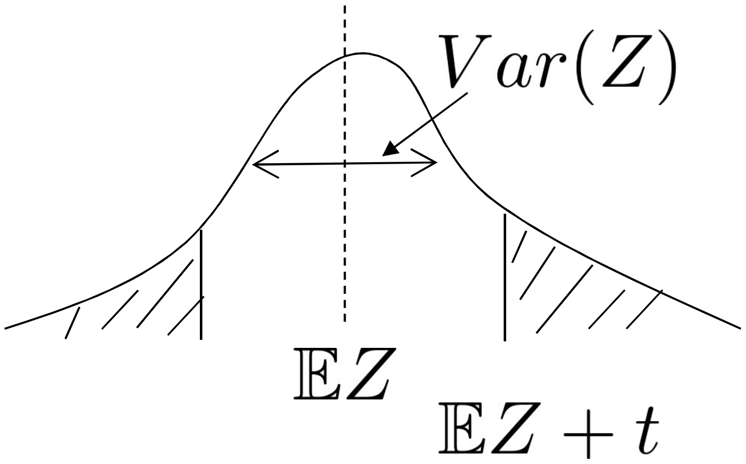
$$Var(Z) = \mathbb{E}(Z - \mathbb{E}Z)^2$$

$$P(|Z - \mathbb{E}Z| > t) = P((Z - \mathbb{E}Z)^2 > t^2) \leq \frac{Var(Z)}{t^2}$$

Chebyshev's Inequality

$$Var(Z) = \mathbb{E}(Z - \mathbb{E}Z)^2$$

$$P(|Z - \mathbb{E}Z| > t) = P((Z - \mathbb{E}Z)^2 > t^2) \leq \frac{Var(Z)}{t^2}$$



Meaningless if

$t^2 < Var(Z)$

Example of Concentration (Chebyshev)

- If we are interested in

$$Z = \frac{1}{N} \sum_{i=1}^{N} X_i$$

with independent R.V.s $X_1, \ldots, X_N$

Example of Concentration (Chebyshev)

- If we are interested in

$$Z = \frac{1}{N}\sum_{i=1}^{N} X_i$$

with independent R.V.s $X_1, \ldots, X_N$

$$Var(\sum X_i) = \sum_{i=1}^{N} Var(X_i) = NVar(X_1)$$

$$P\left(\left|\sum_{i=1}^{N} X_i - N\mathbb{E}X_1\right| > t\right) \leq \frac{NVar(X_1)}{t^2}$$

Example of Concentration (Chebyshev)

- We are interested in

$$Z = \frac{1}{N}\sum_{i=1}^{N} X_i$$

with independent R.V.s $X_1, \ldots, X_N$

- Equivalently,

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}X_1\right| > \frac{t}{\sqrt{N}}\right) \leq \frac{Var(X_1)}{t^2}$$

$$P\left(\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}X_1\right| > t\right) \leq \frac{Var(X_1)}{Nt^2}$$

Chernoff Bounds

- Motivation:
  - Central limit theorem:

$$X \sim \mathcal{N}(0, \sigma^2)$$

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right) \rightsquigarrow X$$

Chernoff Bounds

- Motivation:
  - Central limit theorem:

$$\lim_{N\to\infty} P\left\{ \frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}X_i > \frac{t}{\sqrt{N}} \right\} = 1 - \Psi(t/\sqrt{Var(X_1)})$$

$$\Psi(\mathbf{x}) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mathbf{x}^2} d\mathbf{x}$$

