Theoretical Stat/Machine Learning (Part-I)

1. Linear Algebra

Cauchy–Schwarz: (special case of Hölder)

$$|\langle x, y\rangle| \leqslant \|x\|_2\|y\|_2$$

Hölder:

$$|\langle x, y\rangle| \leqslant \|x\|_p\|y\|_{p'}, \quad \frac{1}{p}+\frac{1}{p'}=1$$

Dual norm of ℓ_p is $\ell_{p'}$:

$$\|x\|_p = \sup_{\|y\|_{p'}=1} \langle x, y\rangle$$

Minkowski's inequality:

$$\|x+y\|_p \leqslant \|x\|_p + \|y\|_p$$

ℓ_p comparisons: For $1 \leqslant p \leqslant q \leqslant \infty$:

$$\|x\|_q \leqslant \|x\|_p \leqslant n^{1/p-1/q}\|x\|_q$$

Matrix basics:

- $\operatorname{tr}(AB) = \operatorname{tr}(BA)$
- $\langle A, B\rangle_F = \operatorname{tr}(A^\top B)$

Trace and determinant:

$$\operatorname{tr}(A) = \sum_i \lambda_i \qquad \det(A) = \prod_i \lambda_i$$

Extreme eigenvalues:

$$\lambda_1 = \max_{\|x\|_2=1} x^\top Ax \qquad \lambda_p = \min_{\|x\|_2=1} x^\top Ax$$

PSD through eigenvalues: $A \in \mathbb{S}^p$

$$A \succeq 0 \iff \lambda_i(A) \geqslant 0 \quad \forall\, i$$

Induced operator norm: $A \in \mathbb{R}^{m\times n}$

$$\|A\|_{p\to q} = \sup_{\|x\|_p\leqslant 1} \|Ax\|_q = \sup_{x\neq 0} \frac{\|Ax\|_q}{\|x\|_p}$$

Composition bound (Submultiplicativity):

$$\|AB\|_{p\to r} \leqslant \|A\|_{q\to r}\,\|B\|_{p\to q}$$

Spectral norm: $p = q = 2$

$$\|A\|_{\mathrm{op}} = \|A\|_{2\to 2} = \sup_{\|x\|_2=1}\|Ax\|_2 = \sigma_1(A) = \sqrt{\lambda_1(A^\top A)}$$

Frobenius norm:

$$\|A\|_{\mathrm{F}}^2 = \sum_{i,j} A_{ij}^2 = \operatorname{tr}(A^\top A) = \sum_{i=1}^{\operatorname{rank}(A)} \sigma_i(A)^2$$

Spectral vs Frobenius norm:

$$\|A\|_{\mathrm{op}} \leqslant \|A\|_{\mathrm{F}} \leqslant \sqrt{\operatorname{rank}(A)}\,\|A\|_{\mathrm{op}}$$

Frobenius submultiplicativity: A, B conformable matrices,

$$\|AB\|_{\mathrm{F}} \leqslant \|A\|_{\mathrm{op}}\,\|B\|_{\mathrm{F}} \leqslant \|A\|_{\mathrm{F}}\|B\|_{\mathrm{F}}$$

Rank one case: if $A = uv^\top$ then

$$\|A\|_{\mathrm{op}} = \|A\|_{\mathrm{F}} = \|u\|_2\,\|v\|_2$$

2. Probability Basics

Variance:

$$\operatorname{Var}(aX+bY) = a^2\operatorname{Var}(X) + b^2\operatorname{Var}(Y) + 2ab\operatorname{Cov}(X,Y)$$

L_p−norms:

$$\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p}, \qquad \|X\|_{L_\infty} = \operatorname{ess\,sup}|X|$$

Key inequalities

Hölder: p, p' are conjugate exponents,

$$|\mathbb{E}[XY]| \leqslant \|X\|_{L_p}\,\|Y\|_{L_{p'}}$$

Minkowski:

$$\|X+Y\|_{L_p} \leqslant \|X\|_{L_p} + \|Y\|_{L_p}$$

Jensen (convex ϕ):

$$\phi(\mathbb{E}X) \leqslant \mathbb{E}[\phi(X)]$$

Since **norms are convex**, $\|\mathbb{E}X\| \leqslant \mathbb{E}\,\|X\|$

Monotonicity of L_p−norms: for $1 \leqslant p \leqslant q \leqslant \infty$,

$$\|X\|_{L_p} \leqslant \|X\|_{L_q}$$

L_p vs ℓ_p norms: Let $x = (x_1, \ldots, x_n)$ and X be a random variable that takes values x_i each with prob. $1/n$ then,

