Information Theory for Statistics and Learning

(Transcribed from handwritten course notes by GPT 5.2 Pro. Beware of transcription errors.)

Yanjun Han

January 17, 2026

Contents

Lecture 1: Entropy & Mutual Information

Remark 1.1 (Logarithm convention)*.* For information-theoretic quantities in this lecture, log denotes $\log_2$ (so entropies are measured in *bits*).

1.1 Entropy

Definition 1.2 (Entropy)**.** Let X be a discrete random variable taking values in an alphabet $\mathcal{X}$ with pmf p. Its entropy $H(X)$ (or $H(p)$) is

$$H(X) = \sum_{x\in\mathcal{X}} p(x)\,\log\frac{1}{p(x)}. \tag{1.1}$$

Remark 1.3.

1. $0 \leq H(X) \leq \log|\mathcal{X}|$. (One way to see the upper bound is via Jensen's inequality.)
2. If $|\mathcal{X}| = \infty$, then $H(X)$ may be finite or infinite.
3. For continuous (or more general) random variables, one chooses a reference measure μ such that X has a density f w.r.t. μ, and defines the *differential entropy*

   $$h(X) = \int f(x)\,\log\frac{1}{f(x)}\,d\mu(x). \tag{1.2}$$

   The value of $h(X)$ depends on the choice of μ.
4. (Base of log.) For IT applications in this lecture, use base 2 (bits). In many other settings one uses natural logs (nats).

1.1.1 Why entropy? Source coding (i.i.d. case)

Shannon (1948) showed that entropy characterizes the fundamental limit of source coding.

Source coding problem. Given:

(1) an input alphabet $\mathcal{X}$ (e.g. English letters $\{a, b, \ldots, z\}$),

(2) a known pmf p on $\mathcal{X}$ (the *source distribution*),

find a map (code)

$$f : \mathcal{X} \to \{0,1\}^* \;:=\; \bigcup_{n\geq 1}\{0,1\}^n, \tag{1.3}$$

such that

(1) f is *uniquely decodable*: from the concatenation $f(x_1)\cdots f(x_n)$ one can uniquely recover both n and $(x_1,\dots,x_n)\in\mathcal{X}^n$;

(2) the expected code length is minimized:

$$\mathbb{E}\big[\ell(f(X))\big] = \sum_{x\in\mathcal{X}} p(x)\,\ell\big(f(x)\big), \tag{1.4}$$

where $\ell(\cdot)$ is the length (in bits) of a binary string.

Example 1.4. Let $\mathcal{X}=\{a,b,c\}$ and $p=(\frac14,\frac12,\frac14)$.

(a) The code $a\mapsto 0$, $b\mapsto 10$, $c\mapsto 11$ is uniquely decodable (e.g. 1001011 decodes to *babc*).

(b) The code $a\mapsto 0$, $b\mapsto 1$, $c\mapsto 10$ is *not* uniquely decodable (e.g. 10 could be c or ba).

(c) The code $a\mapsto 10$, $b\mapsto 0$, $c\mapsto 11$ is uniquely decodable and has smaller expected length:

$$2\cdot\tfrac14 + 1\cdot\tfrac12 + 2\cdot\tfrac14 = 1.5 \text{ bits} < 1.75 \text{ bits for (a)}.$$

1.1.2 Kraft–McMillan theorem

Given a length profile $\{\ell_x\}_{x\in\mathcal{X}}$, when does there exist a uniquely decodable code f with $\ell(f(x))=\ell_x$?

Theorem 1.5 (Kraft–McMillan (Kraft inequality))**.** *A necessary and sufficient condition is*

$$\sum_{x\in\mathcal{X}} 2^{-\ell_x} \le 1. \tag{1.5}$$

Proof sketch. Sufficiency. If $\sum_x 2^{-\ell_x}\le 1$, one can construct a full binary tree whose leaves include $\mathcal{X}$ with $\mathrm{depth}(x)=\ell_x$. Assigning each symbol the bitstring along its root-to-leaf path produces a *prefix code*, hence uniquely decodable.

Necessity. Assume w.l.o.g. that $|\mathcal{X}|<\infty$ and $\ell_{\max} := \max_x \ell(f(x)) < \infty$. For any $m\in\mathbb{N}$,

$$\begin{aligned}\Big(\sum_{x\in\mathcal{X}} 2^{-\ell(f(x))}\Big)^m &= \sum_{x_1,\dots,x_m\in\mathcal{X}} 2^{-(\ell(f(x_1))+\cdots+\ell(f(x_m)))}\\ &= \sum_{x_1,\dots,x_m\in\mathcal{X}} 2^{-\ell(f(x_1)\cdots f(x_m))}\\ &= \sum_{t=1}^{m\ell_{\max}} 2^{-t}\,N_t,\end{aligned}$$

where N_t counts the number of m-tuples whose concatenated codeword has total length t. By unique decodability, distinct m-tuples yield distinct binary strings, so $N_t\le 2^t$. Hence

$$\Big(\sum_{x\in\mathcal{X}} 2^{-\ell(f(x))}\Big)^m \le \sum_{t=1}^{m\ell_{\max}} 1 = m\ell_{\max}.$$

Taking $m\to\infty$ gives $\sum_x 2^{-\ell(f(x))}\le 1$. □

1.1.3 Source coding theorem (uniquely decodable codes)

Theorem 1.6 (Source coding theorem for uniquely decodable codes)**.**

$$H(X) \;\le\; \min_{\textit{uniquely decodable } f} \mathbb{E}\big[\ell(f(X))\big] \;<\; H(X)+1. \tag{1.6}$$

Proof sketch. Upper bound. Set $\ell_x := \left\lceil \log \frac{1}{p(x)} \right\rceil$. Then $\{\ell_x\}$ satisfies Kraft's inequality, and

$$\sum_{x\in\mathcal{X}} p(x)\,\ell_x \;<\; \sum_{x\in\mathcal{X}} p(x)\left(\log\tfrac{1}{p(x)}+1\right) \;=\; H(X)+1.$$

Lower bound. Minimizing $\sum_x p(x)\ell_x$ subject to $\sum_x 2^{-\ell_x} \le 1$ (allowing real lengths) yields the optimum $\ell_x^\star = \log\frac{1}{p(x)}$, giving value $H(X)$. (One can verify this via Lagrange multipliers.) □

Remark 1.7.
1. The gap between $H(X)$ and $H(X)+1$ can be significant (e.g. if $H(X) = 1.5$ bits). In practice the alphabet is often a "super-symbol" alphabet, e.g. $\mathcal{X} = \{a,\dots,z\}^{256}$, in which case $H(X) \gg 1$ bit.

2. Information theory often provides *robust* results when a small error probability is allowed; purely combinatorial arguments (like Kraft counting) typically do not extend as cleanly.

1.2 Asymptotic equipartition property (AEP)

Another way to write entropy is

$$H(X) = \mathbb{E}_{X\sim p}\Big[\log\frac{1}{p(X)}\Big]. \tag{1.7}$$

(A mild warning: the distribution p appears both in the expectation and inside the logarithm.)

Let $X_1,\dots,X_n$ be i.i.d. $\sim p$. If $H(X) < \infty$, then by the law of large numbers,

$$\frac{1}{n}\log\frac{1}{p(X_1,\dots,X_n)} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{p(X_i)} \xrightarrow[n\to\infty]{\text{a.s.}} \mathbb{E}\Big[\log\frac{1}{p(X)}\Big] = H(X). \tag{1.8}$$

For $\varepsilon > 0$, define the *typical set*

$$T_n^\varepsilon := \Big\{x^n \in \mathcal{X}^n : p(x^n) \in \big[2^{-n(H(X)+\varepsilon)},\, 2^{-n(H(X)-\varepsilon)}\big]\Big\}. \tag{1.9}$$

Theorem 1.8 (AEP)**.** *The typical set T_n^ε satisfies:*

(1) $\mathbb{P}\big((X_1,\dots,X_n) \in T_n^\varepsilon\big) \to 1$ *as* $n\to\infty$.

(2) $(1-o(1))\,2^{n(H(X)-\varepsilon)} \le |T_n^\varepsilon| \le 2^{n(H(X)+\varepsilon)}$.

Remark 1.9. In words: for i.i.d. samples, the joint distribution of $X_1,\dots,X_n$ is "roughly" uniform over about $2^{nH(X)}$ typical sequences.

1.3 Source coding with error probability

Consider an encoder/decoder pair

$$(X_1, \dots, X_n) \xrightarrow{\text{encoder}} Y \in \{0,1\}^* \xrightarrow{\text{decoder}} (\widehat{X}_1, \dots, \widehat{X}_n),$$

with a block error guarantee $\mathbb{P}((X_1, \dots, X_n) \neq (\widehat{X}_1, \dots, \widehat{X}_n)) \leq \delta$.

Theorem 1.10 (Source coding theorem with error probability)**.** *(1)* ***Achievability.*** *There exist encoder/decoder pairs such that* $\frac{1}{n}\mathbb{E}[\ell(Y)] \leq H(p) + o(1)$ *and* $\delta = o(1)$*.*

(2) ***Converse.*** *If* $\delta = o(1)$*, then any encoder/decoder pair must satisfy* $\frac{1}{n}\mathbb{E}[\ell(Y)] \geq H(p) - o(1)$*.*

Proof sketch. Achievability. Encode only the typical sequences in T_n^ε: enumerate the elements of T_n^ε and transmit the index; declare an error otherwise. By AEP, $\mathbb{P}((X_1, \dots, X_n) \notin T_n^\varepsilon) \to 0$. Moreover,

$$\ell(Y) \leq \log |T_n^\varepsilon| \leq n(H(p) + \varepsilon) \quad \text{deterministically.}$$

Since $\varepsilon > 0$ is arbitrary, this gives $\frac{1}{n}\mathbb{E}[\ell(Y)] \leq H(p) + o(1)$.

Converse. Fix $\varepsilon > 0$. Define

$$\begin{aligned} A &:= \{X^n : \ell(Y) > n(H(p) - 2\varepsilon)\}, \\ B &:= \{X^n : X^n = \widehat{X}^n\}. \end{aligned}$$

By AEP and a union bound, $\mathbb{P}(T_n^\varepsilon \cap B) \geq 1 - \delta - o(1)$. Also,

$$\begin{aligned} |T_n^\varepsilon \cap B \cap A^c| &\leq |\{y \in \{0,1\}^* : \ell(y) \leq n(H(p) - 2\varepsilon)\}| \\ &= \sum_{k=1}^{n(H(p)-2\varepsilon)} 2^k < 2 \cdot 2^{n(H(p)-2\varepsilon)}. \end{aligned}$$

Each $x^n \in T_n^\varepsilon$ has probability at most $2^{-n(H(p)-\varepsilon)}$, hence

$$\mathbb{P}(T_n^\varepsilon \cap B \cap A^c) \leq 2^{-n(H(p)-\varepsilon)} \, |T_n^\varepsilon \cap B \cap A^c| < 2 \cdot 2^{-n\varepsilon}.$$

Therefore $\mathbb{P}(T_n^\varepsilon \cap A \cap B) \geq 1 - \delta - o(1) - 2 \cdot 2^{-n\varepsilon} = 1 - o(1)$, so $\mathbb{P}(A) \geq 1 - o(1)$. Finally,

$$\frac{1}{n}\mathbb{E}[\ell(Y)] \geq (H(p) - 2\varepsilon)\,\mathbb{P}(A) \geq (1 - o(1))(H(p) - 2\varepsilon).$$

Letting $\varepsilon \downarrow 0$ yields the converse. □

1.4 Joint entropy and mutual information

Definition 1.11 (Joint and conditional entropies)**.** For a pair (X, Y), define

$$H(X, Y) := \mathbb{E}_{X,Y}\Big[\log \frac{1}{p(X, Y)}\Big], \tag{1.10}$$

$$H(Y \mid X) := \mathbb{E}_{X,Y}\Big[\log \frac{1}{p(Y \mid X)}\Big] = H(X, Y) - H(X). \tag{1.11}$$

Definition 1.12 (Mutual information)**.**

$$I(X;Y) := H(X) + H(Y) - H(X,Y) \tag{1.12}$$

$$= H(Y) - H(Y \mid X) \tag{1.13}$$

$$= \mathbb{E}_{X,Y}\Big[\log \frac{p(X,Y)}{p(X)p(Y)}\Big]. \tag{1.14}$$

Lemma 1.13 (Non-negativity of mutual information)**.** *$I(X;Y) \geq 0$. Equivalently, conditioning reduces entropy: $H(X) \geq H(X \mid Y)$.*

Typicality/AEP proof sketch. A one-line proof uses KL divergence, but we follow the typicality argument.

For $\varepsilon > 0$, define

$$T_n^\varepsilon(X) := \Big\{(x^n, y^n) : \Big|\frac{1}{n}\sum_{i=1}^n \log\frac{1}{p_X(x_i)} - H(X)\Big| \leq \varepsilon\Big\},$$

$$T_n^\varepsilon(Y) := \Big\{(x^n, y^n) : \Big|\frac{1}{n}\sum_{i=1}^n \log\frac{1}{p_Y(y_i)} - H(Y)\Big| \leq \varepsilon\Big\},$$

$$T_n^\varepsilon(X,Y) := \Big\{(x^n, y^n) : \Big|\frac{1}{n}\sum_{i=1}^n \log\frac{1}{p_{XY}(x_i, y_i)} - H(X,Y)\Big| \leq \varepsilon\Big\},$$

and the joint typical set $T_n^\varepsilon := T_n^\varepsilon(X) \cap T_n^\varepsilon(Y) \cap T_n^\varepsilon(X,Y)$.

If $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. $\sim p_{XY}$, then LLN + union bound give $\mathbb{P}((X^n, Y^n) \in T_n^\varepsilon) \to 1$, and hence $|T_n^\varepsilon| \geq (1 - o(1))2^{n(H(X,Y)-\varepsilon)}$.

Now draw $(\widetilde{X}_1, \widetilde{Y}_1), \dots, (\widetilde{X}_n, \widetilde{Y}_n)$ i.i.d. $\sim p_X p_Y$ (independent). Then

$$\begin{aligned} 1 &\geq \mathbb{P}((\widetilde{X}^n, \widetilde{Y}^n) \in T_n^\varepsilon) \\ &= \sum_{(x^n, y^n) \in T_n^\varepsilon} p_X(x^n) p_Y(y^n) \\ &\geq (1 - o(1))2^{n(H(X,Y)-\varepsilon)} \cdot 2^{-n(H(X)+\varepsilon)} \cdot 2^{-n(H(Y)+\varepsilon)} \\ &= (1 - o(1))2^{-n(I(X;Y)+3\varepsilon)}. \end{aligned}$$

Thus $I(X;Y) + 3\varepsilon \geq 0$, and letting $\varepsilon \downarrow 0$ yields $I(X;Y) \geq 0$. □

1.4.1 Some consequences (Shannon-type inequalities)

The non-negativity $I(X;Y) \geq 0$ is a fundamental inequality used to prove many others. For instance:

(1) $H(X_1, \dots, X_n) = \sum_{k=1}^n H(X_k \mid X_1, \dots, X_{k-1}) \leq \sum_{k=1}^n H(X_k)$.

