Midterm

Law of total Expectation: $\mathbb{E}(Y) = \mathbb{E}_X[\mathbb{E}_Y(Y|X)]$

Law of total Variance:

$$\mathrm{Var}(Y) = \mathrm{Var}_X[\mathbb{E}_Y(Y|X)] + \mathbb{E}_X[\mathrm{Var}_Y(Y|X)]$$

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

$$= \mathbb{E}_X[\mathbb{E}_Y(Y|X)^2] - (\mathbb{E}_X\mathbb{E}_Y(Y|X))^2 + \mathbb{E}_X[\mathrm{Var}_Y(Y|X)]$$

$$= \mathbb{E}_X[\cancel{\mathbb{E}_Y(Y|X)^2}] - (\mathbb{E}Y)^2 + \underbrace{\mathbb{E}_X[\mathrm{Var}_Y(Y|X)]}$$

$$\downarrow$$

$$\underbrace{\mathbb{E}_X[\mathbb{E}_Y[Y^2|X] - \mathbb{E}_Y(Y|X)^2]}$$

$$\downarrow$$

$$\mathbb{E}Y^2 - \cancel{\mathbb{E}_X[\mathbb{E}_Y(Y|X)^2]}$$

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

Moment generating function:

$$M_X(t) = \mathbb{E}[\exp(Xt)] = \mathbb{E}[e^{Xt}]$$

$$M_X^n(t)\Big|_{t=0} = \mathbb{E}(X^n)$$

Markov Inequality:

R.V. $X \geq 0$ $\quad \mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$

proof: $\int_0^\infty f(x)x\,dx = \int_0^t f(x)x\,dx + \int_t^\infty f(x)x\,dx$

$\mathbb{E}[X] \geq t\,\mathbb{P}(X \geq t)$

$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$

Chebyshev Inequality:

$$\mathbb{P}(|X - \mathbb{E}[X]| > k\sigma) \leq \frac{1}{k^2}$$

$\hat{\mu}_n = \frac{1}{n}\sum^n X_i \quad X_i \sim N(\mu, \sigma^2)$

$\hat{\mu}_n \sim N(\mu, \frac{\sigma^2}{n})$

[illegible]

$$\mathbb{P}(|\hat{\mu}_n - \mu| > \frac{k\sigma}{\sqrt{n}}) \le \frac{1}{k^2}$$

proof: $\mathbb{P}(|X-\mathbb{E}[X]| > k\sigma) = \mathbb{P}(|X-\mathbb{E}[X]|^2 > k^2\sigma^2)$

$$= \frac{\mathbb{E}(X-\mathbb{E}[X])^2}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

Chernoff bound:

when mgf exists in a neighborhood around 0. $\Rightarrow$ mgf is finite, when $0<t<b$

$$\mathbb{P}(|X-\mu|>u) = \mathbb{P}(\exp(t(X-\mu)) > \exp(tu)) < \frac{\mathbb{E}[\exp(t(X-\mu))]}{\exp(tu)}$$

$$\mathbb{P}(|X-\mu|>u) < \inf_{0\le t\le b} \frac{\mathbb{E}[\exp(tX)]}{\exp(t\mu+tu)}$$

Gaussian tail bound via Chernoff

$X \sim N(\mu,\sigma^2)$, then $M(Xt) = \mathbb{E}[\exp(Xt)] = \exp(t\mu + t^2\sigma^2/2)$

$$\mathbb{P}(X-\mu>u) \le \inf_{0\le t} \frac{\exp(t\mu+t^2\sigma^2/2)}{\exp(t\mu+tu)} = \inf_{0\le t} \exp(t^2\sigma^2/2 - tu)$$

