Real Analysis Cheatsheet

Hongyi Guo*

November 26, 2020

1 Basics of Information Theory

1.1 Entropy, KL divergence, and mutual information

Definition 1.1 (Entropy). $H(X) = -\sum_X P(X) \log P(X)$.

Theorem 1.2 (Shannon's source coding theorem). Encoding X with a k-ary string needs minimal expected length of $-\sum_X P(X) \log_k P(X)$.

Definition 1.3 (Conditional entropy). $H(X \mid Y = y) = -\sum_X P(X \mid y) \log(X \mid y)$

Definition 1.4 (KL divergence). $\mathrm{KL}(P \,\|\, Q) = -\mathbb{E}[\log(q(X)/p(X))] \geq -\log \mathbb{E}[q(x)/p(x)]$.

Remark 1.5. $H(P) = H(q) + \langle \Delta H(q), p - q \rangle + \mathrm{KL}(p \,\|\, q)$.

Definition 1.6 (Mutual information). $I(X; Y) = \sum_{x,y} p(x, y) \log(p(x, y)/p(x)p(y))$

Proposition 1.7. $I(X; Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$.

Definition 1.8. $I(X; Y \mid Z) = \sum_z I(X; Y \mid Z = z)p(z)$.

Proposition 1.9. $I(X; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = H(Y \mid Z) - H(Y \mid X, Z)$

Proposition 1.10 (Chain rule). $H(X_1^n) = \sum_{i=1}^n H(X_i \mid X_1^{i-1})$.

Proposition 1.11 (Chain rule). $I(X; Y_1^n) = \sum_{i=1}^n I(X_i; Y_i \mid Y_1^{i-1})$.

Proposition 1.12 (Chain rule). $\mathrm{KL}(X_1^n \,\|\, Y_1^n) = \sum_{i=1}^n \mathrm{KL}(X_i \,\|\, Y_i \mid X_1^{i-1})$.

Proposition 1.13 (Data processing inequality). Given a Markov chain $X \to Y \to Z$, it follows that $I(X; Z) \leq I(X; Y)$.

Remark 1.14. $X \to y(X)$ does not change KL or mutual information.

$\mathrm{KL}(P \,\|\, Q)$	$f(t) = t \log t$
$\mathrm{KL}(Q \,\|\, P)$	$f(t) = -\log t$
$\mathrm{TV}(P; Q)$	$f(t) = \frac{1}{2}\lvert t-1\rvert$
$\mathrm{Hel}(P, Q)$	$f(t) = (\sqrt{t}-1)^2$
$\chi^2(P \,\|\, Q)$	$f(t) = (t-1)^2$

Definition 1.15 (f-divergence)**.** $D_f(P \,\|\, Q) = \sum_x q(x) f(p(x)/q(x))$.

Remark 1.16. $\mathrm{TV}(P; Q) = \frac{1}{2} \int |p(x)/q(X) - 1| q(x) dx = \sup_{A \subset F} |P(A) - Q(A)|$.

Proposition 1.17 (Pinsker's inequality)**.** $\mathrm{TV}(P; Q)^2 \leq \frac{1}{2} \mathrm{KL}(P \,\|\, Q)$.

Definition 1.18 (Bregman divergence)**.** $\mathrm{KL}(P \,\|\, Q) = H(p) - H(q) - \langle \Delta H(q), p - q \rangle$.

Proposition 1.19. $\mathrm{KL}(P \,\|\, Q) \leq \log(1 + \chi^2(P \,\|\, Q)) \leq \chi^2(P \,\|\, Q)$.

Proposition 1.20. Let $V \in \{0, 1\}$ be uniform, and draw $X \sim P_v$ conditioning on $V = v$. Then, $I(X; V) = \frac{1}{2} D_f(P_0 \,\|\, P_1) + \frac{1}{2} D_f(P_1 \,\|\, P_0)$, where $f(t) = t \log \frac{2t}{t+1}$. Also, $\mathrm{Hel}^2(P_0, P_1) \leq I(X; V) \leq 2\mathrm{Hel}^2(P_0, P_1)$.