(2) If $P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i}$, then

$$I(X^n; Y^n) = H(Y^n) - H(Y^n \mid X^n) \leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i \mid X_i) = \sum_{i=1}^n I(X_i; Y_i).$$

(3) If $P_{X^n} = \prod_{i=1}^n P_{X_i}$, then

$$I(X^n; Y^n) = H(X^n) - H(X^n \mid Y^n) \geq \sum_{i=1}^n H(X_i) - \sum_{i=1}^n H(X_i \mid Y_i) = \sum_{i=1}^n I(X_i; Y_i).$$

All inequalities that can be derived from *monotonicity* $H(X) \leq H(X,Y)$ and *submodularity* $H(X_A) + H(X_B) \geq H(X_{A\cup B}) + H(X_{A\cap B})$ are often called *Shannon-type inequalities.*

1.5 Channel coding and channel capacity

Channel model. A message $m \sim \mathrm{Unif}(\{1,\dots,M\})$ is encoded into a channel input $X^n \in \mathcal{X}^n$, passed through a memoryless channel $P_{Y|X}$ so that $P_{Y^n|X^n} = \prod_{i=1}^n P_{Y_i|X_i}$, and decoded into $\widehat{m} \in \{1,\dots,M\}$.

Given a block error guarantee $\mathbb{P}(m \neq \widehat{m}) \le \delta$, the goal is to maximize the communication rate

$$R_n := \frac{\log M}{n} \qquad \text{(bits per channel use).} \tag{1.15}$$

Definition 1.14 (Channel capacity)**.** The (Shannon) channel capacity is

$$C = C(P_{Y|X}) := \max_{P_X} I(X;Y), \qquad \text{where } P_{XY} = P_X P_{Y|X}. \tag{1.16}$$

Equivalently: choose an input distribution P_X that maximizes the mutual information between input and output.

1.5.1 Examples

Binary symmetric channel (BSC). For $\mathcal{X} = \mathcal{Y} = \{0,1\}$ and crossover probability $\varepsilon \in [0,1]$,

$$P_{Y|X} = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}.$$

One has

$$I(X;Y) = H(Y) - H(Y \mid X) \le 1 - h_2(\varepsilon), \tag{1.17}$$

with equality iff $P_X = (\frac{1}{2}, \frac{1}{2})$. Here

$$h_2(\varepsilon) := \varepsilon \log \frac{1}{\varepsilon} + (1-\varepsilon) \log \frac{1}{1-\varepsilon} \tag{1.18}$$

is the binary entropy function.

Binary erasure channel (BEC). For $\mathcal{X} = \{0,1\}$, $\mathcal{Y} = \{0,1,\perp\}$, and erasure probability ε,

$$P_{Y|X} = \begin{array}{c|ccc} & 0 & 1 & \perp \\ \hline 0 & 1-\varepsilon & 0 & \varepsilon \\ 1 & 0 & 1-\varepsilon & \varepsilon \end{array}.$$

Then

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= H(X) - \mathbb{P}(Y \neq \perp)\, H(X \mid Y \neq \perp) - \mathbb{P}(Y = \perp)\, H(X \mid Y = \perp) \\ &= H(X) - 0 - \varepsilon H(X) \\ &= (1-\varepsilon) H(X) \le 1 - \varepsilon, \end{aligned}$$

with equality iff $P_X = (\frac{1}{2}, \frac{1}{2})$.

1.5.2 Shannon's channel coding theorem (statement)

Theorem 1.15 (Shannon's channel coding theorem)**.** *Fix any $\varepsilon > 0$.*

(1) ***Achievability.*** *If $R_n < C - \varepsilon$, then there exist encoders/decoders such that $\mathbb{P}(m \neq \widehat{m}) \to 0$ as $n \to \infty$.*

(2) ***(Weak) converse.*** *If $R_n > C+\varepsilon$, then for every encoder/decoder sequence, $\liminf_{n\to\infty} \mathbb{P}(m \neq \widehat{m}) > 0$.*

(A strong converse *strengthens the second statement to $\mathbb{P}(m \neq \widehat{m}) \to 1$; see later lectures.)*

1.5.3 Achievability idea: random coding and typicality

Generate a random codebook $X^n_{(1)}, \dots, X^n_{(M)}$ i.i.d. $\sim P_X^{\otimes n}$. To send message m, transmit $X^n_{(m)}$. Given the channel output Y^n, decode by finding the *unique* $\widehat{m}$ such that $(X^n_{(\widehat{m})}, Y^n)$ is jointly typical; if none (or not unique), declare an error.

Assuming the true message is $m = 1$, successful decoding occurs if:

(1) $(X^n_{(1)}, Y^n)$ is jointly typical;

(2) none of $(X^n_{(2)}, Y^n), \dots, (X^n_{(M)}, Y^n)$ is jointly typical.

By LLN, event (1) holds with probability $1 - o(1)$. Moreover, since $(X^n_{(2)}, Y^n) \sim P_X^{\otimes n} \otimes P_Y^{\otimes n}$ (independent), the typicality bound implies $\mathbb{P}((X^n_{(2)}, Y^n) \text{ jointly typical}) \le 2^{-n(I(X;Y)-3\varepsilon)}$. A union bound gives

$$\mathbb{P}(\text{event (2)}) \ge 1 - M \cdot 2^{-n(I(X;Y)-3\varepsilon)}.$$

If $\log M < n(I(X;Y) - 4\varepsilon)$, then $\mathbb{P}(\widehat{m} = 1) \ge 1 - o(1)$. Optimizing over P_X yields rates below capacity.

Remark 1.16. Random coding was historically surprising (algebraic codes dominated early intuition) and helped motivate the probabilistic method. The typicality decoder is computationally expensive; capacity-achieving efficient codes (e.g. spatially coupled LDPC codes and polar codes) were developed much later.

1.6 Weak converse via Fano's inequality

Lemma 1.17 (Data processing inequality for mutual information)**.** *If $X - Y - Z$ forms a Markov chain (i.e. $P_{XYZ} = P_X P_{Y|X} P_{Z|Y}$), then*

$$I(X;Y) \ge I(X;Z). \tag{1.19}$$

Proof. Using Shannon-type identities,

$$I(X;Y) - I(X;Z) = H(X \mid Z) - H(X \mid Y) = H(X \mid Z) - H(X \mid Y, Z) = I(X;Y \mid Z) \ge 0.$$

□

Theorem 1.18 (Fano's inequality (one convenient form))**.** *If $X \sim \mathrm{Unif}([M])$, then*

$$\mathbb{P}(X \neq Y) \;\ge\; 1 - \frac{I(X;Y) + \log 2}{\log M}. \tag{1.20}$$

Proof. Let $E := \mathbb{1}\{X \neq Y\}$ and $p_e := \mathbb{P}(E = 1) = \mathbb{P}(X \neq Y)$. Then

$$\begin{aligned} H(X \mid Y) &= H(X \mid Y, E) + I(X; E \mid Y) \\ &\leq H(X \mid Y, E) + H(E) \\ &\leq \mathbb{P}(E = 1)\, H(X \mid Y, E = 1) + \mathbb{P}(E = 0)\, H(X \mid Y, E = 0) + \log 2 \\ &\leq p_e \log M + \log 2. \end{aligned}$$

On the other hand, since X is uniform, $H(X \mid Y) = H(X) - I(X;Y) = \log M - I(X;Y)$. Rearranging yields the claim. □

1.6.1 Applying Fano to channel coding (weak converse)

If the communication rate satisfies $R_n > C + \varepsilon$, then applying Fano's inequality to $(m, \widehat{m})$ gives

$$\begin{aligned} \mathbb{P}(m \neq \widehat{m}) &\geq 1 - \frac{I(m; \widehat{m}) + \log 2}{\log M} \\ &\geq 1 - \frac{I(X^n; Y^n) + \log 2}{\log M} \quad \text{(Markov chain } m - X^n - Y^n - \widehat{m}\text{)} \\ &\geq 1 - \frac{\sum_{i=1}^n I(X_i; Y_i) + \log 2}{\log M} \quad \text{(memoryless channel bound)} \\ &\geq 1 - \frac{nC + \log 2}{\log M} \quad \text{(definition of } C\text{)}. \end{aligned}$$

Since $\log M = nR_n > n(C + \varepsilon)$, the right-hand side tends to $\varepsilon/(C + \varepsilon) > 0$, establishing the weak converse.

Lecture 2: KL Divergence

2.1 Kullback–Leibler (KL) divergence

Definition 2.1 (KL divergence / relative entropy)**.** Let P and Q be probability measures on the same measurable space. The *Kullback–Leibler divergence* (or *relative entropy*) of P with respect to Q is

$$D_{\mathrm{KL}}(P\|Q) := \begin{cases} \mathbb{E}_{X\sim P}\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}(X)\right], & \text{if } P \ll Q, \\ +\infty, & \text{otherwise.} \end{cases}$$

Remark 2.2.

1. The definition covers both discrete and continuous cases. If p, q are pmfs on a countable set $\mathcal{X}$,

   $$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

   If p, q are densities with respect to a common reference measure μ,

   $$D_{\mathrm{KL}}(P\|Q) = \int p(x) \log \frac{p(x)}{q(x)}\, \mu(\mathrm{d}x).$$

2. This is a *divergence* rather than a distance: in general $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$. Hence we write $D_{\mathrm{KL}}(P\|Q)$ instead of $D_{\mathrm{KL}}(P, Q)$.

3. Information-theoretic origin (redundancy). In the discrete case,

   $$D_{\mathrm{KL}}(P\|Q) = \sum_{x} p(x) \log \frac{1}{q(x)} - H(P),$$

   where $H(P) = \sum_x p(x) \log \frac{1}{p(x)}$ is Shannon entropy. Thus $D_{\mathrm{KL}}(P\|Q)$ equals the expected code length when using a code optimal for Q minus the optimal expected code length for source P.

2.1.1 Basic properties

Proposition 2.3 (Property I: nonnegativity)**.** *For any P, Q, $D_{\mathrm{KL}}(P\|Q) \geq 0$, with equality if and only if $P = Q$.*

Proof. Assume $P \ll Q$ and write $Z := \frac{\mathrm{d}P}{\mathrm{d}Q}$. Then $\mathbb{E}_Q[Z] = 1$ and

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_P[\log Z] = \mathbb{E}_Q[Z \log Z].$$

The function $\varphi(t) = t \log t$ is convex on $\mathbb{R}_+$ and satisfies $\varphi(1) = 0$. By Jensen's inequality,

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_Q[\varphi(Z)] \geq \varphi(\mathbb{E}_Q[Z]) = \varphi(1) = 0.$$

Moreover, equality holds if and only if $Z = 1$ Q-a.s., i.e. $P = Q$. If $P \not\ll Q$, then $D_{\mathrm{KL}}(P\|Q) = +\infty$ by definition. □

Remark 2.4 (Mutual information as a KL divergence). For random variables (X, Y) with joint law P_{XY} and marginals P_X, P_Y,

$$I(X;Y) = \mathbb{E}\left[\log \frac{\mathrm{d}P_{XY}}{\mathrm{d}(P_X \otimes P_Y)}(X,Y)\right] = D_{\mathrm{KL}}\big(P_{XY}\|P_X \otimes P_Y\big) \geq 0.$$

Equality holds if and only if $P_{XY} = P_X \otimes P_Y$, i.e. X and Y are independent.

Proposition 2.5 (Property II: joint convexity)**.** *The map* $(P, Q) \mapsto D_{\mathrm{KL}}(P\|Q)$ *is jointly convex.*

Proof sketch. In the discrete/density setting, $D_{\mathrm{KL}}(P\|Q) = \int \phi\Big(\frac{p}{q}\Big)\, q$, where $\phi(u) = u \log u$. Equivalently, one may use the joint convexity of $(x, y) \mapsto x \log \frac{x}{y}$ on $\mathbb{R}^2_+$. Indeed, for $f(x, y) = x \log(x/y)$,

$$\nabla^2 f(x,y) = \begin{pmatrix} \frac{1}{x} & -\frac{1}{y} \\ -\frac{1}{y} & \frac{x}{y^2} \end{pmatrix} \succeq 0.$$

Proposition 2.6 (Property III: chain rule)**.** *Let* $X^n = (X_1, \dots, X_n)$*. Then*

$$D_{\mathrm{KL}}(P_{X^n}\|Q_{X^n}) = \sum_{i=1}^{n} \mathbb{E}_{P_{X^{i-1}}}\Big[D_{\mathrm{KL}}\big(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}\big)\Big].$$

Proof. Write the likelihood ratio via conditional distributions: $\frac{P_{X^n}}{Q_{X^n}} = \prod_{i=1}^{n} \frac{P_{X_i|X^{i-1}}}{Q_{X_i|X^{i-1}}}$. Taking logs and expectations under P_{X^n} gives

$$D_{\mathrm{KL}}(P_{X^n}\|Q_{X^n}) = \mathbb{E}_{P_{X^n}}\left[\sum_{i=1}^{n} \log \frac{P_{X_i|X^{i-1}}(X_i \mid X^{i-1})}{Q_{X_i|X^{i-1}}(X_i \mid X^{i-1})}\right] = \sum_{i=1}^{n} \mathbb{E}_{P_{X^{i-1}}}\Big[D_{\mathrm{KL}}\big(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}\big)\Big].$$

□

2.1.2 Data processing inequality

Theorem 2.7 (Data processing inequality (DPI))**.** *Let* P_X, Q_X *be distributions on* $\mathcal{X}$ *and let* $P_{Y|X}$ *be a Markov kernel (channel) from* $\mathcal{X}$ *to* $\mathcal{Y}$*. Let* P_Y, Q_Y *be the output laws induced by the same channel:* $P_Y = P_X P_{Y|X}$ *and* $Q_Y = Q_X P_{Y|X}$*. Then*

$$D_{\mathrm{KL}}(P_X\|Q_X) \geq D_{\mathrm{KL}}(P_Y\|Q_Y).$$

(In words: distributions become "closer" after processing.)

Proof (method 1: convexity / Jensen). Assume $P_X \ll Q_X$ (otherwise $D_{\mathrm{KL}}(P_X\|Q_X) = +\infty$ and the claim is trivial). Let

$$L_X(x) := \frac{\mathrm{d}P_X}{\mathrm{d}Q_X}(x).$$

Define the joint laws $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X P_{Y|X}$. Then $P_{XY} \ll Q_{XY}$ and

$$\frac{\mathrm{d}P_{XY}}{\mathrm{d}Q_{XY}}(x, y) = \frac{\mathrm{d}P_X}{\mathrm{d}Q_X}(x) = L_X(x).$$

Let

$$L_Y(y) \coloneqq \frac{\mathrm{d}P_Y}{\mathrm{d}Q_Y}(y).$$

We claim that

$$L_Y(Y) = \mathbb{E}_{X \sim Q_{X|Y}}\big[L_X(X) \mid Y\big] \qquad Q_Y\text{-a.s.}$$

(this is the "exercise" step in the handwritten notes). Indeed, for any bounded measurable g,

$$\begin{aligned} \mathbb{E}_{Q_Y}[g(Y)L_Y(Y)] &= \mathbb{E}_{P_Y}[g(Y)] = \mathbb{E}_{P_{XY}}[g(Y)] \\ &= \mathbb{E}_{Q_{XY}}[g(Y)L_X(X)] = \mathbb{E}_{Q_Y}\Big[g(Y)\,\mathbb{E}_{Q_{X|Y}}[L_X(X) \mid Y]\Big], \end{aligned}$$

which identifies L_Y as the conditional expectation.