Chernoff Bounds

- If we use

$$\int_a^\infty e^{-\frac{\mathbf{x}^2}{2}} d\mathbf{x} \le \frac{e^{-\frac{a^2}{2}}}{a}$$

we obtain

$$\lim_{N\to\infty} P\left\{\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}X_i > \frac{t}{\sqrt{N}}\right\} \le \frac{\sqrt{Var(X_1)}}{\sqrt{2\pi}t} e^{-\frac{t^2}{2Var(X_1)}}$$

$$\le e^{-\frac{t^2}{2Var(X_1)}}$$

$$\left(t \quad \text{for} \quad \frac{\sqrt{Var(X_1)}}{t} < 1\right)$$

Chernoff Bounds

- Chernoff Bounds

$$P\{Z - \mathbb{E}Z > t\} = P\left\{e^{\lambda(Z-\mathbb{E}Z)} > e^{\lambda t}\right\} \qquad \lambda > 0$$

$$\leq \frac{\mathbb{E}e^{\lambda(Z-\mathbb{E}Z)}}{e^{\lambda t}}$$

Example of Concentration (Chernoff)

$$Z = \frac{1}{N}\sum X_i$$

$$\mathbb{E}e^{\lambda \sum X_i} = \mathbb{E}\prod_{i=1}^{N} e^{\lambda X_i} = \prod_{i=1}^{N}\mathbb{E}e^{\lambda X_i}$$

Independence

From Hoeffding's inequality,

$$X_1, \ldots, X_N \in [0,1] \qquad \mathbb{E}e^{\lambda(X_i - \mathbb{E}X_i)} \leq e^{\lambda^2/8}$$

$$\frac{\mathbb{E}e^{\lambda(X_i - \mathbb{E}X_i)}}{e^{\lambda t}} \leq \frac{e^{\lambda^2/8}}{e^{\lambda t}} = e^{\frac{\lambda^2}{8} - \lambda t}$$

$$\min_{\lambda} e^{\frac{\lambda^2}{8} - \lambda t} = e^{-2t^2}$$

Example of Concentration (Chernoff)

$$P\left\{\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right]\right| > t\right\} \leq 2e^{-2nt^2}$$

Martingale and Efron-Stein Inequality

$$\mathbb{E}_i[Z] = \mathbb{E}[Z|X_1, \ldots, X_i]$$

$$\mathbb{E}_0[Z] = \mathbb{E}[Z]$$

$$Z = f(X_1, \ldots, X_N)$$

Ex)

$$Z = X_1 + \cdots + X_N$$

For $\mathbb{E}_i[Z]$

$$Z = \underbrace{X_1 + \cdots + X_i}_{\text{fixed}} + \underbrace{X_{i+1} + \cdots + X_N}_{\text{R.V.}}$$

Martingale and Efron-Stein Inequality

$$\mathbb{E}_i[Z] = \int f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_i) \, d\mathbf{x}_{i+1} \ldots d\mathbf{x}_N$$

$$= g_i(\mathbf{x}_1, \ldots, \mathbf{x}_i)$$

$\rightarrow$ function of $\mathbf{x}_1, \ldots, \mathbf{x}_i$

Martingale and Efron-Stein Inequality

- Excercise

$$j > i, \qquad \mathbb{E}_i \mathbb{E}_j Z = \mathbb{E}_i Z$$

$$\mathbb{E}_i \mathbb{E}_j Z = \mathbb{E}_i \underline{\int f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_{j+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_j) d\mathbf{x}_{j+1} \ldots d\mathbf{x}_N}$$

$$\rightarrow \text{function of } \mathbf{x}_1, \ldots, \mathbf{x}_j$$

$$= \mathbb{E}_i g_j(\mathbf{x}_1, \ldots, \mathbf{x}_j)$$

$$= \int g_j(\mathbf{x}_1, \ldots, \mathbf{x}_j) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_N$$

$$= \int g_j(\mathbf{x}_1, \ldots, \mathbf{x}_j) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_j | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_j$$