Proposition 1.21. $D_f(\lambda P_1 + (1-\lambda) P_2 \,\|\, \lambda Q_1 + (1-\lambda) Q_2) \leq \lambda D_f(P_1 \,\|\, Q_1) + (1-\lambda) D_f(P_2 \,\|\, Q_2)$.

Remark 1.22. KL is jointly convex in P and Q.

Proposition 1.23 (Data processing inequality)**.** Consider Markov chain $X \to Z$. Let $K(\cdot \mid X)$ be the transition kernel. Then, $K_P(A) = \int_x K(A \mid x) p(x) \mathrm{d}x$. It follows that $D_f(K_P \,\|\, K_Q) \leq D_f(P \,\|\, Q)$.

1.2 Hypothesis testing

Lemma 1.24 (Binary testing with Le Cam's lemma)**.** Nature picks $v \in \{1, 2\}$ uniformly random. Conditioning on $V = v \in \{1, 2\}$, nature generates $X \sim P_v$. Let $\mathbb{P}$ be the joint distribution of X, V. Consider $\mathbb{P}(\varphi(x) \neq v) = \frac{1}{2} P_1(\varphi(x) \neq 1) + \frac{1}{2} P_2(\varphi(x) \neq 2)$. Then, $\inf_\varphi \{P_1(\varphi(x) \neq 1) + P_2(\varphi(x) \neq 2)\} = 1 - \mathrm{TV}(P_1; P_2)$.

Lemma 1.25 (Multiple testing with Fano's lemma)**.** Consider the Markov $X \to Y \to \widehat{X}$, where $X, \widehat{X} \in \mathcal{X}$ (discrete space). Let $E = \mathbb{1}(\widehat{X} \neq X)$. Then $H(E) + \mathbb{P}(\widehat{X} \neq X) \log(|\mathcal{X}| - 1) \geq H(X \mid \widehat{X})$, which implies $\mathbb{P}(\widehat{X} \neq X) \geq 1 - \frac{I(X;Y) + \log 2}{\log |\mathcal{X}|}$.

*Northwestern University; hongyiguo2025@u.northwestern.edu

2 Information, Concentration, Stability, and Generalization

2.1 Concentration inequality

Proposition 2.1 (Markov inequality)**.** $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[x]}{t}, \forall t$, where $X \geq 0$ w.p.1.

Proposition 2.2 (Chebyshev)**.** $\mathbb{P}(X - \mathbb{E}[X] > t) \leq \frac{\mathrm{Var}(X)}{t^2}, \mathbb{P}(X - \mathbb{E}[X] < -t) \leq \frac{\mathrm{Var}(X)}{t^2}, \forall t$, where $\mathrm{Var}(X) < \infty$.

Proposition 2.3 (Chernoff)**.** $\mathbb{P}(X > t) \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda t}} = \varphi_X(\lambda) e^{-\lambda t}$.

Proposition 2.4 (Standard Chenoff)**.** $\mathbb{P}(X > t) \leq \inf_{\lambda \geq 0} \{\varphi_X(\lambda) e^{-\lambda t}\}$.

Remark 2.5. Both Markov inequality and Chebyshev inequality gives us polynomial tails, but Chenoff gives us an exponential tail.

Example 2.6 (Gaussian)**.** Gaussian variable X has $\varphi_X(\lambda) = \mathbb{E}[\exp(\lambda X)] = \exp(\lambda^2 \sigma^2/2)$. Chernoff gives $\mathbb{P}(X \geq \mathbb{E}[X] + t) \leq \exp(-\lambda^2\sigma^2/2), \mathbb{P}(X \leq \mathbb{E}[X] - t) \leq \exp(-\lambda^2\sigma^2/2)$.

2.2 Sub-Gaussian

Definition 2.7 (Sub-Gaussian random variable)**.** X is sub-Gaussian with "variance proxy" σ^2 iff $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\frac{\lambda^2\sigma^2}{2})$. Gaussian with variance σ^2 attains the "=".