Now use $\varphi(t) = t \log t$ (convex on $\mathbb{R}_+$). Then

$$\begin{aligned} D_{\mathrm{KL}}(P_Y \| Q_Y) &= \mathbb{E}_{P_Y}[\log L_Y(Y)] = \mathbb{E}_{Q_Y}[L_Y(Y) \log L_Y(Y)] = \mathbb{E}_{Q_Y}[\varphi(L_Y(Y))] \\ &= \mathbb{E}_{Q_Y}\Big[\varphi\big(\mathbb{E}_{Q_{X|Y}}[L_X(X) \mid Y]\big)\Big] \le \mathbb{E}_{Q_Y}\mathbb{E}_{Q_{X|Y}}[\varphi(L_X(X)) \mid Y] \qquad \text{(Jensen)} \\ &= \mathbb{E}_{Q_X}[\varphi(L_X(X))] = \mathbb{E}_{P_X}[\log L_X(X)] = D_{\mathrm{KL}}(P_X \| Q_X). \end{aligned}$$

□

Proof (method 2: chain rule). Form the joint laws $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X P_{Y|X}$. Since the conditional distributions coincide, the chain rule gives

$$D_{\mathrm{KL}}(P_{XY} \| Q_{XY}) = D_{\mathrm{KL}}(P_X \| Q_X) + \mathbb{E}_{P_X}\big[D_{\mathrm{KL}}(P_{Y|X} \| P_{Y|X})\big] = D_{\mathrm{KL}}(P_X \| Q_X).$$

Applying the chain rule in the other direction,

$$D_{\mathrm{KL}}(P_{XY} \| Q_{XY}) = D_{\mathrm{KL}}(P_Y \| Q_Y) + \mathbb{E}_{P_Y}\big[D_{\mathrm{KL}}(P_{X|Y} \| Q_{X|Y})\big] \ge D_{\mathrm{KL}}(P_Y \| Q_Y).$$

Combining yields $D_{\mathrm{KL}}(P_X \| Q_X) \ge D_{\mathrm{KL}}(P_Y \| Q_Y)$. □

2.1.3 Applications of DPI

Example 2.8 (DPI for mutual information)**.** If $X - Y - Z$ is a Markov chain, then

$$I(X;Y) \ge I(X;Z).$$

Proof. Using the KL representation of mutual information, $I(X;Y) = D_{\mathrm{KL}}(P_{XY} \| P_X \otimes P_Y)$. Under the Markov condition $X - Y - Z$, both P_{XZ} and $P_X \otimes P_Z$ are obtained from P_{XY} and $P_X \otimes P_Y$, respectively, by applying the same channel $P_{Z|Y}$. By DPI,

$$D_{\mathrm{KL}}(P_{XY} \| P_X \otimes P_Y) \ge D_{\mathrm{KL}}(P_{XZ} \| P_X \otimes P_Z),$$

which is $I(X;Y) \ge I(X;Z)$. □

Example 2.9 (Fano's inequality (one form))**.** If $X \sim \mathrm{Unif}([M])$ and Y is any estimator of X, then

$$\mathbb{P}(X \neq Y) \geq 1 - \frac{I(X;Y) + \log 2}{\log M}.$$

Proof. Let $A = \mathbb{1}\{X = Y\}$. Under P_{XY}, $A \sim \mathrm{Bern}(\mathbb{P}(X = Y))$. Under $P_X \otimes P_Y$, since X is uniform and independent of Y, $\mathbb{P}_{P_X \otimes P_Y}(X = Y) = 1/M$, hence $A \sim \mathrm{Bern}(1/M)$. By DPI,

$$I(X;Y) = D_{\mathrm{KL}}(P_{XY} \| P_X \otimes P_Y) \geq D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathbb{P}(X = Y)) \| \mathrm{Bern}(1/M)\big).$$

Writing $p = \mathbb{P}(X = Y) = 1 - \mathbb{P}(X \neq Y)$ and expanding the Bernoulli KL,

$$D_{\mathrm{KL}}(\mathrm{Bern}(p) \| \mathrm{Bern}(1/M)) = p \log \frac{p}{1/M} + (1-p) \log \frac{1-p}{1 - 1/M} \geq p \log M - \log 2.$$

Rearranging yields the stated bound. □

Example 2.10 (A contiguity bound)**.** For any event A,

$$\mathbb{P}_P(A) \, \log \frac{\mathbb{P}_P(A)}{e \, \mathbb{P}_Q(A)} \leq D_{\mathrm{KL}}(P \| Q).$$

In particular, if $D_{\mathrm{KL}}(P \| Q) = O(1)$ and $\mathbb{P}_Q(A) \to 0$, then $\mathbb{P}_P(A) \to 0$.

Proof. Apply DPI to the mapping $x \mapsto \mathbb{1}\{x \in A\}$. Then

$$D_{\mathrm{KL}}(P \| Q) \geq D_{\mathrm{KL}}\big(\mathrm{Bern}(\mathbb{P}_P(A)) \| \mathrm{Bern}(\mathbb{P}_Q(A))\big) \geq \mathbb{P}_P(A) \log \frac{\mathbb{P}_P(A)}{e \, \mathbb{P}_Q(A)},$$

where the last inequality is a standard lower bound on Bernoulli KL. □

2.1.4 Dual representations of KL

Theorem 2.11 (Donsker–Varadhan variational formula)**.**

$$D_{\mathrm{KL}}(P \| Q) = \sup_f \Big\{ \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f] \Big\},$$

where the supremum is over measurable f such that $\mathbb{E}_Q[e^f] < \infty$.

Proof. ($\leq$) Take $f = \log \frac{\mathrm{d}P}{\mathrm{d}Q}$. Then $\mathbb{E}_Q[e^f] = 1$ and the objective equals $\mathbb{E}_P[f] = D_{\mathrm{KL}}(P \| Q)$.

($\geq$) For any f, replace f by $f - c$ so that $\mathbb{E}_Q[e^f] = 1$. Define a probability measure $\widetilde{Q}$ by $\widetilde{Q}(\mathrm{d}x) = e^{f(x)} Q(\mathrm{d}x)$. Then

$$D_{\mathrm{KL}}(P \| Q) - \mathbb{E}_P[f] = \mathbb{E}_P \left[\log \frac{\mathrm{d}P}{e^f \, \mathrm{d}Q} \right] = D_{\mathrm{KL}}(P \| \widetilde{Q}) \geq 0.$$

Thus $\mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f] \leq D_{\mathrm{KL}}(P \| Q)$ for all f. □

Theorem 2.12 (Gibbs variational principle)**.** *For any measurable f with $\mathbb{E}_Q[e^f] < \infty$,*

$$\log \mathbb{E}_Q[e^f] = \sup_P \Big\{ \mathbb{E}_P[f] - D_{\mathrm{KL}}(P \| Q) \Big\}.$$

Proof. ($\geq$) Take $P(\mathrm{d}x) = \frac{e^{f(x)} Q(\mathrm{d}x)}{\mathbb{E}_Q[e^f]}$.

($\leq$) Follows from the Donsker–Varadhan formula. □

2.2 Applications

2.2.1 Application 1: transportation inequalities

Example 2.13 (Pinsker's inequality from Donsker–Varadhan)**.** Restrict Donsker–Varadhan to functions of the form $f = \lambda g$ with $\|g\|_\infty \le 1$. Then

$$D_{\mathrm{KL}}(P\|Q) \ge \sup_{\lambda\in\mathbb{R},\, \|g\|_\infty \le 1} \Big\{ \lambda \mathbb{E}_P[g] - \log \mathbb{E}_Q[e^{\lambda g}] \Big\}.$$

Using Hoeffding's lemma for bounded g, $\log \mathbb{E}_Q[e^{\lambda g}] \le \lambda \mathbb{E}_Q[g] + \frac{\lambda^2}{2}$. Hence

$$D_{\mathrm{KL}}(P\|Q) \ge \sup_{\lambda, \|g\|_\infty \le 1} \Big\{ \lambda(\mathbb{E}_P[g] - \mathbb{E}_Q[g]) - \tfrac{\lambda^2}{2} \Big\} = \frac{1}{2}\Big(\sup_{\|g\|_\infty \le 1} (\mathbb{E}_P[g] - \mathbb{E}_Q[g]) \Big)^2 = 2\,\mathrm{TV}(P,Q)^2.$$

This is Pinsker's inequality.

Example 2.14 (Bobkov–Gotze: a T_1 transportation inequality)**.** Let $(\mathcal{X}, d)$ be a metric space. The following are equivalent:

(1) For all 1-Lipschitz f and all $\lambda \in \mathbb{R}$,

$$\mathbb{E}_Q\big[\exp(\lambda(f - \mathbb{E}_Q[f]))\big] \le \exp\Big(\tfrac{\lambda^2 C}{2}\Big).$$

(2) For all $P \ll Q$,

$$W_1(P,Q) \le \sqrt{2C\, D_{\mathrm{KL}}(P\|Q)}.$$

Here the Wasserstein-1 distance is

$$W_1(P,Q) = \inf_{\pi\in\Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}[d(X,Y)] = \sup_{f \text{ 1-Lip}} \big\{ \mathbb{E}_P[f] - \mathbb{E}_Q[f] \big\}.$$

Proof (sketch). (1)⇒(2): Restrict Donsker–Varadhan to $f = \lambda\, f_0$ with f_0 1-Lipschitz and apply (1) to control $\log \mathbb{E}_Q[e^{\lambda f_0}]$. Optimizing over λ gives $D_{\mathrm{KL}}(P\|Q) \ge W_1(P,Q)^2/(2C)$.

(2)⇒(1): By Gibbs variational principle,

$$\log \mathbb{E}_Q[e^{\lambda(f - \mathbb{E}_Q f)}] = \sup_P \Big\{ \lambda(\mathbb{E}_P f - \mathbb{E}_Q f) - D_{\mathrm{KL}}(P\|Q) \Big\} \le \sup_P \Big\{ \lambda(\mathbb{E}_P f - \mathbb{E}_Q f) - \tfrac{(\mathbb{E}_P f - \mathbb{E}_Q f)^2}{2C} \Big\} \le \tfrac{\lambda^2 C}{2}.$$

□

2.2.2 Application 2: variational inference

Setting. Consider a model family $p_\theta(x^n, y^n)$ where both $p_\theta(x^n)$ and $p_\theta(y^n \mid x^n)$ are tractable. We observe y^n but not x^n (missing data / latent variables). The marginal likelihood

$$p_\theta(y^n) = \int p_\theta(x^n)\, p_\theta(y^n \mid x^n)\, \mathrm{d}x^n$$

is often intractable.

Theorem 2.15 (Evidence lower bound (ELBO))**.**

$$\log p_\theta(y^n) = \sup_q \; \mathbb{E}_{X^n\sim q}\left[\log \frac{p_\theta(X^n, y^n)}{q(X^n)} \right].$$

Proof. Apply the Gibbs variational principle with $f(x^n) = \log p_\theta(y^n \mid x^n)$ and base measure $p_\theta(x^n)$:

$$\begin{aligned}\log p_\theta(y^n) &= \log \mathbb{E}_{p_\theta(x^n)}\left[e^{\log p_\theta(y^n|x^n)}\right] \\ &= \sup_q \Big\{\mathbb{E}_q[\log p_\theta(y^n \mid X^n)] - D_{\mathrm{KL}}(q \| p_\theta(x^n))\Big\} \\ &= \sup_q \mathbb{E}_q\left[\log \frac{p_\theta(X^n, y^n)}{q(X^n)}\right].\end{aligned}$$

□

Example 2.16 (Ising model: variational lower bound on $\log Z$)**.** Let $y \in \{\pm 1\}^n$ and

$$p(y) = \frac{1}{Z}\exp\left(\sum_{i<j} A_{ij} y_i y_j + \sum_i b_i y_i\right).$$

Then

$$\begin{aligned}\log Z &= \log\left(2^n \mathbb{E}_{Y\sim\mathrm{Unif}(\{\pm1\}^n)} \exp\left(\sum_{i<j} A_{ij} Y_i Y_j + \sum_i b_i Y_i\right)\right) \\ &= n\log 2 + \sup_p \Big\{\mathbb{E}_p\big[\sum_{i<j} A_{ij} Y_i Y_j + \sum_i b_i Y_i\big] - D_{\mathrm{KL}}\big(p \| \mathrm{Unif}(\{\pm1\}^n)\big)\Big\} \\ &= \sup_p \Big\{\mathbb{E}_p\big[\sum_{i<j} A_{ij} Y_i Y_j + \sum_i b_i Y_i\big] + H(p)\Big\}.\end{aligned}$$

Relaxing p to a product form $p = \prod_{i=1}^n \mathrm{Bern}(p_i)$ yields a tractable lower bound.

Example 2.17 (EM algorithm as coordinate ascent on the ELBO)**.** The maximum-likelihood estimator satisfies

$$\arg\max_\theta \log p_\theta(y^n) = \arg\max_\theta \; \sup_q \; \mathbb{E}_q\left[\log \frac{p_\theta(X^n, y^n)}{q(X^n)}\right].$$

Successive maximization gives:

- **E-step:** for fixed $\theta = \theta^{(t)}$, the maximizer is $q^{(t)}(x^n) = p_{\theta^{(t)}}(x^n \mid y^n)$.
- **M-step:** for fixed $q = q^{(t)}$, update

$$\theta^{(t+1)} \in \arg\max_\theta \; \mathbb{E}_{X^n \sim q^{(t)}}\big[\log p_\theta(X^n, y^n)\big].$$

For exponential families $p_\theta(x, y) \propto \exp(\langle \theta, T(x, y)\rangle - A(\theta))$, this reduces to computing conditional sufficient statistics in the E-step and a standard MLE update in the M-step.

Example 2.18 (VAE: ELBO with reparameterization)**.** A typical VAE generative model is

$$X_i \sim \mathcal{N}(0, I), \qquad Y_i \mid X_i \sim \mathcal{N}(\mu_\theta(X_i), \sigma_\theta^2(X_i) I),$$

with $\mu_\theta, \sigma_\theta$ parameterized by neural nets. Choose an approximate posterior

$$q_\phi(x \mid y) = \mathcal{N}(\mu_\phi(y), \sigma_\phi^2(y) I).$$

Then the ELBO suggests optimizing over (θ, ϕ). In practice one:

1. Replaces expectations under q_ϕ by Monte Carlo samples $X_{ij} \sim \mathcal{N}(\mu_\phi(y_i), \sigma^2_\phi(y_i)I)$, $j = 1, \dots, M$.
2. Computes ∇_θ from the explicit form of $\log p_\theta(x, y)$.
3. Computes ∇_ϕ via the reparameterization trick: if $X \sim \mathcal{N}(\mu_\phi, \sigma^2_\phi I)$ and $\varepsilon \sim \mathcal{N}(0, I)$, then $X = \mu_\phi + \sigma_\phi \varepsilon$ and

$$\nabla_\phi \mathbb{E}[f(X)] = \nabla_\phi \mathbb{E}_\varepsilon[f(\mu_\phi + \sigma_\phi\varepsilon)] = \mathbb{E}_\varepsilon[\nabla_\phi f(\mu_\phi + \sigma_\phi\varepsilon)] \approx \frac{1}{M}\sum_{j=1}^{M} \nabla_\phi f(\mu_\phi + \sigma_\phi\varepsilon_j).$$

2.2.3 Application 3: adaptive data analysis

Problem. Let $X^n = (X_1, \dots, X_n)$ be i.i.d. from P, and let $\{\phi_t : \mathcal{X} \to \mathbb{R}\}$ be a class of functions. For fixed t define

$$P_n\phi_t := \frac{1}{n}\sum_{i=1}^{n} \phi_t(X_i), \qquad P\phi_t := \mathbb{E}[\phi_t(X)].$$

What if the index T is chosen adaptively, i.e. $T = T(X^n)$?