$t = \frac{u}{\sigma^2}$ $\qquad \frac{1}{2}(t\sigma - \frac{u}{\sigma})^2 - \frac{1}{2}\frac{u^2}{\sigma^2}$

$$\le \exp(-\frac{1}{2}\frac{u^2}{\sigma^2})$$

$$\mathbb{P}(-X+\mu>u) \le \exp(-\frac{1}{2}\frac{u^2}{\sigma})$$

$$\mathbb{P}(|X-\mu|>u) \le 2\exp(-\frac{1}{2}\frac{u^2}{\sigma^2})$$

$\hat{\mu}_n \sim N(\mu, \frac{\sigma^2}{n})$ $\qquad \mathbb{P}(|\hat{\mu}_n - \mu| > \frac{1}{\sqrt{n}}\sigma k) \le 2\exp(-\frac{k^2}{2})$

sub-Gaussian: $\mathbb{E}(t(X-\mu)) \le \exp(t^2\sigma^2/2)$ $\qquad$ ☆ for all t

$$\mathbb{P}(|X-\mu|>u) \le 2\exp(-\frac{1}{2}\frac{u^2}{2\sigma^2})$$

$$\mathbb{P}(|\hat{\mu}-\mu| > \frac{k\sigma}{\sqrt{n}}) \le 2\exp(-\frac{k^2}{2})$$

Bounded R.V. – Hoeffding's

proof: $\quad X'$ denote an independant copy of X, $X\in[a,b]$ zero mean

$$\mathbb{E}_X[\exp(tX)] = \mathbb{E}_X[\exp(t(X-\mathbb{E}[X']))] \le \mathbb{E}_{X,X'}[\exp(t(X-X'))]$$

using Jensen, and convexity of exp( )

* Rademacher R.V. ϵ (+1, −1) equiprobably. $\mathbb{E}\exp(tX) \le \exp(\frac{t^2\sigma^2}{2}) = \exp(\frac{t^2}{2})$

$$X - X' = X' - X = \epsilon(X - X')$$

$$\mathbb{E}_{X,X'}[\exp(t(X-X'))] = \mathbb{E}_{X,X'}[\mathbb{E}_\epsilon[\exp(t\epsilon(X-X'))]]$$

$$\le \mathbb{E}_{X,X'}[\exp(t^2(X-X')^2/2)]$$

$$\mathbb{E}_X[\exp(tX)] \le \exp(t^2(b-a)^2/2)$$

$$\mathbb{P}(|\frac{1}{n}\sum^n X_i - \mu| \ge t) \le \exp(-\frac{k^2}{2})$$

$$t = \frac{k(b-a)}{\sqrt{n}} \quad \le 2\exp(-\frac{t^2 n}{2(b-a)^2})$$

$$k = \frac{t\sqrt{n}}{b-a}$$

Berstein's inequality (refinement of Hoeffding)

$X_1, \dots, X_n$ μ, bounded support $[a,b]$, $\mathbb{E}[(X-\mu)^2] = \sigma^2$

$$\mathbb{P}(|\hat{\mu} - \mu| > t) \le 2\exp(-\frac{nt^2}{2(\sigma^2 + (b-a)t)})$$

McDiarmid's inequality (concentration of Lipshitz functions of iid R.V.)

R.V.s $X_1, \dots, X_n$, $f: \mathbb{R}^n \to \mathbb{R}$

bounded difference condition: $|f(X_1, \dots, X_n) - f(X_1, \dots, X_{k-1}, X_k', X_{k+1}, \dots X_n)| \le L_k$

for all $t \ge 0$ $\mathbb{P}(|f(X_1, \dots X_n) - \mathbb{E} f(X_1, \dots, X_n)| > t) \le 2\exp(-\frac{2t^2}{\sum_{k=1}^n L_k^2})$

directly implies Hoeffding: $f(X_1, \dots, X_n) = \frac{1}{n}\sum^n X_i$, $L_k = \frac{b-a}{n}$ $\quad \| \quad 2\exp(-\frac{2nt^2}{(b-a)^2})$

Levy's inequality

$$|f(X_1, \dots, X_n) - f(Y_1, \dots, Y_n)| \le L\sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

if $X_1, \dots, X_n \sim N(0,1)$

$$\mathbb{P}(\ |f(X_1, \dots X_n) - \mathbb{E} f(X_1, \dots X_n)| \ge t) \le \dots \frac{t^2}{\ }$$

$\mathbb{P}(|f(x_1,\dots,x_n) - \mathbb{E} f(x_1,\dots,x_n)| > t) \le 2\exp(-\frac{\quad}{2(\ )})$