$$\|X\|_{L_p} \leqslant 1 \iff \|x\|_{\ell_p} \leqslant n^{1/p}$$

Integrated tail identities:

$$\mathbb{E}|X|^p = \int_0^\infty pt^{p-1}\mathbb{P}\{|X|>t\}\,dt$$

For a non-decreasing $\Phi \geqslant 0$ with $\Phi(0) < \infty$,

$$\mathbb{E}[\Phi(|X|)] = \Phi(0) + \int_0^\infty \Phi'(t)\mathbb{P}\{|X|>t\}\,dt$$

Moment growth bound: Random variable X, assume $\forall\, t \geqslant 0$,

$$\mathbb{P}\{|X| \geqslant t\} \leqslant C\exp(-ct^\alpha),$$

then there exists a constant $C' = C'(C,\alpha)$ such that

$$\|X\|_{L_p} = (\mathbb{E}|X|^p)^{1/p} \leqslant C'(p/c)^{1/\alpha}, \quad \text{for all } p \geqslant 1.$$

Markov: if $Y \geqslant 0$, then for $t > 0$,

$$\mathbb{P}\{Y \geqslant t\} \leqslant \frac{\mathbb{E}[Y]}{t}.$$

Chebyshev:

$$\mathbb{P}\{|X-\mathbb{E}X| \geqslant t\} \leqslant \frac{\operatorname{Var}(X)}{t^2}.$$

Cantelli (one-sided Chebyshev):

$$\mathbb{P}\{X-\mathbb{E}X \geqslant t\} \leqslant \frac{\operatorname{Var}(X)}{\operatorname{Var}(X)+t^2}.$$

Paley–Zygmund. $X \geqslant 0$ and $\mathbb{E}X^2 < \infty$, then $\forall\, \lambda \in (0,1)$,

$$\mathbb{P}\{X \geqslant \lambda\mathbb{E}X\} \geqslant (1-\lambda^2)\frac{(\mathbb{E}X)^2}{\mathbb{E}X^2}$$

3. Variance Bounds

Here, unique minimizer $a^* = \mathbb{E}Z$, and Z' is iid copy of Z,

$$\operatorname{Var}(Z) = \mathbb{E}[(Z-\mathbb{E}Z)^2] = \min_a \mathbb{E}[(Z-a)^2] = \frac{1}{2}\mathbb{E}[(Z-Z')^2],$$

For every $v \in \mathbb{R}^d$,

$$\operatorname{Var}(v^\top X) = v^\top \operatorname{Cov}(X)v$$

And,

$$\operatorname{tr}(\operatorname{Cov}(X)) = \mathbb{E}\,\|X-\mathbb{E}X\|_2^2$$

Bounded range: If $a \leqslant Z \leqslant b$ (a.s.):

$$\operatorname{Var}(Z) \leqslant \frac{(b-a)^2}{4}$$

Mean versus median If M_X is a median of X, then

$$|M_X - \mathbb{E}X| \leqslant \sqrt{\operatorname{Var}(X)}.$$

Tensorization of Variance:
If $Z = f(X_1, \ldots, X_n)$ with independent $X_1, \ldots, X_n$, and

$$X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n), \qquad \mathbb{E}^{(i)}[\cdot] = \mathbb{E}\left[\cdot \mid X^{(i)}\right],$$

then

$$\operatorname{Var}(Z) \leqslant \sum_{i=1}^n \mathbb{E}\left[\operatorname{Var}(Z \mid X^{(i)})\right] = \sum_{i=1}^n \mathbb{E}\left[(Z-\mathbb{E}^{(i)}Z)^2\right]$$

Bounded differences:

$$\operatorname{Var}(Z) \leqslant \frac{1}{4}\mathbb{E}\left[\sum_{i=1}^n (D_i f)^2\right] \leqslant \frac{1}{4}\sum_{i=1}^n c_i^2,$$

if changing only coordinate i can change f by at most c_i.