Example 2.19 (Russo–Zou (2016))**.** Assume each $\phi_t(X)$ is σ^2-sub-Gaussian under P. Then

$$\Big|\mathbb{E}[P_n\phi_T] - \mathbb{E}[P\phi_T]\Big| \le \sqrt{\frac{2\sigma^2}{n} I(X^n; T)}.$$

Proof. Condition on T.

$$\mathbb{E}[P_n\phi_T \mid T] = \mathbb{E}_{P_{X^n|T}}\left[\frac{1}{n}\sum_{i=1}^{n}\phi_T(X_i)\right], \qquad \mathbb{E}[P\phi_T \mid T] = \mathbb{E}_{P_{X^n}}\left[\frac{1}{n}\sum_{i=1}^{n}\phi_T(X_i)\right].$$

By Donsker–Varadhan, for any $\lambda \in \mathbb{R}$,

$$\begin{aligned} D_{\mathrm{KL}}(P_{X^n|T} \| P_{X^n}) &\ge \lambda\, \mathbb{E}\Big[P_n\phi_T \mid T\Big] - \log \mathbb{E}\left[\exp\Big(\lambda P_n\phi_T\Big)\Big| T\right] \\ &= \lambda\big(\mathbb{E}[P_n\phi_T \mid T] - \mathbb{E}[P\phi_T \mid T]\big) - \log \mathbb{E}\left[\exp\Big(\lambda(P_n\phi_T - \mathbb{E}[P\phi_T \mid T])\Big)\Big| T\right]. \end{aligned}$$

The sub-Gaussian assumption implies $\log \mathbb{E}[\exp(\lambda(P_n\phi_T - \mathbb{E}[P\phi_T \mid T])) \mid T] \le \frac{\lambda^2\sigma^2}{2n}$. Therefore,

$$D_{\mathrm{KL}}(P_{X^n|T} \| P_{X^n}) \ge \sup_{\lambda\in\mathbb{R}}\left\{\lambda\Delta - \frac{\lambda^2\sigma^2}{2n}\right\} = \frac{n\Delta^2}{2\sigma^2}, \qquad \Delta := \mathbb{E}[P_n\phi_T \mid T] - \mathbb{E}[P\phi_T \mid T].$$

Taking expectations in T and using $I(X^n; T) = \mathbb{E}\big[D_{\mathrm{KL}}(P_{X^n|T} \| P_{X^n})\big]$ yields the claim. □

2.3 Special topic: PAC-Bayes

Theorem 2.20 (PAC-Bayes inequality)**.** *Let $X \sim P$ and let $\{f_\theta : \mathcal{X} \to \mathbb{R}\}$ be a class of functions indexed by θ. Fix any prior distribution π on θ. Then with probability at least $1 - \delta$ (over $X \sim P$), for* all *distributions ρ on θ,*

$$\mathbb{E}_{\theta\sim\rho}\big[f_\theta(X) - \psi(\theta)\big] \le D_{\mathrm{KL}}(\rho\|\pi) + \log\frac{1}{\delta}, \qquad \psi(\theta) := \log \mathbb{E}_{X\sim P}\big[e^{f_\theta(X)}\big].$$

Proof. By Markov's inequality, it suffices to show

$$\mathbb{E}_{X\sim P}\left[\sup_{\rho}\exp\Big(\mathbb{E}_{\theta\sim\rho}[f_\theta(X)-\psi(\theta)]-D_{\mathrm{KL}}(\rho\|\pi)\Big)\right]\le 1.$$

By the Gibbs variational principle, the inner supremum equals $\log\mathbb{E}_{\theta\sim\pi}e^{f_\theta(X)-\psi(\theta)}$. Hence the left-hand side becomes

$$\begin{aligned}\mathbb{E}_{X\sim P}\left[\exp\left(\log\mathbb{E}_{\theta\sim\pi}e^{f_\theta(X)-\psi(\theta)}\right)\right]&=\mathbb{E}_{X\sim P}\mathbb{E}_{\theta\sim\pi}\big[e^{f_\theta(X)-\psi(\theta)}\big]\\&=\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P}\big[e^{f_\theta(X)-\psi(\theta)}\big]\right]=1.\end{aligned}$$

□

Example 2.21 (Why call it "PAC-Bayes"? (a quadratic PAC-Bayes bound))**.** Let $\mathcal{F}$ be a class of functions $f:\mathcal{X}\to[0,1]$ and let $X_1,\dots,X_n\overset{\text{i.i.d.}}{\sim}P$. Write

$$P_nf\coloneqq\frac{1}{n}\sum_{i=1}^{n}f(X_i),\qquad Pf\coloneqq\mathbb{E}_{X\sim P}[f(X)].$$

For fixed f, Hoeffding's inequality implies the usual concentration bound

$$(P_nf-Pf)^2\le\frac{1}{2n}\log\frac{2}{\delta}\qquad\text{with prob.}\ge 1-\delta.$$

PAC-Bayes gives a "soft" uniform version: fixing any prior π on $\mathcal{F}$, with probability at least $1-\delta$, *simultaneously for all* distributions ρ on $\mathcal{F}$,

$$\mathbb{E}_{f\sim\rho}\big[(P_nf-Pf)^2\big]\le\frac{D_{\mathrm{KL}}(\rho\|\pi)+\log\frac{2}{\delta}}{2n}.$$

Proof. Apply the PAC-Bayes inequality to the random variable $X^n=(X_1,\dots,X_n)$ and to the function class

$$F_f(X^n)\coloneqq\lambda\,(P_nf-Pf)^2,\qquad f\in\mathcal{F},$$

where $\lambda>0$ is a parameter. Then, with probability at least $1-\delta$, for all posteriors ρ,

$$\mathbb{E}_{f\sim\rho}\Big[\lambda(P_nf-Pf)^2-\log\mathbb{E}\exp\big(\lambda(P_nf-Pf)^2\big)\Big]\le D_{\mathrm{KL}}(\rho\|\pi)+\log\frac{1}{\delta}.\tag{2.1}$$

Now fix f. Since $f(X_i)\in[0,1]$, the centered average $Z_f\coloneqq P_nf-Pf$ is sub-Gaussian at scale $1/n$. A standard computation for sub-Gaussian random variables yields the "square-mgf" bound

$$\log\mathbb{E}\exp\big(\lambda Z_f^2\big)\le\frac{1}{2}\log\frac{1}{1-\lambda/(4n)}\qquad\text{for }\lambda<4n.$$

Plug this into (2.1). Choosing $\lambda=2n$ (so that $\lambda<4n$) gives

$$\mathbb{E}_{f\sim\rho}\big[2n\,(P_nf-Pf)^2\big]\le D_{\mathrm{KL}}(\rho\|\pi)+\log\frac{1}{\delta}+\frac{1}{2}\log 2\le D_{\mathrm{KL}}(\rho\|\pi)+\log\frac{2}{\delta}.$$

Dividing by $2n$ yields the claim. □

Example 2.22 (A Gaussian norm bound via PAC-Bayes)**.** If $X \sim \mathcal{N}(0, \Sigma)$, then with probability at least $1 - \delta$,

$$\|X\|_2 \leq \sqrt{\operatorname{Tr}(\Sigma)} + \sqrt{2 \, \|\Sigma\|_{\mathrm{op}} \log \frac{1}{\delta}}.$$

Proof. We start from the dual characterization

$$\|X\|_2 = \sup_{\|v\|_2 = 1} \langle v, X \rangle .$$

Fix parameters $\lambda > 0$ and $\sigma^2 > 0$. Let the prior on $\theta \in \mathbb{R}^d$ be $\pi = \mathcal{N}(0, \sigma^2 I)$. For each v with $\|v\|_2 = 1$, define a posterior

$$\rho_v := \mathcal{N}(v, \sigma^2 I).$$

Apply PAC-Bayes with the function $f_\theta(X) = \lambda \langle \theta, X \rangle$. Then, with probability at least $1 - \delta$, simultaneously for all v,

$$\mathbb{E}_{\theta \sim \rho_v}\Big[\lambda \langle \theta, X \rangle - \log \mathbb{E} \exp\big(\lambda \langle \theta, X \rangle\big)\Big] \leq D_{\mathrm{KL}}(\rho_v \| \pi) + \log \frac{1}{\delta}. \tag{2.2}$$

We now compute the three terms explicitly. First, since $X \sim \mathcal{N}(0, \Sigma)$,

$$\log \mathbb{E} \exp\big(\lambda \langle \theta, X \rangle\big) = \frac{\lambda^2}{2} \, \theta^\top \Sigma \theta.$$

Second,

$$D_{\mathrm{KL}}(\rho_v \| \pi) = D_{\mathrm{KL}}\big(\mathcal{N}(v, \sigma^2 I) \, \| \, \mathcal{N}(0, \sigma^2 I)\big) = \frac{\|v\|_2^2}{2\sigma^2} = \frac{1}{2\sigma^2}.$$

Third, $\mathbb{E}_{\theta \sim \rho_v}[\langle \theta, X \rangle] = \langle \mathbb{E}\theta, X \rangle = \langle v, X \rangle$ and

$$\mathbb{E}_{\theta \sim \rho_v}[\theta^\top \Sigma \theta] = v^\top \Sigma v + \sigma^2 \operatorname{Tr}(\Sigma).$$

Plugging into (2.2) yields, for all v with $\|v\|_2 = 1$,

$$\lambda \langle v, X \rangle - \frac{\lambda^2}{2}\big(v^\top \Sigma v + \sigma^2 \operatorname{Tr}(\Sigma)\big) \leq \frac{1}{2\sigma^2} + \log \frac{1}{\delta},$$

or equivalently

$$\langle v, X \rangle \leq \frac{\lambda}{2}\big(v^\top \Sigma v + \sigma^2 \operatorname{Tr}(\Sigma)\big) + \frac{1}{\lambda}\Big(\frac{1}{2\sigma^2} + \log \frac{1}{\delta}\Big). \tag{2.3}$$

Now optimize over σ^2. For fixed λ, the σ^2-dependent part in (2.3) is

$$\frac{\lambda}{2} \sigma^2 \operatorname{Tr}(\Sigma) + \frac{1}{\lambda} \cdot \frac{1}{2\sigma^2}.$$

This is minimized at

$$\sigma^2 = \frac{1}{\lambda \sqrt{\operatorname{Tr}(\Sigma)}},$$

in which case the minimum value equals $\sqrt{\operatorname{Tr}(\Sigma)}$. Hence (2.3) becomes

$$\langle v, X \rangle \leq \sqrt{\operatorname{Tr}(\Sigma)} + \frac{\lambda}{2} v^\top \Sigma v + \frac{\log(1/\delta)}{\lambda}.$$

Using $v^\top \Sigma v \leq \|\Sigma\|_{\mathrm{op}}$ and taking the supremum over $\|v\|_2 = 1$ gives

$$\|X\|_2 \leq \sqrt{\mathrm{Tr}(\Sigma)} + \frac{\lambda}{2} \|\Sigma\|_{\mathrm{op}} + \frac{\log(1/\delta)}{\lambda}.$$

Finally, optimize over $\lambda > 0$ by choosing

$$\lambda = \sqrt{\frac{2\log(1/\delta)}{\|\Sigma\|_{\mathrm{op}}}},$$

which yields

$$\|X\|_2 \leq \sqrt{\mathrm{Tr}(\Sigma)} + \sqrt{2 \|\Sigma\|_{\mathrm{op}} \log \frac{1}{\delta}}.$$

□

Example 2.23 (Sample covariance (effective rank bound))**.** Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i X_i^\top] = \Sigma$, and assume $\langle v, X_i \rangle$ is sub-Gaussian with variance proxy $v^\top \Sigma v$ for every $v \in \mathbb{R}^d$. Let $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. Then with probability at least $1 - \delta$,

$$\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}} \leq C \|\Sigma\|_{\mathrm{op}} \left(\sqrt{\frac{r(\Sigma) + \log(1/\delta)}{n}} + \frac{r(\Sigma) + \log(1/\delta)}{n} \right),$$

where $r(\Sigma) = \mathrm{Tr}(\Sigma)/\|\Sigma\|_{\mathrm{op}}$ is the *effective rank* and C is a universal constant.

Proof. Throughout the proof, C denotes a large universal constant which may change from line to line.

Step 1: reduce to bilinear forms. Recall

$$\left\|\widehat{\Sigma} - \Sigma\right\|_{\mathrm{op}} = \sup_{\|u\|_2 = \|v\|_2 = 1} u^\top (\widehat{\Sigma} - \Sigma) v.$$

Fix u, v with $\|u\|_2 = \|v\|_2 = 1$.