χ^2 tail bound

$Z_1, \dots, Z_n \sim N(0,1)$ $\quad \mathbb{P}(|\frac{1}{n}\sum^n Z_i^2 - 1| \geq t) \leq 2\exp(-nt^2/8)$

$\mathbb{E}[Z_i^2] = 1$ $\quad$ for all $t \in (0,1)$

χ^2 is sub-exponential RVs. tail bound only holds for small deviation t

Johnson-Lindenstrauss Lemma

$x_1, \dots, x_n \in \mathbb{R}^d$ create a map $F: \mathbb{R}^d \to \mathbb{R}^m$, $m \ll d$.

$$(1-\epsilon)\|x_i - x_j\|_2^2 \leq \|F(x_i) - F(x_j)\|_2^2 \leq (1+\epsilon)\|x_i - x_j\|_2^2$$

$$m \geq \frac{16\log(n/\delta)}{\epsilon^2}$$

$$F(x_i) = \frac{Z x_i}{\sqrt{m}}$$

$Z: \mathbb{R}^{m\times d}$, where each entry of Z is iid $N(0,1)$

Asymptotic Convergence

Convergence in probability: $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$

$\hat{\theta}_n \xrightarrow{P} \theta$ is consistency

WLLN: $\lim_{n\to\infty} \mathbb{P}(|\frac{1}{n}\sum^n X_i - \mathbb{E}[X]| \geq \epsilon) = 0$, $\quad Var(X_i) = \sigma^2 < \infty$

or first absolute moment finite

Convergence in quadratic mean

$$\mathbb{E}(X_n - X)^2 \to 0 \text{ as } n \to \infty$$

Convergence in distribution

$\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$ for all points t, where CDF F_X is continuous

$qm \xrightarrow{(1)} p \xrightarrow{(2)} d$

(1) $\lim \mathbb{P}(|X_n - X| > \epsilon) \leq \frac{\mathbb{E}(X_n - X)^2}{\quad} \to 0$

$n\to\infty$ ϵ^2

(2) $F_{X_n}(x) = \mathbb{P}(X_n \le x) = \mathbb{P}(X_n \le x, X \le x+\epsilon) + \mathbb{P}(X_n \le x, X \ge x+\epsilon)$

$\le \mathbb{P}(X \le x+\epsilon) + \mathbb{P}(|X - X_n| \ge \epsilon)$

$= F_X(x+\epsilon) + \mathbb{P}(|X - X_n| \ge \epsilon)$

$F_X(x-\epsilon) = \mathbb{P}(X \le x-\epsilon) = \mathbb{P}(X \le x-\epsilon, X_n \le x) + \mathbb{P}(X \le x-\epsilon, X_n \ge x)$

$\le F_{X_n}(x) + \mathbb{P}(|X - X_n| \ge \epsilon)$

$F_X(x-\epsilon) - \mathbb{P}(|X-X_n| \ge \epsilon) \le F_{X_n}(x) \le F_X(x+\epsilon) + \mathbb{P}(|X-X_n| \ge \epsilon)$

$\mathbb{P} \to 0$ as $n \to \infty$

$\epsilon \to 0$, use continuity of $F_X(x)$ at x $F_{X_n}(x) \to F_X(x)$

$$F_X(x-\epsilon) \le \liminf_{n\to\infty} F_{X_n}(x) \le F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x) \le F_X(x+\epsilon)$$

$d \not\Rightarrow p$, except, X is deterministic

$\mathbb{P}(|X_n - c| \ge \epsilon) = \mathbb{P}(X_n \ge c+\epsilon) + \mathbb{P}(X_n \le c-\epsilon)$

$= F_{X_n}(c-\epsilon) + 1 - F_{X_n}(c+\epsilon)$

$n \to \infty$ $= F_X(c-\epsilon) + 1 - F_X(c+\epsilon) = 0 + 1 - 1 = 0$

Continuous mapping theorem:

$X_1, \cdots, X_n \xrightarrow{P} X$

$h(X_1), \dots, h(X_n) \xrightarrow{P} h(X)$

also true for convergence in distribution.