Efron–Stein–Steele (ESS):

$$Z^{(i)} := f(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n),$$

where X_i' is an independent copy of X_i, then

$$\operatorname{Var}(Z) \leqslant \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[(Z-Z^{(i)})^2\right].$$

One-sided ESS form:

$$\operatorname{Var}(Z) \leqslant \sum_{i=1}^n \mathbb{E}\left[(Z-Z^{(i)})_+^2\right]$$

This one-sided version is especially useful for suprema.

good luck :)

4. Poincaré Inequality

Gaussian Poincaré:
if $g \sim \mathcal{N}(0, I_n)$ and f is smooth enough, then

$$\mathrm{Var}(f(g)) \leqslant \mathbb{E}\|\nabla f(g)\|_2^2$$

Convex Poincaré:
if $X_1, \ldots, X_n \in [0,1]$ are indep., f is separately convex,

$$\mathrm{Var}(f(X)) \leqslant \mathbb{E}\|\nabla f(X)\|_2^2$$

L-Lipschitz:
$f : [0,1]^n \to \mathbb{R}$ is L–Lipschitz with respect to $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leqslant L\,\|x-y\|_2\,, \qquad \forall x, y \in [0,1]^n$$

Poincaré + L-Lipschitz: if f is L-Lipschitz, then

$$\|\nabla f(x)\|_2 \leqslant L \quad \text{(a.e.)} \implies \mathrm{Var}(f(X)) \leqslant L^2$$

Scaled Poincaré: if $X_i \in [a,b]$ almost surely,

$$\mathrm{Var}(f(X)) \leqslant (b-a)^2 \mathbb{E}\|\nabla f(X)\|_2^2$$

5. Exponential Concentration

Polynomial moment method:

$$\mathbb{P}\{X - EX \geqslant t\} \leqslant \inf_{p \in \mathbb{N}} \frac{\mathbb{E}\left[(X - \mathbb{E}X)_+^p\right]}{t^p}$$

Exponential moment method (Chernoff bound):

$$\mathbb{P}\{X - \mathbb{E}X \geqslant t\} \leqslant \inf_{\lambda > 0} \exp\left(-\lambda t + \log \mathbb{E}e^{\lambda(X - \mathbb{E}X)}\right)$$

6. MGF and CGF

Moment Generating Function (MGF):

$$m_X(\lambda) = \mathbb{E}e^{\lambda X} \in (0, \infty] \qquad (\lambda \in \mathbb{R})$$

Cumulant generating function (CGF):

$$\kappa_X(\lambda) = \log \mathbb{E}e^{\lambda X} \in (-\infty, \infty] \qquad (\lambda \in \mathbb{R})$$

Centered CGF:

$$\psi_X(\lambda) = \log \mathbb{E}e^{\lambda(X - \mathbb{E}X)} = \kappa_X(\lambda) - \lambda \mathbb{E}X$$

Moments from MGF:

$$m_X(\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}\left[X^k\right] \implies m_X^{(k)}(0) = \mathbb{E}\left[X^k\right]$$

Convexity of CGF: If $m_X(\lambda) < \infty$ in a neighborhood of a point λ, then κ_X is twice differentiable there and

$$\kappa_X''(\lambda) = \mathrm{Var}_{\mathbb{P}_\lambda}(X) \geqslant 0,$$

where $\mathbb{P}_\lambda$ is the tilted measure

$$d\mathbb{P}_\lambda = \frac{e^{\lambda X}}{\mathbb{E}\left[e^{\lambda X}\right]}\, d\mathbb{P}$$

Useful result near 0: if m_X is finite near 0,

$$\psi_X(0) = 0, \qquad \psi_X'(0) = 0, \qquad \psi_X''(0) = \mathrm{Var}(X)$$

7. Sub-Gaussian

Definition: X is σ^2-sub-Gaussian if

$$\psi_X(\lambda) = \log \mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leqslant \frac{\sigma^2 \lambda^2}{2} \qquad \forall \lambda \in \mathbb{R}$$

ψ_2 norm: X is sub-Gaussian iff

$$\|X\|_{\psi_2} = \inf\left\{K > 0 \;:\; \mathbb{E}\exp\left(\frac{X^2}{K^2}\right) \leqslant 2\right\} < \infty$$

Equivalent characterizations: Define $Z = X - \mathbb{E}X$

Gaussian tails:

$$\mathbb{P}\{|Z| \geqslant t\} \leqslant 2\exp\left(-\frac{t^2}{2\sigma^2}\right), \qquad t \geqslant 0$$

Moment growth:

$$\|Z\|_{L_p} = (\mathbb{E}|Z|^p)^{1/p} \leqslant C\sigma\sqrt{p}, \qquad p \geqslant 1$$

Example:
- $X \sim \mathcal{N}(0, \sigma^2) \implies$ X is σ^2–sub-Gaussian w/ $\|X\|_{\psi_2} \asymp \sigma$