Step 2: construct truncated-Gaussian posteriors. Let f_u be the density of $\mathcal{N}(u, \sigma^2 I)$ conditioned on the event

$$(x - u)^\top \Sigma (x - u) \leq r^2,$$

and define the product posterior on $(\theta, \theta') \in \mathbb{R}^d \times \mathbb{R}^d$ by

$$\rho_{u,v} := f_u \otimes f_v.$$

By symmetry of the conditioning set around u (resp. v), $\mathbb{E}_{\theta \sim f_u}[\theta] = u$ and $\mathbb{E}_{\theta' \sim f_v}[\theta'] = v$. Therefore,

$$\mathbb{E}_{(\theta, \theta') \sim \rho_{u,v}} \left[\theta^\top (\widehat{\Sigma} - \Sigma) \theta'\right] = u^\top (\widehat{\Sigma} - \Sigma) v.$$

Let

$$p := \mathbb{P}\big(Z^\top \Sigma Z \leq r^2\big), \qquad Z \sim \mathcal{N}(0, \sigma^2 I).$$

By Markov's inequality,

$$p \geq 1 - \frac{\mathbb{E}[Z^\top \Sigma Z]}{r^2} = 1 - \frac{\sigma^2 \,\mathrm{Tr}(\Sigma)}{r^2}.$$

Step 3: compute the KL term. Let the prior be

$$\pi := \mathcal{N}(0,\sigma^2 I)\otimes\mathcal{N}(0,\sigma^2 I).$$

Write φ_u for the density of $\mathcal{N}(u,\sigma^2 I)$. Then

$$f_u(x) = \frac{\varphi_u(x)\,\mathbb{1}\{(x-u)^\top\Sigma(x-u)\le r^2\}}{p}.$$

A direct calculation gives

$$D_{\mathrm{KL}}\big(f_u\|\mathcal{N}(0,\sigma^2 I)\big) = D_{\mathrm{KL}}\big(\mathcal{N}(u,\sigma^2 I)\|\mathcal{N}(0,\sigma^2 I)\big) + \log\frac{1}{p} = \frac{\|u\|_2^2}{2\sigma^2} + \log\frac{1}{p} = \frac{1}{2\sigma^2} + \log\frac{1}{p}.$$

Hence

$$D_{\mathrm{KL}}(\rho_{u,v}\|\pi) = \frac{1}{\sigma^2} + 2\log\frac{1}{p}.$$

Step 4: apply PAC-Bayes. Apply PAC-Bayes to the random sample $X^n = (X_1,\dots,X_n)$, parameter (θ,θ'), and the function

$$F_{\theta,\theta'}(X^n) := \lambda\,\theta^\top(\widehat{\Sigma}-\Sigma)\theta',$$

where $\lambda > 0$. Then with probability at least $1-\delta$,

$$\mathbb{E}_{(\theta,\theta')\sim\rho_{u,v}}\Big[\lambda\theta^\top(\widehat{\Sigma}-\Sigma)\theta' - \log\mathbb{E}\exp\big(\lambda\theta^\top(\widehat{\Sigma}-\Sigma)\theta'\big)\Big] \le D_{\mathrm{KL}}(\rho_{u,v}\|\pi) + \log\frac{1}{\delta}. \tag{2.4}$$

Step 5: bound the log-mgf term (Bernstein-type control). Under the stated sub-Gaussian assumption on $\langle w, X_i\rangle$, one has the estimate (as in the handwritten notes)

$$\log\mathbb{E}\exp\big(\lambda\theta^\top(\widehat{\Sigma}-\Sigma)\theta'\big) \le \frac{C\lambda^2}{n}\big(\theta^\top\Sigma\theta + \theta'^\top\Sigma\theta'\big)^2, \qquad \text{for } \lambda \le \frac{n}{C\big(\theta^\top\Sigma\theta+\theta'^\top\Sigma\theta'\big)}. \tag{2.5}$$

Using (2.5) in (2.4) and dividing by λ yields

$$u^\top(\widehat{\Sigma}-\Sigma)v \le \frac{C\lambda}{n}\mathbb{E}_{\rho_{u,v}}\big(\theta^\top\Sigma\theta+\theta'^\top\Sigma\theta'\big)^2 + \frac{1}{\lambda}\Big(\frac{1}{\sigma^2} + 2\log\frac{1}{p} + \log\frac{1}{\delta}\Big) \tag{2.6}$$

provided λ satisfies the condition in (2.5).

Step 6: control $\theta^\top\Sigma\theta$ under the truncation. If $\theta\sim f_u$, then $(\theta-u)^\top\Sigma(\theta-u)\le r^2$. Hence, using the triangle inequality in the seminorm $x\mapsto\sqrt{x^\top\Sigma x}$,

$$\sqrt{\theta^\top\Sigma\theta} \le \sqrt{u^\top\Sigma u} + \sqrt{(\theta-u)^\top\Sigma(\theta-u)} \le \sqrt{\|\Sigma\|_{\mathrm{op}}} + r,$$

so

$$\theta^\top\Sigma\theta \le (\sqrt{\|\Sigma\|_{\mathrm{op}}} + r)^2.$$

The same bound holds for $\theta'\sim f_v$. Consequently,

$$\big(\theta^\top\Sigma\theta + \theta'^\top\Sigma\theta'\big)^2 \le C\,(\sqrt{\|\Sigma\|_{\mathrm{op}}} + r)^4.$$

Also, the condition in (2.5) is ensured if

$$\lambda \leq \frac{n}{C(\sqrt{\|\Sigma\|_{\mathrm{op}}}+r)^2}.$$

Taking the supremum over u, v in (2.6), we obtain

$$\left\|\widehat{\Sigma}-\Sigma\right\|_{\mathrm{op}} \leq \frac{C\lambda}{n}(\sqrt{\|\Sigma\|_{\mathrm{op}}}+r)^4+\frac{1}{\lambda}\Big(\frac{1}{\sigma^2}+2\log\frac{1}{p}+\log\frac{1}{\delta}\Big) \tag{2.7}$$

for all $\lambda \leq \frac{n}{C(\sqrt{\|\Sigma\|_{\mathrm{op}}}+r)^2}$.

Step 7: choose r and σ^2. Choose

$$r^2 = 2\|\Sigma\|_{\mathrm{op}}, \qquad \sigma^2 = \frac{\|\Sigma\|_{\mathrm{op}}}{\mathrm{Tr}(\Sigma)} = \frac{1}{r(\Sigma)}.$$

Then

$$p \geq 1-\frac{\sigma^2\,\mathrm{Tr}(\Sigma)}{r^2} = 1-\frac{\|\Sigma\|_{\mathrm{op}}}{2\|\Sigma\|_{\mathrm{op}}} = \frac{1}{2},$$

so $\log\frac{1}{p} \leq \log 2$. Moreover, $(\sqrt{\|\Sigma\|_{\mathrm{op}}}+r)^4 \asymp \|\Sigma\|_{\mathrm{op}}^2$. Absorbing constants into C, (2.7) becomes

$$\left\|\widehat{\Sigma}-\Sigma\right\|_{\mathrm{op}} \leq C\left(\frac{\lambda}{n}\|\Sigma\|_{\mathrm{op}}^2+\frac{1}{\lambda}\Big(r(\Sigma)+\log\frac{1}{\delta}\Big)\right) \qquad \text{for } \lambda \leq \frac{n}{C\|\Sigma\|_{\mathrm{op}}}.$$

Step 8: optimize over λ. Let $a := r(\Sigma)+\log\frac{1}{\delta}$. If $a/n \leq 1$, choose

$$\lambda \asymp \frac{n}{\|\Sigma\|_{\mathrm{op}}}\sqrt{\frac{a}{n}}.$$

If $a/n > 1$, choose $\lambda \asymp \frac{n}{\|\Sigma\|_{\mathrm{op}}}$. In both cases,

$$\left\|\widehat{\Sigma}-\Sigma\right\|_{\mathrm{op}} \leq C\,\|\Sigma\|_{\mathrm{op}}\left(\sqrt{\frac{a}{n}}+\frac{a}{n}\right) = C\,\|\Sigma\|_{\mathrm{op}}\left(\sqrt{\frac{r(\Sigma)+\log(1/\delta)}{n}}+\frac{r(\Sigma)+\log(1/\delta)}{n}\right),$$

which is the claimed bound. □

Lecture 3: f-divergences

3.1 f-divergence: definition and examples

Definition 3.1 (f-divergence (Csiszár, 1963))**.** Let $f : (0, \infty) \to \mathbb{R}$ be convex with $f(1) = 0$. For two probability measures P, Q on the same measurable space with $P \ll Q$, the f-divergence is

$$D_f(P\|Q) \;\triangleq\; \mathbb{E}_Q\Big[f\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\Big].$$

Remark 3.2 (Normalizations and the case $P \not\ll Q$).

1. Some definitions additionally assume $f'(1) = 0$. This is without loss of generality: if $c \in \mathbb{R}$ then $f(x)$ and $f(x) + c(x-1)$ induce the same f-divergence, since $\mathbb{E}_Q[\frac{\mathrm{d}P}{\mathrm{d}Q} - 1] = 0$.

2. If $\frac{\mathrm{d}P}{\mathrm{d}Q} = 0$, define $f(0) \triangleq f(0+)$. If $P \not\ll Q$, pick a dominating measure μ with densities $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$ and $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$, and define

$$D_f(P\|Q) \;\triangleq\; \int_{\{q>0\}} q\, f\Big(\frac{p}{q}\Big)\, \mathrm{d}\mu \;+\; f(\infty)\, P(q=0), \qquad f(\infty) \triangleq \lim_{x\to\infty} \frac{f(x)}{x}.$$

Examples

Below are standard choices of f and the resulting divergences.

1. **Total variation.** $f(x) = \frac{1}{2}|x-1|$.

$$D_f(P\|Q) = \mathrm{TV}(P,Q) = \frac{1}{2}\int |\,\mathrm{d}P - \,\mathrm{d}Q|.$$

2. **Squared Hellinger distance.** $f(x) = (\sqrt{x} - 1)^2$.

$$D_f(P\|Q) = \mathrm{H}^2(P,Q) = \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2.$$

3. **Kullback–Leibler divergence.** $f(x) = x \log x$.

$$D_f(P\|Q) = D_{\mathrm{KL}}(P\|Q) = \int \log\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\, \mathrm{d}P.$$

4. **χ^2-divergence.** $f(x) = (x-1)^2$.

$$D_f(P\|Q) = \chi^2(P\|Q) = \int \frac{(\,\mathrm{d}P - \,\mathrm{d}Q)^2}{\mathrm{d}Q} = \mathbb{E}_Q\Big[\Big(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\Big)^2\Big].$$

5. **Le Cam distance.** $f(x) = \frac{(1-x)^2}{2(1+x)}$.

$$D_f(P\|Q) = \mathrm{LC}(P,Q) = \frac{1}{2}\int \frac{(\,\mathrm{d}P - \,\mathrm{d}Q)^2}{\mathrm{d}P + \,\mathrm{d}Q}.$$

6. **Jensen–Shannon divergence.** $f(x) = x\log x + (x+1)\log\frac{2}{x+1}$.

$$D_f(P\|Q) = \mathrm{JS}(P,Q) = D_{\mathrm{KL}}\Big(P \,\Big\|\, \frac{P+Q}{2}\Big) + D_{\mathrm{KL}}\Big(Q \,\Big\|\, \frac{P+Q}{2}\Big).$$

3.2 Basic properties

Theorem 3.3 (Non-negativity)**.** *For any convex f with $f(1) = 0$ and any $P \ll Q$,*

$$D_f(P\|Q) \geq 0.$$

Proof. By Jensen's inequality,

$$D_f(P\|Q) = \mathbb{E}_Q\Big[f\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\Big] \geq f\Big(\mathbb{E}_Q\Big[\frac{\mathrm{d}P}{\mathrm{d}Q}\Big]\Big) = f(1) = 0.$$

□

Theorem 3.4 (Joint convexity)**.** *The map $(P,Q) \mapsto D_f(P\|Q)$ is jointly convex.*

Proof. Fix a convex $f : (0,\infty) \to \mathbb{R}$ and define its *perspective transform*

$$\psi(x,y) \triangleq y\, f\left(\frac{x}{y}\right), \qquad (x,y) \in \mathbb{R}^2_{>0}.$$

Assume first that f is twice differentiable and write $t \triangleq x/y$. A direct calculation gives

$$\frac{\partial^2\psi}{\partial x^2} = \frac{1}{y}\, f''(t), \qquad \frac{\partial^2\psi}{\partial x\,\partial y} = -\frac{x}{y^2}\, f''(t), \qquad \frac{\partial^2\psi}{\partial y^2} = \frac{x^2}{y^3}\, f''(t).$$

Hence

$$\nabla^2\psi(x,y) = \begin{pmatrix} \frac{1}{y}f''(x/y) & -\frac{x}{y^2}f''(x/y) \\ -\frac{x}{y^2}f''(x/y) & \frac{x^2}{y^3}f''(x/y) \end{pmatrix} \succeq 0,$$

since $f'' \geq 0$ and the matrix has rank one. Therefore ψ is convex on $\mathbb{R}^2_{>0}$. (For a general convex f, the same conclusion holds by standard approximation.)

Now let P, Q admit densities p, q with respect to a common dominating measure μ. For $\lambda \in [0,1]$, define

$$p_\lambda \triangleq \lambda p_1 + (1-\lambda)p_2, \qquad q_\lambda \triangleq \lambda q_1 + (1-\lambda)q_2.$$

Pointwise convexity of ψ gives

$$q_\lambda\, f\left(\frac{p_\lambda}{q_\lambda}\right) = \psi(p_\lambda, q_\lambda) \leq \lambda\psi(p_1,q_1) + (1-\lambda)\psi(p_2,q_2) = \lambda q_1 f\left(\frac{p_1}{q_1}\right) + (1-\lambda)q_2 f\left(\frac{p_2}{q_2}\right).$$

Integrating over μ yields

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

which is joint convexity. □

Theorem 3.5 (Data processing inequality)**.** *Let P_X, Q_X be distributions on $\mathcal{X}$ and let $P_{Y|X}$ be a Markov kernel. If P_Y, Q_Y are the induced marginals, then*

$$D_f(P_X \| Q_X) \geq D_f(P_Y \| Q_Y).$$

Proof. Write $L \triangleq \frac{\mathrm{d}P_X}{\mathrm{d}Q_X}$. For the channel (Markov kernel) $P_{Y|X}$, the induced marginals satisfy

$$\mathrm{d}P_Y(y) = \int P_{Y|X}(y \mid x)\ \mathrm{d}P_X(x), \qquad \mathrm{d}Q_Y(y) = \int P_{Y|X}(y \mid x)\ \mathrm{d}Q_X(x).$$

Using $\mathrm{d}P_X = L\ \mathrm{d}Q_X$, we can rewrite

$$\mathrm{d}P_Y(y) = \int P_{Y|X}(y \mid x)\, L(x)\ \mathrm{d}Q_X(x).$$

Consequently, whenever $\mathrm{d}Q_Y(y) > 0$,

$$\frac{\mathrm{d}P_Y}{\mathrm{d}Q_Y}(y) = \frac{\int P_{Y|X}(y \mid x)\, L(x)\ \mathrm{d}Q_X(x)}{\int P_{Y|X}(y \mid x)\ \mathrm{d}Q_X(x)} = \mathbb{E}_Q[L(X) \mid Y = y],$$

where the conditional expectation is with respect to the joint law $Q_{XY} \triangleq Q_X P_{Y|X}$. Therefore

$$D_f(P_Y \| Q_Y) = \mathbb{E}_{Q_Y}\Big[f\big(\mathbb{E}_Q[L \mid Y]\big)\Big].$$

By Jensen's inequality (since f is convex),

$$\mathbb{E}_{Q_Y}\Big[f\big(\mathbb{E}_Q[L \mid Y]\big)\Big] \leq \mathbb{E}_{Q_Y}\Big[\mathbb{E}_Q\big[f(L) \mid Y\big]\Big] = \mathbb{E}_Q[f(L)] = \mathbb{E}_{Q_X}\Big[f\Big(\frac{\mathrm{d}P_X}{\mathrm{d}Q_X}\Big)\Big] = D_f(P_X \| Q_X).$$

This proves $D_f(P_X \| Q_X) \geq D_f(P_Y \| Q_Y)$. □

$$\begin{tikzcd}
P_X \arrow[r, "P_{Y|X}"] & P_Y \\
Q_X \arrow[r, "P_{Y|X}"'] & Q_Y
\end{tikzcd}$$

3.3 Why f-divergence? Binary hypothesis testing

Consider simple hypothesis testing:

$$H_0 : \ X \sim P, \qquad H_1 : \ X \sim Q,$$

with a (possibly randomized) test $T : \mathcal{X} \to \{0, 1\}$. The errors are

$$\text{Type I: } P(T(X) = 1), \qquad \text{Type II: } Q(T(X) = 0).$$

Theorem 3.6 (Total variation and the best sum of errors)**.**

$$\inf_T \big(P(T(X) = 1) + Q(T(X) = 0)\big) = 1 - \mathrm{TV}(P, Q).$$

Proof. First recall the standard identity

$$\mathrm{TV}(P,Q) = \sup_A \bigl(P(A) - Q(A)\bigr), \tag{3.1}$$

where the supremum ranges over measurable sets A. Indeed, if P, Q have densities p, q w.r.t. a common dominating measure μ, then

$$\mathrm{TV}(P,Q) = \frac{1}{2}\int |p-q| \, \mathrm{d}\mu = \int_{\{p\ge q\}} (p-q)\, \mathrm{d}\mu = P(A^*) - Q(A^*), \qquad A^* \triangleq \{p \ge q\},$$

which proves (3.1).