Martingale and Efron-Stein Inequality

$$= \iint f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_{j+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_j) d\mathbf{x}_{j+1} \ldots d\mathbf{x}_N$$
$$p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_j | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_j$$

$$= \int f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_{j+1}, \mathbf{x}_j, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_i)$$
$$d\mathbf{x}_{i+1} \ldots d\mathbf{x}_j d\mathbf{x}_{j+1} \ldots d\mathbf{x}_N$$

$$= \mathbb{E}_i Z$$

Martingale and Efron-Stein Inequality

- Doob (Joseph Leo Doob) martingale representation

$$\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$$

$$\begin{aligned}
\sum_{i=1}^{N} \Delta_i &= \mathbb{E}_N Z - \cancel{\mathbb{E}_{N-1} Z} \\
&\quad + \cancel{\mathbb{E}_{N-1} Z} - \cancel{\mathbb{E}_{N-2} Z} + \cdots \\
&\quad + \cancel{\mathbb{E}_1 Z} - \mathbb{E} Z \\
&= \mathbb{E}_N Z - \mathbb{E} Z \\
&\boxed{= Z - \mathbb{E} Z}
\end{aligned}$$

Martingale and Efron-Stein Inequality

$$
\begin{aligned}
Var(Z) &= \mathbb{E}\left[(Z - \mathbb{E}Z)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N} \Delta_i\right)^2\right] \\
&= \sum_{i=1}^{N} \mathbb{E}[\Delta_i^2] \; + \; 2 \sum_{i,j;j>i} \mathbb{E}[\Delta_i \Delta_j]
\end{aligned}
$$

$\mathbb{E}[\Delta_i \Delta_j]$?

Martingale and Efron-Stein Inequality

$$\mathbb{E}[\Delta_i\Delta_j] = \int \Delta_i\Delta_j p(\mathbf{x}_1,\ldots,\mathbf{x}_N)d\mathbf{x}_1\ldots d\mathbf{x}_N$$

$$= \iint \Delta_i\Delta_j p(\mathbf{x}_{i+1},\ldots,\mathbf{x}_N|\mathbf{x}_1,\ldots,\mathbf{x}_i)d\mathbf{x}_{i+1}\ldots d\mathbf{x}_N \; p(\mathbf{x}_1,\ldots,\mathbf{x}_i)d\mathbf{x}_1\ldots d\mathbf{x}_i$$

$$= \int \underline{\mathbb{E}_i[\Delta_i\Delta_j]}p(\mathbf{x}_1,\ldots,\mathbf{x}_i)d\mathbf{x}_1\ldots d\mathbf{x}_i$$

$$\mathbb{E}_i[\Delta_i\Delta_j] =$$

$$\int \underbrace{g_i(\mathbf{x}_1,\ldots,\mathbf{x}_i)}_{\text{function of } \mathbf{x}_1,\ldots,\mathbf{x}_i} \underbrace{g_j(\mathbf{x}_1,\ldots,\mathbf{x}_j)}_{\text{function of } \mathbf{x}_1,\ldots,\mathbf{x}_j} p(\mathbf{x}_{i+1},\ldots,\mathbf{x}_N|\mathbf{x}_1,\ldots,\mathbf{x}_i)d\mathbf{x}_{i+1}\ldots d\mathbf{x}_N$$

$$= g_i(\mathbf{x}_1,\ldots,\mathbf{x}_i)\int g_j(\mathbf{x}_1,\ldots,\mathbf{x}_j)\, p(\mathbf{x}_{i+1},\ldots,\mathbf{x}_N|\mathbf{x}_1,\ldots,\mathbf{x}_i)d\mathbf{x}_{i+1}\ldots d\mathbf{x}_N$$

$$= \Delta_i\mathbb{E}_i[\Delta_j]$$

Martingale and Efron-Stein Inequality

$j > i,$

$$\mathbb{E}_i \Delta_j = \mathbb{E}_i(\mathbb{E}_j Z - \mathbb{E}_{j-1} Z)$$

$$= \mathbb{E}_i \left(\mathbb{E} f(\mathbf{x}_{j+1}, \ldots, \mathbf{x}_N) - \mathbb{E} f(\mathbf{x}_j, \ldots, \mathbf{x}_N) \right)$$

$$= \int (g_j(\mathbf{x}_1, \ldots, \mathbf{x}_j) - g_{j-1}(\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}))$$

$$p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_N$$

$$= \int g_j(\mathbf{x}_1, \ldots, \mathbf{x}_j) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_j | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_j$$

$$- \int g_{j-1}(\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_{j-1} | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_{j-1}$$