Remark 2.8 (Rademacher random variable)**.** $X \in \{-1, 1\}$, $\mathbb{E}[\exp(\lambda X)] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \exp(\frac{\lambda^2}{2})$.

Proof. Use <u>Taylor expansion</u>. □

Proposition 2.9 (Hoeffding bound)**.** If $X \in [a, b]$, then $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\frac{\lambda^2(b-a)^2}{8})$.

Proposition 2.10 (Chernoff bound for sub-Gaussian)**.** Let X be σ^2 sub-Gaussian. $\forall t \geq 0, \mathbb{P}(X - \mathbb{E}[X] \geq t \vee X - \mathbb{E}[X] \leq -t) \leq \exp(-\frac{t^2}{2\sigma^2})$. Tensorization of MGF gives $\mathbb{E}[\exp(\lambda \sum_{i=1}^n (X_i - \mathbb{E}[X_i]))] \leq \exp(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2})$. Hence, $\sum_{i=1}^n X_i$ is $\sum_{i=1}^n \sigma_i^2$-sub-Gaussian. $\mathbb{P}(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t \vee \sum_{i=1}^n x_i - \mathbb{E}[X_i] \leq t) \leq \exp(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2})$.

Remark 2.11. If X_i is bounded by $[a_i, b_i]$, then $\mathbb{P}(\sum_{i=1}^n X_i - \mathbb{E}[X_i] \geq t) \leq \exp(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2})$.

Definition 2.12 (Equivalent definition of sub-Gaussian)**.** Orlicz norm $\|X\|_{\varphi_2} = \sup_{k \geq 1} \frac{1}{\sqrt{k}} \mathbb{E}[|X|^k]^{1//k}$ Let X has mean zero and $\sigma^2 \geq 0$ be a constant. The following statements are true up to constant scaling of k

1. sub-Gaussian tail $\mathbb{P}(|X| \geq t) \leq 2\exp(-\frac{t^2}{k\sigma^2}), \forall t \geq 0$,

2. sub-Gaussian moment $\frac{1}{\sqrt{k}} \mathbb{E}[|X|^k]^{1/k} \leq k\sigma, \forall k, \|X\|_{\varphi_2} = \sigma$,

3. sub-Gaussian moment $\mathbb{E}[\exp(\frac{X^2}{k\sigma^2})] \leq e$,

4. sub-Gaussian MGF $\mathbb{E}[\exp(\lambda X)] \leq \exp(k\lambda^2\sigma^2), \forall\lambda$.

Remark 2.13 (Sub-Gaussian squared)**.** $\mathbb{E}[\exp(\lambda X^2)] \leq \frac{1}{[1-2\delta^2\lambda]_+^{1/2}}$.

2.3 Sub-exponential

Definition 2.14 (Sub-exponential)**.** X is sub-exponential with (σ^2, b) iff $\forall\lambda$ with $|\lambda| \leq 1/b$, $\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(\frac{\lambda^2\sigma^2}{2})$.

Remark 2.15. σ^2-sub-Gaussian is $(\sigma^2, 0)$-sub-exponential.

Remark 2.16. Let $X = Z^2$, where $Z \sim N(0,1) \Rightarrow \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp(2\lambda^2), \forall\lambda \leq \frac{1}{4}$.

Remark 2.17 (Bounded random variables are sub-exponential)**.** $X \in [-b, b], \mathbb{E}[X] = 0, \sigma^2 = \mathbb{E}[X^2] \Rightarrow \mathbb{E}[\exp(\lambda X)] \leq \exp(3\lambda^2\sigma^2/5), \forall|\lambda| \leq \frac{1}{2b}$.

Proof. Tricks: $\mathbb{E}[|X|^k] \leq \mathbb{E}[X^2 b^{k-2}] = \sigma^2 b^{k-2}$, and when dealing with $\sum_{k=1}^{\infty}(\lambda b)^k/(k+2)!$, calculate the first two terms and use the series after that. □

3 Information theory, reinforcement learning, regret