8. Hoeffding Bounds

Hoeffding's lemma: if $X \in [a,b]$ (a.s.), then for all $\lambda \in \mathbb{R}$,

$$\psi_X(\lambda) = \log \mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leqslant \frac{\lambda^2(b-a)^2}{8}$$

Hoeffding's inequality: (X_i) are independent w/ $X_i \in [a_i, b_i]$ (a.s.) $\forall i$. Let $S_n = \sum_i X_i$ and define the **variance proxy**:

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

then for all $t \geqslant 0$,

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| \geqslant t\} \leqslant 2\exp\left(-\frac{t^2}{2v}\right)$$

Sub-gaussian Hoeffding inequality: (X_i) are indep, mean-zero, and sub-Gaussian w/ $K_i = \|X_i\|_{\psi_2}$. Then, for every $t \geqslant 0$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| \geqslant t\right\} \leqslant 2\exp\left(-c\frac{t^2}{\sum_i K_i^2}\right)$$

Equivalently: $\left\|\sum_i X_i\right\|_{\psi_2}^2 \leqslant C \sum_i \|X_i\|_{\psi_2}^2$

9. Sub-Exponential

Definition: X is sub-exponential if

$$\|X\|_{\psi_1} = \inf\left\{K > 0 \;:\; \mathbb{E}\exp\left(\frac{|X|}{K}\right) \leqslant 2\right\} < \infty$$

Equivalent characterizations: Define $Z = X - \mathbb{E}X$

Exponential tails:

$$\mathbb{P}\{|Z| \geqslant t\} \leqslant 2\exp\left(-c\frac{t}{K}\right), \qquad t \geqslant 0$$

Local CGF Control:

$$\psi_Z(\lambda) = \log \mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leqslant CK^2\lambda^2 \qquad |\lambda| \leqslant \frac{c}{K}$$

Moment growth:

$$\|Z\|_{L_p} = (\mathbb{E}|Z|^p)^{1/p} \leqslant CKp, \qquad p \geqslant 1$$

Closure Properties:
- X is sub-Gaussian, X^2 sub-exponential, $\|X^2\|_{\psi_1} \asymp \|X\|_{\psi_2}^2$
- X, Y is sub-Gaussian, then $\|XY\|_{\psi_1} \lesssim \|X\|_{\psi_2}\|Y\|_{\psi_2}$

10. Bernstein inequality

(X_i) are independent, mean-zero sub-exponential. Set

$$V = \sum_{i=1}^n \|X_i\|_{\psi_1}^2\,, \qquad B = \max_i \|X_i\|_{\psi_1}$$

Then for all $t \geqslant 0$,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| \geqslant t\right\} \leqslant 2\exp\left[-c\min\left(\frac{t^2}{V}, \frac{t}{B}\right)\right]$$

Structure in L_p bound: $\displaystyle \left\|\sum_i X_i\right\|_{L_p} \lesssim \sqrt{p}\sqrt{V} + pB$

11. Common Distributions

1. $X \sim \mathrm{Bernoulli}(p)$, $m_X(\lambda) = 1 - p + pe^\lambda$
2. $X \sim \mathrm{Normal}(\mu, \sigma^2), m_X(\lambda) = \exp\left(\mu\lambda + \sigma^2\lambda^2/2\right)$
3. $X \sim \chi_k^2$, $\mathbb{E}X = k$, $\mathrm{Var}(X) = 2k$, $m_X(\lambda) = (1 - 2\lambda)^{-k/2}$
4. $X \sim \mathrm{Gamma}(\alpha, \beta)$, $m_X(\lambda) = \left(1 - \frac{\lambda}{\beta}\right)^{-\alpha}$, and density is

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\beta x}, \quad \mathbb{E}X = \frac{\alpha}{\beta}$$

12. Good to know

1. $p^{1/p} \leqslant e^{1/e} < e$
2. $\Gamma(z) \leqslant z^z$ for all $z \geqslant 1$
3. $e^{-\min\{a,b\}} \leqslant e^{-a} + e^{-b}$
4. $\min\{t^2/a^2, t/b\} \gtrsim \frac{t^2}{a^2 + bt}$
5. $t \leqslant 2a\sqrt{u} + 2bu \implies \min\{t^2/a^2, t/b\} \lesssim u$
6. Stirling's approximation: $n! \sim \sqrt{2\pi n}(n/e)^n$
7. $\delta^2 = \min\{t, t^2\} \iff t = \max\{\delta, \delta^2\} \leqslant \delta + \delta^2$

good luck :)