Slutsky's theorem:

$Y_n \xrightarrow{d} c$, $X_n \to X$ then

$X_n + Y_n \to X + c$

$X_n Y_n \to cX$

Stochastic order notation: $a_n = o(1)$ if $a_n \to 0$, as $n \to \infty$

$a_n = O(1)$ if $|a_n| \le C$ for constant $C \ge 0$

$a_n = O(b_n)$ if $a_n/b_n = O(1)$

$\hat{\mu} - \mu = o_p(1)$ (WLLN)

$\hat{\mu} - \mu = O_p(1/\sqrt{n})$ (CLT)

Central Limit Theorem

$X_1, \dots, X_n$, iid μ, σ^2, $\mathbb{E}[\exp(tX_i)]$ finite for t in a neighborhood near zero

$$S_n = \frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma} \xrightarrow{d} N(0,1)$$

most general, finite variance

Fact: (1) $M_Z(t) = M_X(t) M_Y(t)$,

X, Y independent with M_X, M_Y, then $Z = X+Y$

(2) $Y = a+bX$

$M_Y(t) = \exp(at) + M_X(bt)$

(3) ☆ If for all t in an open interval around 0 we have that,

$M_{X_n}(t) \to M_X(t)$, then $X_n \xrightarrow{d} X$

Proof: mgf of a standard gaussian is $M_Z(t) = \exp(t^2/2)$

$$M_{S_n}(t) = \left[M_{(X-\mu)}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \quad \text{using fact (1) (2)}$$

$$S_n = \frac{\sqrt{n}\left(\frac{1}{n}\sum^n X_i - \mu\right)}{\sigma} = \sum_{i=1}^n \frac{1}{\sqrt{n}\sigma}(X_i - \mu)$$

imagine t is small, close to zero

☆ Taylor expanding:

$$M_{S_n}(t) = \left[1 + \frac{t}{\sigma\sqrt{n}}\mathbb{E}(X-\mu) + \frac{t^2}{2\sigma^2 n}\mathbb{E}(X-\mu)^2 + \frac{t^3}{6n^{3/2}\sigma^3}\mathbb{E}(X-\mu)^3 + \dots\right]^n$$

$$\approx \left[1 + \frac{t^2}{2n}\right]^n \to \exp(t^2/2),$$

using the fact that $\lim_{n\to\infty}(1+x/n)^n \to \exp(x)$

Lyapunov CLT: $X_1, \dots, X_n$ independent but not necessarily identically dist

$\mu_i = \mathbb{E}[X_i]$, $\sigma_i^2 = Var(X_i)$

Lyapnov Condition: $\lim_{n\to\infty} \frac{1}{S_n^3}\sum_{i=1}^n \mathbb{E}|X_i - \mu_i|^3 = 0$, $S_n^2 = \sum_{i=1}^n \sigma_i^2$

then:

$$\frac{1}{S_n}\sum_{i=1}^{n}(X_i-\mu_i) \Rightarrow N(0,1)$$

third moment $\sum_{i=1}^{n}\mathbb{E}|X_i-\mu|^3 \le Cn$

$$S_n^2 = \sum \sigma_i^2 \ge n\sigma_{min}^2$$

Lyapunov ratio : $\dfrac{nC}{\sqrt{n^{3}}^{2}\,\sigma_{min}^3} = \dfrac{C}{\sqrt{n}\,\sigma_{min}^3} \to 0$

Multivariate CLT:

$X_1, \dots, X_n$ iid $\mu \in \mathbb{R}^d$ covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ with finite entries

$$\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{d} N(0,\Sigma)$$

CLT with estimated variance

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum^{n}(X_i-\bar{x})^2$$

$$\frac{\sqrt{n}(\hat{\mu}-\mu)}{\hat{\sigma}_n} \xrightarrow{d} N(0,1)$$

$$\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma}\cdot\frac{\sigma}{\hat{\sigma}_n} \xrightarrow{d} N(0,1)\cdot 1$$

with slutsky's if $\dfrac{\sigma}{\hat{\sigma}_n} \xrightarrow{d} 1$

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum^{n}(X_i-\bar{x})^2 \xrightarrow{P} \frac{1}{n}\sum^{n}(X_i-\bar{x})^2 \xrightarrow{P} \mathbb{E}(X-\bar{X})^2 = \sigma^2$$