Martingale and Efron-Stein Inequality

$$= \iint f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_{j+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_j)$$
$$p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_j | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_j d\mathbf{x}_{j+1} \ldots d\mathbf{x}_N$$
$$- \iint f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_j, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_{j-1})$$
$$p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_{j-1} | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_{j-1} d\mathbf{x}_j \ldots d\mathbf{x}_N$$

$$= \int f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_N$$
$$- \int f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_{i+1}, \ldots, \mathbf{x}_N | \mathbf{x}_1, \ldots, \mathbf{x}_i) d\mathbf{x}_{i+1} \ldots d\mathbf{x}_N$$

$$= 0$$

Martingale and Efron-Stein Inequality

$$Var(Z) = \sum_{i=1}^{N} \mathbb{E}[\Delta_i^2]$$

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

$$\mathbb{E}^{(i)} Z = \int f(\mathbf{x}_1, \ldots, \mathbf{x}_N) p(\mathbf{x}_i | \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N) d\mathbf{x}_i$$

Show that

$$\mathbb{E}_i \mathbb{E}^{(i)} Z = \mathbb{E}_{i-1} Z$$

Martingale and Efron-Stein Inequality

$$\Delta_i \;=\; \mathbb{E}_i[Z - \mathbb{E}^{(i)} Z]$$

From Jensen's inequality
(Square of expectation vs.
Expectation of square),

$$\Delta_i^2 \;\leq\; \mathbb{E}_i \left[(Z - \mathbb{E}^{(i)} Z)^2 \right]$$

$\Bigg($ Note

$$Var(Z) = \sum_{i=1}^{N} \mathbb{E}[\Delta_i^2]$$

$\Bigg)$

Martingale and Efron-Stein Inequality

Show that $\mathbb{E}\left[\mathbb{E}_i\left[a\right]\right] = \mathbb{E}[a]$.

$$\Rightarrow \mathbb{E}[\Delta_i^2] \leq \mathbb{E}\left[\mathbb{E}_i[(Z - \mathbb{E}^{(i)}Z)^2]\right]$$

$$= \underline{\mathbb{E}\left[(Z - \mathbb{E}^{(i)}Z)^2\right]}$$

Martingale and Efron-Stein Inequality

$$Var^{(i)}Z \equiv \mathbb{E}^{(i)}\left[(Z - \mathbb{E}^{(i)}Z)^2\right]$$

Conditional variance operator conditioned on

$$X^{(i)} \quad (= X_1, , \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)$$

$$\mathbb{E}[Var^{(i)}Z] = \mathbb{E}\left[\mathbb{E}^{(i)}\left[(Z - \mathbb{E}^{(i)}Z)^2\right]\right]$$

$$= \underline{\mathbb{E}\left[(Z - \mathbb{E}^{(i)}Z)^2\right]}$$

(from $\mathbb{E}\left[\mathbb{E}^{(i)}[a]\right] = \mathbb{E}[a]$)

Also show this!!

Martingale and Efron-Stein Inequality

$$\begin{aligned}\mathbb{E}[\Delta_i^2] &= \mathbb{E}[(Z - \mathbb{E}^{(i)}Z)^2] \\ &= \mathbb{E}[Var^{(i)}Z]\end{aligned}$$

$$\boxed{\begin{aligned}Var(Z) &\leq \sum_{i=1}^{N} \mathbb{E}\left[(Z - \mathbb{E}^{(i)}Z)^2\right] \\ &= \sum_{i=1}^{N} \mathbb{E}\left[Var^{(i)}(Z)\right]\end{aligned}}$$

Martingale and Efron-Stein Inequality

- How can we use the inequality?

① $$Var(Z) \leq \frac{1}{2}\sum_{i=1}^{N}\mathbb{E}\left[(Z - Z_i')^2\right]$$

Z_i' : independent copy of Z conditionally on $X^{(i)}$

We use $$Var^{(i)}(Z) = \frac{1}{2}\mathbb{E}^{(i)}\left[(Z - Z_i')^2\right]$$

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

Independent realizations $X, Y \sim P$

Martingale and Efron-Stein Inequality

- How can we use the inequality?