Now fix any (possibly randomized) test $T : \mathcal{X} \to \{0,1\}$ and let

$$A \triangleq \{x : T(x) = 0\}.$$

Then

$$P(T(X)=1) + Q(T(X)=0) = P(A^c) + Q(A) = 1 - P(A) + Q(A) = 1 - (P(A) - Q(A)).$$

Taking the infimum over tests T is therefore equivalent to taking the supremum over sets A:

$$\inf_T \bigl(P(T(X)=1) + Q(T(X)=0)\bigr) = 1 - \sup_A (P(A) - Q(A)) = 1 - \mathrm{TV}(P,Q),$$

where we used (3.1). Finally, equality is attained by the deterministic test $T^*(x) = \mathbb{1}\{x \notin A^*\}$ for any set A^* achieving the supremum in (3.1). □

Remark 3.7 (Interpretation of total variation).

1. $\mathrm{TV}(P,Q) = 0$ iff $P = Q$ (totally indistinguishable).
2. $\mathrm{TV}(P,Q) = 1$ iff $P \perp Q$ (perfectly distinguishable).
3. $\mathrm{TV}(P,Q) < 1$ means partially indistinguishable.

This quantity is central in minimax lower bounds.

3.3.1 Why not just total variation? Tensorization

1. $\mathrm{TV}(P,Q)$ can be hard to compute.
2. TV does not tensorize well: in general,

$$\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \le n\, \mathrm{TV}(P,Q)$$

is the best possible inequality in full generality, but it is often loose.

Example. How large is $\mathrm{TV}(\mathrm{Ber}(\frac{1}{2})^{\otimes n}, \mathrm{Ber}(\frac{1}{2}+\delta)^{\otimes n})$? The bound $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \le n\mathrm{TV}(P,Q)$ yields an $n\delta$-type upper bound, whereas Pinsker's inequality gives

$$\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) \le \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P^{\otimes n} \| Q^{\otimes n})} = \sqrt{\tfrac{n}{2} D_{\mathrm{KL}}(P\|Q)} = O(\sqrt{n}\,\delta),$$

which is much tighter for small δ.

3.3.2 Popular f-divergences that *do* tensorize

For product measures $\bigotimes_i P_i$ and $\bigotimes_i Q_i$:

1. **Squared Hellinger:**

$$1 - \frac{1}{2}\mathrm{H}^2\Big(\bigotimes_i P_i, \bigotimes_i Q_i\Big) = \prod_i \Big(1 - \frac{1}{2}\mathrm{H}^2(P_i, Q_i)\Big).$$

2. **KL:**

$$D_{\mathrm{KL}}\Big(\bigotimes_i P_i \,\Big\|\, \bigotimes_i Q_i\Big) = \sum_i D_{\mathrm{KL}}(P_i \| Q_i).$$

3. χ^2**:**

$$\chi^2\Big(\bigotimes_i P_i \,\Big\|\, \bigotimes_i Q_i\Big) + 1 = \prod_i \big(\chi^2(P_i \| Q_i) + 1\big).$$

Remark 3.8 (Optional: Rényi divergences unify these). Rényi divergence (order $\lambda \neq 1$) is

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda - 1} \log \mathbb{E}_Q\Big[\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)^\lambda\Big].$$

It tensorizes:

$$D_\lambda\big(\bigotimes_i P_i \| \bigotimes_i Q_i\big) = \sum_i D_\lambda(P_i \| Q_i).$$

For $\lambda = \frac{1}{2}, 1, 2$ this relates to Hellinger affinity, KL, and χ^2, respectively.

3.4 Similarities and differences between f-divergences

3.4.1 Locally χ^2-like

Assume $f''(1)$ exists and P and Q are "close" (heuristically, $\frac{\mathrm{d}P}{\mathrm{d}Q} \approx 1$). Then a Taylor expansion gives

$$D_f(P\|Q) = \mathbb{E}_Q\Big[f\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\Big] \approx \mathbb{E}_Q\Big[f(1) + f'(1)\Big(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\Big) + \frac{f''(1)}{2}\Big(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\Big)^2\Big].$$

Since $f(1) = 0$ and $\mathbb{E}_Q[\frac{\mathrm{d}P}{\mathrm{d}Q} - 1] = 0$,

$$D_f(P\|Q) \approx \frac{f''(1)}{2}\, \chi^2(P\|Q).$$

3.4.2 In parametric models: Fisher information

Let $(P_\theta)_{\theta\in\Theta}$ be a regular parametric model with $\theta \in \mathbb{R}^d$ and (for a dominating μ) densities $f_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$. For $h \in \mathbb{R}^d$ and small t,

$$\chi^2(P_{\theta+th}\|P_\theta) = \int \frac{(f_{\theta+th} - f_\theta)^2}{f_\theta}\, \mathrm{d}\mu \approx t^2\, h^\top \Big(\int \frac{\dot f_\theta(x) \dot f_\theta(x)^\top}{f_\theta(x)}\, \mathrm{d}\mu(x)\Big) h \;=\; t^2\, h^\top I(\theta) h,$$

where $\dot f_\theta(x) = \nabla_\theta f_\theta(x)$ and $I(\theta) \in \mathbb{R}^{d\times d}$ is the Fisher information matrix:

$$I(\theta) = \int \frac{\dot f_\theta(x) \dot f_\theta(x)^\top}{f_\theta(x)}\, \mathrm{d}\mu(x) = \mathbb{E}\big[(\nabla_\theta \log f_\theta(X))(\nabla_\theta \log f_\theta(X))^\top\big] = \mathbb{E}\big[-\nabla_\theta^2 \log f_\theta(X)\big].$$

3.5 f-divergence as "average statistical information"

3.5.1 Bayes error and statistical information

In binary hypothesis testing with prior $\mathbb{P}(H_0) = \pi \in (0,1)$, the Bayes error is

$$B_\pi(P,Q) = \inf_T \Big(\pi\, P(T(X)=1) + (1-\pi)\, Q(T(X)=0)\Big) = \int (\pi \; \mathrm{d}P) \wedge \big((1-\pi) \; \mathrm{d}Q\big),$$

where $x \wedge y \triangleq \min\{x,y\}$.

The associated *statistical information* is the improvement from prior to posterior:

$$I_\pi(P,Q) \triangleq \pi \wedge (1-\pi) - B_\pi(P,Q).$$

One can check that $I_\pi(P,Q)$ is an f-divergence:

$$I_\pi(P,Q) = \mathbb{E}_Q\Big[f_\pi\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\Big], \qquad f_\pi(t) \triangleq \pi \wedge (1-\pi) - \big(\pi t\big) \wedge (1-\pi).$$

Theorem 3.9 (Liese–Vajda, 2006)**.** *For any f-divergence, there exists a (finite) measure Γ_f on $(0,1)$ such that for all P,Q,*

$$D_f(P\|Q) = \int_0^1 I_\pi(P,Q)\, \Gamma_f(\, \mathrm{d}\pi).$$

Remark 3.10. Every f-divergence is an *average* statistical information, with different weights placed on π.

Proof. Assume $f(1)=0$ and, without loss of generality, $f'(1)=0$. For a convex f, its (distributional) second derivative is a nonnegative measure $f''(\, \mathrm{d}x)$ on $(0,\infty)$. (When $f \in C^2$, one has $f''(\, \mathrm{d}x) = f''(x)\, \mathrm{d}x$.)

A standard calculus identity (which one can check first for $f \in C^2$ and then extend by approximation) is:

$$f(t) = \int_1^t (t-x)\, f''(\, \mathrm{d}x) = \int_0^1 (x - t\wedge x)\, f''(\, \mathrm{d}x) + \int_1^\infty (t - t\wedge x)\, f''(\, \mathrm{d}x). \tag{3.2}$$

Define

$$\tilde{f}(t) \triangleq \int_0^1 (x - t\wedge x)\, f''(\, \mathrm{d}x) + \int_1^\infty (1 - t\wedge x)\, f''(\, \mathrm{d}x).$$

Then, for any $t>0$, subtracting from (3.2) gives

$$(f-\tilde{f})(t) = \int_1^\infty (t-1)\, f''(\, \mathrm{d}x) = (t-1)\, f(\infty),$$

which is affine in t. Hence for $L \triangleq \frac{\mathrm{d}P}{\mathrm{d}Q}$ (so that $\mathbb{E}_Q[L]=1$),

$$\mathbb{E}_Q\big[(f-\tilde{f})(L)\big] = f(\infty)\, \mathbb{E}_Q[L-1] = 0,$$

and therefore

$$D_f(P\|Q) = \mathbb{E}_Q[f(L)] = \mathbb{E}_Q[\tilde{f}(L)]. \tag{3.3}$$

Next, for any $x > 0$ and any $t > 0$,

$$(1 \wedge x) - (t \wedge x) = (1+x)\left(\frac{1}{1+x} \wedge \frac{x}{1+x} - t\frac{1}{1+x} \wedge \frac{x}{1+x}\right) = (1+x)\, f_{\frac{1}{1+x}}(t), \tag{3.4}$$

where $f_\pi(t) = \pi \wedge (1-\pi) - (\pi t) \wedge (1-\pi)$ is the f generating I_π. Combining (3.3) and (3.4) gives

$$\int_0^\infty (1+x)\, I_{\frac{1}{1+x}}(P,Q)\, f''(\,\mathrm{d}x) = \mathbb{E}_Q\Big[\int_0^\infty (1+x)\, f_{\frac{1}{1+x}}(L)\, f''(\,\mathrm{d}x)\Big] = \mathbb{E}_Q[\tilde{f}(L)] = D_f(P\|Q).$$

Finally, define Γ_f as the pushforward of the measure $(1+x)f''(\,\mathrm{d}x)$ under the map

$$(0,\infty) \ni x \longmapsto \frac{1}{1+x} \in (0,1).$$

Then the last display is exactly

$$D_f(P\|Q) = \int_0^1 I_\pi(P,Q)\, \Gamma_f(\,\mathrm{d}\pi),$$

which proves the theorem. □

3.6 Different guarantees on contiguity

Definition 3.11 (Contiguity)**.** A sequence of measures $\{P_n\}$ is *contiguous* with respect to $\{Q_n\}$ (written $\{P_n\} \triangleleft \{Q_n\}$) if for any events A_n,

$$Q_n(A_n) \to 0 \quad \Longrightarrow \quad P_n(A_n) \to 0.$$

Remark 3.12.

- **TV.** If $\mathrm{TV}(P_n, Q_n) \to 0$ then $\{P_n\} \triangleleft \{Q_n\}$. Indeed, for any event A_n,

$$P_n(A_n) = Q_n(A_n) + \big(P_n(A_n) - Q_n(A_n)\big) \le Q_n(A_n) + \sup_A |P_n(A) - Q_n(A)| = Q_n(A_n) + \mathrm{TV}(P_n, Q_n).$$

  So $Q_n(A_n) \to 0$ and $\mathrm{TV}(P_n, Q_n) \to 0$ imply $P_n(A_n) \to 0$.

- **KL.** If $D_{\mathrm{KL}}(P_n\|Q_n) \le C$, contiguity already holds. In fact, for any event A_n,

$$P_n(A_n)\, \log \frac{P_n(A_n)}{e\, Q_n(A_n)} \le D_{\mathrm{KL}}(P_n\|Q_n) \le C. \tag{3.5}$$

  Proof of (3.5)*:* Let $p \triangleq P_n(A_n)$ and $q \triangleq Q_n(A_n)$. Apply the data processing inequality for KL to the mapping $x \mapsto \mathbb{1}\{x \in A_n\}$ to get

$$D_{\mathrm{KL}}(P_n\|Q_n) \ge D_{\mathrm{KL}}\big(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(q)\big) = p \log \frac{p}{q} + (1-p)\log\frac{1-p}{1-q}.$$

  Using the inequality $\log u \ge 1 - \frac{1}{u}$ (valid for all $u > 0$) with $u = \frac{1-p}{1-q}$, we have

$$(1-p)\log\frac{1-p}{1-q} \ge (1-p)\left(1 - \frac{1-q}{1-p}\right) = q - p.$$

  Therefore

$$D_{\mathrm{KL}}(P_n\|Q_n) \ge p\log\frac{p}{q} + q - p = p\log\frac{p}{e\,q} + q \ge p\log\frac{p}{e\,q},$$

which is exactly (3.5). To conclude contiguity, suppose $q_n \to 0$. If $p_n \not\to 0$, then there exists $\varepsilon > 0$ and infinitely many n such that $p_n \geq \varepsilon$. For those n,

$$C \geq p_n \log \frac{p_n}{eq_n} \geq \varepsilon \log \frac{\varepsilon}{eq_n} \xrightarrow[n\to\infty]{} \infty,$$

a contradiction. Hence $p_n \to 0$.

- χ^2**.** If $\chi^2(P_n\|Q_n) \leq C$, one gets a stronger quantitative control. Let $p \triangleq P_n(A_n)$ and $q \triangleq Q_n(A_n)$. By data processing for χ^2 under the same indicator map,

$$\chi^2(P_n\|Q_n) \geq \chi^2\big(\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(q)\big) = \frac{(p-q)^2}{q(1-q)}.$$

Thus

$$\frac{(P_n(A_n) - Q_n(A_n))^2}{Q_n(A_n)(1 - Q_n(A_n))} \leq \chi^2(P_n\|Q_n) \leq C.$$

Since $1 - q \leq 1$, this implies $(p-q)^2 \leq Cq$, and hence

$$P_n(A_n) \leq Q_n(A_n) + \sqrt{C\, Q_n(A_n)}.$$

Different f-divergences give different “powers” for establishing contiguity results, due to different growth of $f(t)$ as $t \to \infty$. In this context, a popular choice is to upper bound $\chi^2(P_n\|Q_n)$, known as the *second moment method*.

3.7 Dual representations of f-divergence

Definition 3.13 (Convex conjugate)**.** For a convex function f on $\mathbb{R}$, its convex conjugate is

$$f^*(y) \triangleq \sup_x \,(xy - f(x)).$$

Remark 3.14 (Standard properties). 1. f^* is convex.

2. $f^{**} = f$.

3. (Young) $f(x) + f^*(y) \geq xy$.