Rate of convergence in CLT Berry-Essen

$$\sup_x |F_n(x)-\Phi(x)| \le \frac{9\mu_3}{\sigma^3\sqrt{n}} \qquad \mu_3 = \mathbb{E}[|X_1-\mu|^3]$$

$$F_n(x) = \mathbb{P}\left(\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sqrt{\sigma}} \le x\right) \qquad \sigma^2 = \mathbb{E}[(X_1-\mu)^2]$$

Delta method:

$\dfrac{\sqrt{n}(X_n-\mu)}{\sigma} \xrightarrow{d} N(0,1)$, g is continuously differentiable $g'(\mu)\neq 0$

$$\frac{\sqrt{n}(g(X_n)-g(\mu))}{\sigma} \xrightarrow{d} N(0, g'(\mu)^2)$$

$$g(X_n) = g(\mu) + g'(\mu)(X_n-\mu)$$

$$\frac{\sqrt{n}\,(g(X_n)-g(\mu))}{\sigma} \approx \frac{\sqrt{n}\,(g'(\mu)(X_n-\mu))}{\sigma} \xrightarrow{d} N(0, g'(\mu)^2)$$

Uniform Laws of Large Numbers

$$\Delta = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_X(x)|$$

Glivenko-Cantelli: $\Delta \xrightarrow{P} 0$

$$\Delta(\mathcal{A}) = \sup_{A\in\mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)|$$ Vapnik-Cervonenkis theory

$$\Delta(\mathcal{F}) = \sup_{F\in\mathcal{F}} \left|\frac{1}{n}\sum^{n} F(X_i) - \mathbb{E}[F]\right|.$$ empirical process

$$\hat{R}_n(f) = \frac{1}{n}\sum^{n} \mathbb{1}(f(X_i) \neq y_i)$$ Binary classification

$$\mathbb{P}(|\hat{R}_n(f) - \mathbb{P}(f(X)\neq y)| \geq t) \leq 2\exp(-2nt^2)$$

$$\underbrace{\hat{f} = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\ \hat{R}_n(f)}$$ Empirical risk minimization

Excess risk of the chosen classifier

$$\Delta = \mathbb{P}(\hat{f}(X)\neq y) - \mathbb{P}(f^*(X)\neq y)$$

$$= \underbrace{\mathbb{P}(\hat{f}(X)\neq y) - \hat{R}_n(\hat{f})} + \underbrace{\hat{R}_n(\hat{f}) - \hat{R}_n(f^*)} + \underbrace{\hat{R}_n(f^*) - \mathbb{P}(f^*(x)\neq y)}$$

$\hat{R}_n(\hat{f})$ is not sum of iid. can't use Hoeffding

$\downarrow$

uniform convergence bound

$T_2 \leq 0$ because $\hat{f}$ minimise empirical risk $\hat{R}_n$

Hoeffding $T_3 \leq \sqrt{\frac{2\log 2/\ldots}{n}}$

Shattering: the max of # different subsets of n points that can be picked out by the collection $\mathcal{A}$

$$N_{\mathcal{A}}(z_1,\dots,z_n) = \{\{z_1,\dots,z_n\}\cap A : A\in\mathcal{A}\} \leq 2^n$$

$$s(\mathcal{A}, n) = \max_{\{z_1, \dots, z_n\}} N_{\mathcal{A}}(z_1, \dots, z_n)$$

VC Theorem:

$$\mathbb{P}(\Delta(\mathcal{A}) \geq t) \leq 8 s(\mathcal{A}, n) \exp(-nt^2/32)$$

Vc dimension: largest d. for which $s(A, d) = 2^d$

Sauer's Lemma:

Empirical Rademacher complexity: $\mathcal{R}(X_1, \dots, X_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right]$

Rademacher complexity: $\mathcal{R}(\mathcal{F}) = \mathbb{E}_\epsilon \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \right]$

Rademacher Theorem: $\mathbb{E}[\Delta(\mathcal{F})] \leq 2\mathcal{R}(\mathcal{F})$