② $$Var(Z) \le \sum_{i=1}^{N} \inf_{Z_i} \mathbb{E}\left[(Z - Z_i)^2\right]$$

We use

$$Var^{(i)}(Z) = \inf_{Z_i} \mathbb{E}^{(i)}\left[(Z - Z_i)^2\right]$$

Conditioned on $X^{(i)}$

Functions with Bounded Differences

$$f: \mathcal{X}^N \to \mathbb{R}$$

$$\sup_{\mathbf{x}_1,\ldots,\mathbf{x}_N,\mathbf{x}_i' \in \mathcal{X}} |f(\mathbf{x}_1,\ldots,\mathbf{x}_N) - f(\mathbf{x}_1,\ldots,\mathbf{x}_{i-1},\mathbf{x}_i',\mathbf{x}_{i+1},\ldots,\mathbf{x}_N)| \le c_i$$

for some nonnegtive constants $c_1, \ldots, c_N$

(= If we change the *i*-th variable of *f* while keeping all the others fixed, the value of the function cannot change by more than c_i.)

$$Var(Z) \le \frac{1}{4} \sum_{i=1}^{N} c_i^2$$

Functions with Bounded Differences

$$Var(Z) \le \frac{1}{4}\sum_{i=1}^{N} c_i^2$$

Proof:

$$Var(Z) \le \sum_{i=1}^{N} \inf_{Z_i} \mathbb{E}\left[(Z - Z_i)^2\right]$$

If we let

$$Z_i = \frac{1}{2}\left(\sup_{\mathbf{x}_i' \in \mathcal{X}} f(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i', \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N) + \inf_{\mathbf{x}_i' \in \mathcal{X}} f(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i', \mathbf{x}_{i+1}, \ldots, \mathbf{x}_N)\right)$$

$$(Z - Z_i)^2 \le \left(\frac{c_i}{2}\right)^2 = \frac{c_i^2}{4}$$

Middle between max and min

Kernel Density Estimation

$$\phi_N(\mathbf{x}) = \frac{1}{N h_N} \sum_{i=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right) \qquad \int K = 1$$

$$Z(N) = f(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \int |\phi(\mathbf{x}) - \phi_N(\mathbf{x})| d\mathbf{x} \qquad \text{L}_1 \text{ error}$$

$$\begin{aligned} &|f(\mathbf{x}_1, \ldots, \mathbf{x}_N) - f(\mathbf{x}_1, \ldots, \mathbf{x}_i', \ldots, \mathbf{x}_N)| \\ &\qquad \le \frac{1}{N h_N} \int \left| K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right) - K\left(\frac{\mathbf{x}-\mathbf{x}_i'}{h_N}\right) \right| d\mathbf{x} \quad \le \frac{2}{N} \end{aligned}$$

$$Var(Z(N)) \le \frac{1}{N}$$

Kernel Density Estimation

$$\phi_N(\mathbf{x}) = \frac{1}{Nh_N} \sum_{i=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_N}\right) \qquad \int K = 1$$

By Chebyshev's inequality

$$P\left\{\left|\frac{Z(N)}{\mathbb{E}Z(N)} - 1\right| \geq \epsilon\right\} = P\Big\{|Z(N) - \mathbb{E}Z(N)| \geq \epsilon\mathbb{E}Z(N)\Big\}$$

$$\leq \frac{Var(Z(N))}{\epsilon^2(\mathbb{E}Z(N))^2} \leq \frac{1}{N\epsilon^2(\mathbb{E}Z(N))^2} \longrightarrow 0$$