Theorem 3.15 (Dual form of f-divergence)**.**

$$D_f(P\|Q) = \sup_{g\colon \mathbb{E}_Q[f^*(g)]<\infty} \Big\{\mathbb{E}_P[g] - \mathbb{E}_Q[f^*(g)]\Big\}.$$

Proof. Using $f(x) = \sup_y(xy - f^*(y))$,

$$\begin{aligned} D_f(P\|Q) &= \mathbb{E}_Q\Big[f\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}\Big)\Big] \\ &= \mathbb{E}_Q\Big[\sup_y\Big(\frac{\mathrm{d}P}{\mathrm{d}Q}y - f^*(y)\Big)\Big] \\ &= \sup_g\Big\{\mathbb{E}_Q\Big[\frac{\mathrm{d}P}{\mathrm{d}Q}g\Big] - \mathbb{E}_Q[f^*(g)]\Big\} \\ &= \sup_g\{\mathbb{E}_P[g] - \mathbb{E}_Q[f^*(g)]\}. \end{aligned}$$

□

Example 1: total variation

For $f(x) = \frac{1}{2}|x-1|$, one has

$$f^*(y) = \begin{cases} y, & |y| \le \frac{1}{2}, \\ +\infty, & |y| > \frac{1}{2}. \end{cases}$$

Hence

$$\mathrm{TV}(P,Q) = \sup_{\|g\|_\infty \le 1/2} \big(\mathbb{E}_P[g] - \mathbb{E}_Q[g]\big) = \frac{1}{2} \sup_{\|g\|_\infty \le 1} \big|\mathbb{E}_P[g] - \mathbb{E}_Q[g]\big|.$$

Example 2: KL and Donsker–Varadhan

For KL, one convenient normalization is $f(x) = x\log x - x + 1$ (equivalent to $x\log x$ up to an affine term), whose conjugate is $f^*(y) = e^y - 1$. Then

$$D_{\mathrm{KL}}(P\|Q) = \sup_g \Big\{\mathbb{E}_P[g] - \big(\mathbb{E}_Q[e^g] - 1\big)\Big\}.$$

Since $u - 1 \ge \log u$, this is weaker than the Donsker–Varadhan variational form. A standard way to recover Donsker–Varadhan is to optimize over constant shifts:

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &= \sup_g \sup_{a\in\mathbb{R}} \Big\{\mathbb{E}_P[g+a] - \mathbb{E}_Q[e^{g+a}] + 1\Big\} \\ &= \sup_g \Big\{\mathbb{E}_P[g] - \inf_{a\in\mathbb{R}} \big(e^a \mathbb{E}_Q[e^g] - a\big)\Big\} \\ &= \sup_g \Big\{\mathbb{E}_P[g] - \log \mathbb{E}_Q[e^g]\Big\}, \end{aligned}$$

where the infimum is attained at $a = -\log \mathbb{E}_Q[e^g]$.

Example 3: χ^2 and a variance representation

For $f(x) = (x-1)^2$, the conjugate is $f^*(y) = y + y^2/4$. Hence

$$\chi^2(P\|Q) = \sup_g \Big\{\mathbb{E}_P[g] - \mathbb{E}_Q\Big[g + \frac{g^2}{4}\Big]\Big\}.$$

By a scaling/centering trick (optimize over $\lambda(g+c)$), one can show

$$\chi^2(P\|Q) = \sup_g \frac{(\mathbb{E}_P[g] - \mathbb{E}_Q[g])^2}{\mathrm{Var}_Q(g)}.$$

Corollary 3.16 (Hammersley–Chapman–Robbins (HCR) lower bound)**.** *In a (scalar) parametric family $(P_\theta)_{\theta\in\mathbb{R}}$, if an estimator $\widehat{\theta}$ is unbiased, then*

$$\mathrm{Var}_\theta(\widehat{\theta}) \ge \sup_{\theta' \ne \theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}\|P_\theta)}.$$

In particular, taking $\theta' \to \theta$ recovers the Cramér–Rao bound

$$\mathrm{Var}_\theta(\widehat{\theta}) \ge \frac{1}{I(\theta)}.$$

Example 4: Jensen–Shannon and GANs

For $f(x) = x\log x + (x+1)\log\frac{2}{x+1}$, the conjugate is

$$f^*(y) = \begin{cases} -\log(2-e^y), & y < \log 2, \\ +\infty, & y \geq \log 2. \end{cases}$$

Therefore

$$\mathrm{JS}(P,Q) = \sup_{g \leq \log 2} \Big\{ \mathbb{E}_P[g] + \mathbb{E}_Q[\log(2-e^g)] \Big\}.$$

With the reparameterization $h = e^g/2 \in (0,1)$,

$$\mathrm{JS}(P,Q) = \sup_{0<h<1} \Big\{ \mathbb{E}_P[\log h] + \mathbb{E}_Q[\log(1-h)] \Big\} + \log 2.$$

This is closely related to the classical objective for generative adversarial networks (GANs):

$$\min_G \mathrm{JS}\big(P, P_{G(Z)}\big) = \min_G \sup_D \Big(\mathbb{E}_{X\sim P}[\log D(X)] + \mathbb{E}_{Z\sim N}[\log(1-D(G(Z)))] \Big),$$

where G is the generator, D the discriminator, and Z is a noise input.

3.8 Joint range: inequalities between two f-divergences

Definition 3.17 (Joint range)**.** Fix two f-divergences D_f and D_g. Define

$$\mathcal{R} \triangleq \{(D_f(P\|Q), D_g(P\|Q)) : \ P,Q \text{ arbitrary probability measures}\},$$

and for distributions supported on $[k] = \{1,\dots,k\}$,

$$\mathcal{R}_k \triangleq \{(D_f(P\|Q), D_g(P\|Q)) : \ P,Q \text{ probability measures on } [k]\}.$$

Theorem 3.18 (Harremoës–Vajda, 2011)**.**

$$\mathcal{R} = \mathrm{conv}(\mathcal{R}_2) = \mathcal{R}_4.$$

Remark 3.19 (Key implication). To establish an inequality relating D_f and D_g (e.g. Pinsker's inequality), it suffices to prove it for binary distributions

$$P = (p, 1-p), \qquad Q = (q, 1-q).$$

Proof. We follow the argument in the notes.

Step 1: $\mathcal{R} \subseteq \mathrm{conv}(\mathcal{R}_2)$. Fix any point $(D_f(P\|Q), D_g(P\|Q)) \in \mathcal{R}$ and assume $P \ll Q$. Let

$$L \triangleq \frac{\mathrm{d}P}{\mathrm{d}Q},$$

so L is a random variable taking values in $[0,\infty)$ with $\mathbb{E}_Q[L] = 1$. Then

$$\big(D_f(P\|Q), D_g(P\|Q)\big) = \big(\mathbb{E}_Q[f(L)], \mathbb{E}_Q[g(L)]\big).$$

Now consider the set

$$\mathcal{C} \triangleq \Big\{\mu : \ \mu \text{ is a probability measure on } [0,\infty) \text{ with } \int x\,\mu(\,\mathrm{d}x) = 1\Big\}.$$

For each $\mu \in \mathcal{C}$, associate the point

$$\Phi(\mu) \triangleq \Big(\int f(x)\,\mu(\,\mathrm{d}x),\ \int g(x)\,\mu(\,\mathrm{d}x)\Big) \in \mathbb{R}^2.$$

Clearly $\mathcal{C}$ is convex and Φ is affine, and the law of L under Q is an element of $\mathcal{C}$.

We claim that the extreme points of $\mathcal{C}$ are exactly the probability measures with mean 1 and support size at most 2. Indeed, suppose $\mu \in \mathcal{C}$ has support size at least 3. Partition $[0,\infty)$ into three measurable sets A_1, A_2, A_3 such that $\mu(A_i) > 0$ for all i. Define the conditional measures $\mu_i \triangleq \mu(\cdot \mid A_i)$ and write

$$\mu = \lambda_1\mu_1 + \lambda_2\mu_2 + \lambda_3\mu_3, \qquad \lambda_i \triangleq \mu(A_i) > 0.$$

Let $m(\nu) \triangleq \int x\,\nu(\,\mathrm{d}x)$ denote the mean. The constraints that μ is a probability measure with mean 1 are exactly

$$\lambda_1 + \lambda_2 + \lambda_3 = 1, \qquad \lambda_1 m(\mu_1) + \lambda_2 m(\mu_2) + \lambda_3 m(\mu_3) = 1.$$

These are two linear constraints on the three unknowns $(\lambda_1, \lambda_2, \lambda_3)$, hence the feasible set contains a nontrivial line segment passing through $(\lambda_1, \lambda_2, \lambda_3)$. Therefore μ can be written as a nontrivial convex combination of two distinct elements of $\mathcal{C}$, so μ is not extreme. Conversely, if μ has support size ≤ 2 and mean 1, it is straightforward to check it cannot be decomposed nontrivially.

By the Choquet–Bishop–de Leeuw theorem (every point in a metrizable compact convex set is a barycenter of its extreme points), any $\mu \in \mathcal{C}$ is a convex combination (in the barycentric sense) of extreme points, and since Φ is affine we obtain that $\Phi(\mu)$ lies in the convex hull of the set of values attained by Φ on two-point supported measures. Equivalently, every point in $\mathcal{R}$ lies in $\mathrm{conv}(\mathcal{R}_2)$.

Step 2: $\mathrm{conv}(\mathcal{R}_2) \subseteq \mathcal{R}_4$**.** The set $\mathcal{R}_2 \subset \mathbb{R}^2$ is connected, hence by the refined Carathéodory theorem in $\mathbb{R}^2$ (the $d = 2$ case), any point of $\mathrm{conv}(\mathcal{R}_2)$ can be written as a convex combination of two points in $\mathcal{R}_2$. A convex combination of two binary experiments can be realized on an alphabet of size 4, so the point lies in $\mathcal{R}_4$.

Combining the two steps yields $\mathcal{R} = \mathrm{conv}(\mathcal{R}_2) = \mathcal{R}_4$. □

Theorem 3.20 (Carathéodory)**.** *Let $S \subset \mathbb{R}^d$ and $x \in \mathrm{conv}(S)$. Then there exists $S' = \{x_1, \ldots, x_k\} \subset S$ such that $x \in \mathrm{conv}(S')$ with*

1. *$k \leq d + 1$ in general;*
2. *$k \leq d$ if S has at most d connected components.*

Examples of inequalities

1. **TV vs. Hellinger:**
$$\frac{\mathrm{H}^2}{2} \leq \mathrm{TV} \leq \sqrt{\mathrm{H}^2\Big(1 - \frac{\mathrm{H}^2}{4}\Big)}.$$
2. **TV vs. KL:**
$$\mathrm{TV}^2 \leq \frac{1}{2}D_{\mathrm{KL}}, \qquad \mathrm{TV} \leq 1 - \frac{1}{2}e^{-D_{\mathrm{KL}}}.$$
3. **KL vs. χ^2:**
$$D_{\mathrm{KL}} \leq \log(1 + \chi^2).$$

3.9 Special topic: a chain rule for H^2

Theorem 3.21 (Jayram (2009)). *For all joint distributions P_{X^n} and Q_{X^n},*

$$\mathrm{H}^2(P_{X^n}, Q_{X^n}) \le C \sum_{i=1}^{n} \mathbb{E}_P\big[\mathrm{H}^2(P_{X_i|X^{i-1}}, Q_{X_i|X^{i-1}})\big],$$

where

$$C = \prod_{i=1}^{\infty} \frac{1}{1-2^{-i}} \approx 3.46.$$

Remark 3.22 (Proof idea). The proof is surprisingly combinatorial. It suffices to prove the result for $n = 2^k$; for general $2^{k-1} < n \le 2^k$, one can pad with dummy coordinates.

Lemma 3.23 (L^2 geometry). *For arbitrary distributions $P_0, P_1, \ldots, P_m$,*

$$\frac{1}{m} \sum_{1 \le i < j \le m} \mathrm{H}^2(P_i, P_j) \le \sum_{i=1}^{m} \mathrm{H}^2(P_i, P_0).$$

Proof. This holds for any L^2 distance. Writing $\|\cdot\|$ for the L^2 norm,

$$\frac{1}{m} \sum_{1 \le i < j \le m} \|P_i - P_j\|^2 \le \sum_{i=1}^{m} \|P_i - P_0\|^2.$$

Indeed,

$$\begin{aligned} 2 \cdot \mathrm{LHS} &= \frac{1}{m} \sum_{i,j=1}^{m} \|P_i - P_j\|^2 = \frac{1}{m} \sum_{i,j=1}^{m} \|(P_i - P_0) - (P_j - P_0)\|^2 \\ &= \frac{2}{m} \sum_{i=1}^{m} \|P_i - P_0\|^2 - \frac{2}{m} \Big\| \sum_{i=1}^{m} (P_i - P_0) \Big\|^2 \le 2 \sum_{i=1}^{m} \|P_i - P_0\|^2 = 2 \cdot \mathrm{RHS}. \end{aligned}$$

Finally, $\mathrm{H}^2(P,Q) = \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$ is an L^2 distance. □

3.9.1 Interpolating distributions

For $A \subseteq [n] \triangleq \{1, \ldots, n\}$, define an interpolation P^A via the (conditional) product

$$P^A \triangleq \prod_{i=1}^{n} \big(P_{X_i|X^{i-1}}\big)^{\mathbb{1}\{i \notin A\}} \big(Q_{X_i|X^{i-1}}\big)^{\mathbb{1}\{i \in A\}}.$$

Then $P^{\varnothing} = P_{X^n}$ and $P^{[n]} = Q_{X^n}$.

Lemma 3.24 (Cut–paste property). *Let $a, b, c, d \in \{0,1\}^n$ be the indicator vectors of sets $A, B, C, D \subseteq [n]$. If $a + b = c + d$ (entrywise), then*

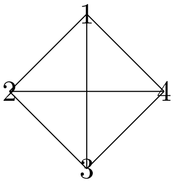
$$\mathrm{H}^2(P^A, P^B) = \mathrm{H}^2(P^C, P^D).$$

Proof. Write densities (or Radon–Nikodym derivatives) for the conditional factors. Then

$$\mathrm{H}^2(P^A, P^B) = 2 - 2\int \sqrt{p^A p^B} = 2 - 2\int \sqrt{\prod_{i=1}^{n} P_{X_i|X^{i-1}}^{2-a_i-b_i}\, Q_{X_i|X^{i-1}}^{a_i+b_i}}.$$

The right-hand side depends on $a + b$ only, hence is invariant under replacing (A, B) by (C, D) whenever $a + b = c + d$. □

Lemma 3.25 (1-factorization of cliques)**.** *For even m, the complete graph K_m can be decomposed into $(m-1)$ edge-disjoint perfect matchings ("round-robin tournaments").*



3.9.2 Completing the proof

Assume $n = 2^k$. We prove by induction on $m = 0, 1, \ldots, k$ that for any partition $A_1, \ldots, A_{2^m}$ of $[n]$ (each of size 2^{k-m}),

$$\sum_{i=1}^{2^m} \mathrm{H}^2(P^{A_i}, P^{\varnothing}) \geq c_m\, \mathrm{H}^2(P^{[n]}, P^{\varnothing}), \qquad c_m \triangleq \prod_{j=1}^{m}(1 - 2^{-j}). \tag{3.6}$$

Base case $m = 0$. The partition is just $A_1 = [n]$, so (3.6) is trivial with $c_0 = 1$.