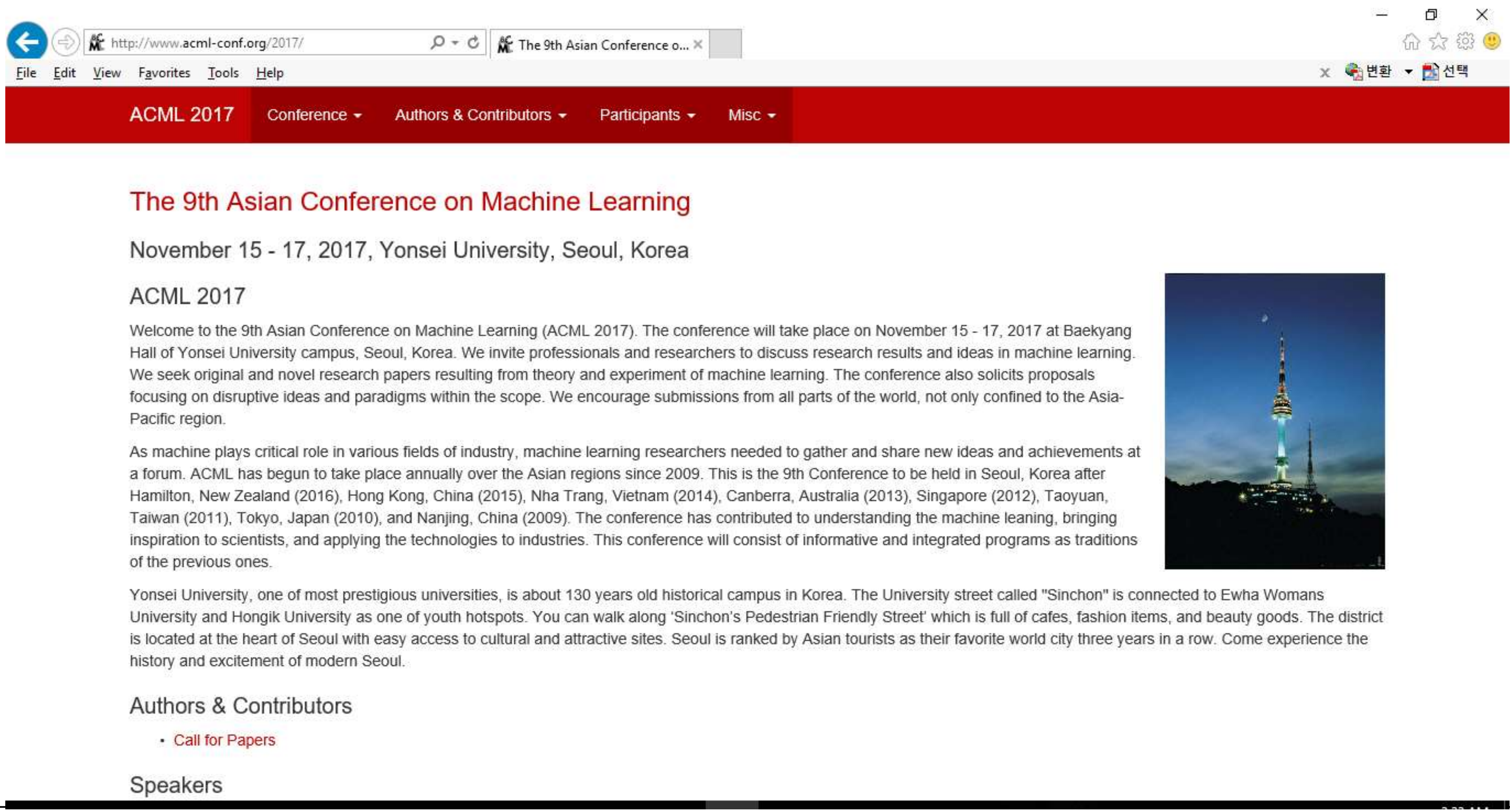
Bounded Difference

- Bounded difference extends to Rademacher average bounding and McDiarmid inequality.

- Foundations of learning theory

Asian Conference on Machine Learning (ACML) in Seoul

- http://www.acml-conf.org/2017/
- Nov. 15 - 17 (Wed. – Fri.), 2017



THANK YOU

Yung-Kyun Noh
nohyung@snu.ac.kr