Induction step $m - 1 \to m$. Assume (3.6) holds for $m - 1$. Let $A_1, \ldots, A_{2^m}$ be any partition of $[n]$. Apply Lemma 3.23 (Lemma 1 in the notes) with $P_0 = P^{\varnothing}$ and $P_i = P^{A_i}$ $(i \in [2^m])$ to get

$$\sum_{i=1}^{2^m} \mathrm{H}^2(P^{A_i}, P^{\varnothing}) \geq \frac{1}{2^m} \sum_{1 \leq s < t \leq 2^m} \mathrm{H}^2(P^{A_s}, P^{A_t}).$$

Using the cut–paste property (Lemma 2 in the notes), for each pair (s, t) we have

$$\mathrm{H}^2(P^{A_s}, P^{A_t}) = \mathrm{H}^2(P^{A_s \cup A_t}, P^{\varnothing}).$$

Hence

$$\sum_{i=1}^{2^m} \mathrm{H}^2(P^{A_i}, P^{\varnothing}) \geq \frac{1}{2^m} \sum_{1 \leq s < t \leq 2^m} \mathrm{H}^2(P^{A_s \cup A_t}, P^{\varnothing}). \tag{3.7}$$

Now consider the complete graph K_{2^m} on vertex set $\{1, \ldots, 2^m\}$. By Lemma 3 (1-factorization of cliques), K_{2^m} can be decomposed into $(2^m - 1)$ edge-disjoint perfect matchings $E_1, \ldots, E_{2^m-1}$. Therefore the sum over all pairs can be written as

$$\sum_{1 \leq s < t \leq 2^m} \mathrm{H}^2(P^{A_s \cup A_t}, P^{\varnothing}) = \sum_{a=1}^{2^m-1} \sum_{(s,t) \in E_a} \mathrm{H}^2(P^{A_s \cup A_t}, P^{\varnothing}).$$

Plugging this into (3.7) yields

$$\sum_{i=1}^{2^m} \mathrm{H}^2(P^{A_i}, P^{\varnothing}) \geq \frac{1}{2^m} \sum_{a=1}^{2^m-1} \sum_{(s,t)\in E_a} \mathrm{H}^2(P^{A_s\cup A_t}, P^{\varnothing}). \tag{3.8}$$

Fix a matching E_a. The sets $\{A_s \cup A_t : (s,t) \in E_a\}$ form a partition of $[n]$ into 2^{m-1} blocks (each of size $2^{k-(m-1)}$). Applying the induction hypothesis (with $m-1$) to this partition gives

$$\sum_{(s,t)\in E_a} \mathrm{H}^2(P^{A_s\cup A_t}, P^{\varnothing}) \geq c_{m-1}\, \mathrm{H}^2(P^{[n]}, P^{\varnothing}).$$

Substituting this bound into (3.8) gives

$$\sum_{i=1}^{2^m} \mathrm{H}^2(P^{A_i}, P^{\varnothing}) \geq \frac{1}{2^m} \sum_{a=1}^{2^m-1} c_{m-1}\, \mathrm{H}^2(P^{[n]}, P^{\varnothing}) = \frac{2^m-1}{2^m} c_{m-1}\, \mathrm{H}^2(P^{[n]}, P^{\varnothing}) = c_m\, \mathrm{H}^2(P^{[n]}, P^{\varnothing}),$$

where $c_m \triangleq \frac{2^m-1}{2^m} c_{m-1} = c_{m-1}(1-2^{-m})$. This is exactly (3.6) for m.

Conclusion. Taking $m = k$ in (3.6) (so the partition is into singletons) gives

$$\mathrm{H}^2(P^{[n]}, P^{\varnothing}) \leq \frac{1}{c_k} \sum_{i=1}^{n} \mathrm{H}^2(P^{\{i\}}, P^{\varnothing}) = \frac{1}{c_k} \sum_{i=1}^{n} \mathbb{E}_P\big[\mathrm{H}^2(P_{X_i|X^{i-1}}, Q_{X_i|X^{i-1}})\big].$$

Letting $k \to \infty$ yields the constant $C = \lim_{k\to\infty} 1/c_k = \prod_{j\geq 1}(1-2^{-j})^{-1}$.

Lecture 4: Large Deviations, Hypothesis Testing

4.1 Large deviations in finite alphabets: method of types

Suppose P is a pmf on $\mathcal{X}$ with $|\mathcal{X}| < \infty$. For $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P$, what is the typical "type" of $(X_1, \dots, X_n)$?

Definition 4.1 (Type)**.** For an "empirical distribution" Q on $\mathcal{X}$, let the *type class*

$$T_Q^n \ := \ \Big\{(x_1, \dots, x_n) \in \mathcal{X}^n : \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i = x\} = Q(x), \ \forall x \in \mathcal{X}\Big\}.$$

(In other words, T_Q^n is the set of all length-n sequences with empirical distribution equal to Q.)

Why types? Types encode all necessary information for $P(x^n)$.

Lemma 4.2. *If $x^n \in T_Q^n$, then*

$$P(x^n) = e^{-n(D_{\mathrm{KL}}(Q\|P) + H(Q))}.$$

Proof. Write $P(x) = P(X = x)$. Then

$$\begin{aligned} P(x^n) &= \prod_{i=1}^n P(x_i) = \prod_{x\in\mathcal{X}} \prod_{i:x_i = x} P(x) = \prod_{x\in\mathcal{X}} P(x)^{nQ(x)} \qquad (x^n \in T_Q^n) \\ &= \exp\Big(n \sum_{x\in\mathcal{X}} Q(x)\log P(x)\Big) = \exp\big(-n(D_{\mathrm{KL}}(Q\|P) + H(Q))\big). \end{aligned}$$

□

Another intriguing property: the number of sequences in a given type is exponential in n, but the number of different types is only polynomial in n.

Lemma 4.3. *The number of different type classes is*

$$\binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \le (n+1)^{|\mathcal{X}|-1}.$$

Proof. The number of types equals the number of nonnegative integer solutions to $\sum_{x\in\mathcal{X}} n_x = n$, which is $\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$. □

Lemma 4.4. *For any type Q,*

$$\frac{e^{nH(Q)}}{(n+1)^{|\mathcal{X}|-1}} \leq |T_Q^n| \leq e^{nH(Q)}.$$

(Equivalently, $|T_Q^n| \doteq e^{nH(Q)}$ ignoring polynomial factors.)

Proof. Under Q, every $x^n \in T_Q^n$ has probability $Q(x^n) = e^{-nH(Q)}$, hence

$$Q(X^n \in T_Q^n) = |T_Q^n| e^{-nH(Q)} \leq 1 \quad \Rightarrow \quad |T_Q^n| \leq e^{nH(Q)}.$$

For the lower bound, note that

$$1 = \sum_{\text{types } P} Q(X^n \in T_P^n) \leq (n+1)^{|\mathcal{X}|-1} Q(X^n \in T_Q^n) = (n+1)^{|\mathcal{X}|-1} |T_Q^n| e^{-nH(Q)},$$

where we used that the mode of a multinomial$(n; Q)$ has type Q and the number of types is at most $(n+1)^{|\mathcal{X}|-1}$. □

Corollary 4.5. *For any type Q,*

$$\frac{e^{-nD_{\mathrm{KL}}(Q\|P)}}{(n+1)^{|\mathcal{X}|-1}} \leq P(X^n \in T_Q^n) \leq e^{-nD_{\mathrm{KL}}(Q\|P)}.$$

Proof. Combine the previous lemma with $P(x^n) = e^{-n(D_{\mathrm{KL}}(Q\|P)+H(Q))}$ for $x^n \in T_Q^n$. □

The above corollary, together with the bound on the number of types, yields Sanov's theorem.

Theorem 4.6 (Sanov's theorem)**.** *Let $|\mathcal{X}| < \infty$ and let $\hat{P}$ be the empirical distribution (type) of $X_1, \ldots, X_n \sim P$ where P is strictly positive on $\mathcal{X}$. Let $\mathcal{E}$ be a closed set of distributions with non-empty interior. Then*

$$\mathbb{P}(\hat{P} \in \mathcal{E}) = \exp\Big(-n \min_{Q \in \mathcal{E}} D_{\mathrm{KL}}(Q\|P) + o(n)\Big).$$

Remark 4.7. The map

$$P \longmapsto \arg\min_{Q \in \mathcal{E}} D_{\mathrm{KL}}(Q\|P)$$

is called the *information projection*.

Proof sketch. Upper bound.

$$\mathbb{P}(\hat{P} \in \mathcal{E}) = \sum_{Q \in \mathcal{E}} P(X^n \in T_Q^n) \leq \sum_{Q \in \mathcal{E}} e^{-nD_{\mathrm{KL}}(Q\|P)} \leq (n+1)^{|\mathcal{X}|-1} e^{-n \min_{Q \in \mathcal{E}} D_{\mathrm{KL}}(Q\|P)}.$$

Lower bound. For any $Q \in \mathcal{E}$, $P(X^n \in T_Q^n) \geq (n+1)^{-(|\mathcal{X}|-1)} e^{-nD_{\mathrm{KL}}(Q\|P)}$. Choose $Q \to Q^*$ and use continuity of $Q \mapsto D_{\mathrm{KL}}(Q\|P)$. □

4.2 Information projection, exponential tilting, and CGF

A corollary of Sanov's theorem is:

Corollary 4.8.

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n X_i \ge v\big)} = \min_{Q:\mathbb{E}_Q[X]\ge v} D_{\mathrm{KL}}(Q\|P).$$

If $\mathbb{E}_P[X] \ge v$, then one can choose $Q = P$ and the right-hand side is 0. Can we find the minimizer Q^* if $\mathbb{E}_P[X] < v$?

Definition 4.9 (Exponential tilt). For $\lambda \in \mathbb{R}$, the exponential tilt of P along X is

$$P_\lambda(dx) = \exp\big(\lambda x - \psi(\lambda)\big)\, P(dx),$$

where

$$\psi(\lambda) := \log \mathbb{E}_P e^{\lambda X}$$

is the cumulant generating function (CGF) of X.

Remark 4.10. The family $\{P_\lambda\}$ is called an *exponential family* in statistics, where $\psi(\lambda)$ is the "log partition function." In particular, $\mathbb{E}_{P_\lambda}[X] = \psi'(\lambda)$, and $\lambda \mapsto \psi(\lambda)$ is convex.

Theorem 4.11 ("Maximum entropy distribution"). *Assume $\mathbb{E}_P[X] < v$, and there exists $\lambda \in \mathbb{R}$ such that $\mathbb{E}_{P_\lambda}[X] = v$. Then*

$$\min_{Q:\mathbb{E}_Q[X]\ge v} D_{\mathrm{KL}}(Q\|P) = D_{\mathrm{KL}}(P_\lambda\|P) = \lambda v - \psi(\lambda) = \psi^*(v),$$

where ψ^* is the convex conjugate of ψ.

Proof sketch. Since $\mathbb{E}_P[X] = \psi'(0) < v = \psi'(\lambda)$, by convexity of ψ we have $\lambda > 0$. If $\mathbb{E}_Q[X] \ge v$, then

$$\begin{aligned} D_{\mathrm{KL}}(Q\|P) &= \mathbb{E}_Q\Big[\log \frac{dQ}{dP}\Big] = \mathbb{E}_Q\Big[\log \frac{dQ}{dP_\lambda} + \log \frac{dP_\lambda}{dP}\Big] \\ &= D_{\mathrm{KL}}(Q\|P_\lambda) + \mathbb{E}_Q[\lambda X - \psi(\lambda)] \ge \lambda v - \psi(\lambda). \end{aligned}$$

Also $D_{\mathrm{KL}}(P_\lambda\|P) = \mathbb{E}_{P_\lambda}[\lambda X - \psi(\lambda)] = \lambda v - \psi(\lambda)$. Finally, since $v = \psi'(\lambda)$,

$$\psi^*(v) = \sup_{t\in\mathbb{R}}\{tv - \psi(t)\} = \lambda v - \psi(\lambda),$$

by convexity of ψ. □

4.3 Large deviations in general alphabets: Cramér's theorem

Theorem 4.12 (Cramér's theorem). *For i.i.d. $X_1, \ldots, X_n \sim P$ with $\mathbb{E}_P[X] < v < \|X\|_\infty$,*

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n X_i > v\big)} = \psi^*(v) = \inf_{Q:\mathbb{E}_Q[X]>v} D_{\mathrm{KL}}(Q\|P),$$

where ψ^* is the convex conjugate of the CGF $\psi(\lambda) = \log \mathbb{E}_P e^{\lambda X}$.

Remark 4.13. This generalizes the previous results to arbitrary alphabets. Two different proofs illustrate the connections between (i) probabilistic large deviations (yielding $\psi^*(v)$) and (ii) information-theoretic arguments (yielding $\inf_{Q:\mathbb{E}_Q[X]>v} D_{\mathrm{KL}}(Q\|P)$).

4.3.1 Probabilistic proof (sketch)

(Lower bound on the exponent) Chernoff inequality.

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n} X_i > v\Big) \le \inf_{\lambda\ge 0} e^{-\lambda n v}\mathbb{E}_P\big[e^{\lambda\sum_{i=1}^n X_i}\big] = \inf_{\lambda\ge 0}\exp\big(-n(\lambda v - \psi(\lambda))\big)$$
$$= \exp\big(-n\psi^*(v)\big).$$

(Upper bound on the exponent) Exponential tilting. Since $\mathbb{E}_P[X] < v < \|X\|_\infty$, there exists $\lambda = \lambda(\varepsilon) > 0$ such that $\mathbb{E}_{P_\lambda}[X] = v + \varepsilon$, where P_λ is the exponential tilt of P. By the law of large numbers,

$$P_\lambda\Big(\tfrac{1}{n}\sum_{i=1}^{n} X_i \in (v, v+2\varepsilon)\Big) = 1 - o(1) \quad (n\to\infty).$$

At the same time, for sequences with $\frac{1}{n}\sum_i X_i \in (v, v+2\varepsilon)$,

$$\frac{dP_\lambda}{dP}(x_1,\dots,x_n) = \exp\Big(\lambda\sum_{i=1}^{n} x_i - n\psi(\lambda)\Big) \le \exp\big(n(\lambda(v+2\varepsilon) - \psi(\lambda))\big),$$

so

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n} X_i \in (v, v+2\varepsilon)\Big) \ge (1-o(1))\exp\big(-n(\lambda(v+2\varepsilon) - \psi(\lambda))\big).$$

Letting $\varepsilon \downarrow 0$ completes the proof sketch.

4.3.2 Information-theoretic proof (sketch)

Let $E_n := \{\frac{1}{n}\sum_{i=1}^n X_i > v\}$.

Upper bound. Fix any Q with $\mathbb{E}_Q[X] > v$. Then $Q(E_n) = 1 - o(1)$ by the law of large numbers. Using the binary relative entropy bound,

$$Q(E_n)\log\frac{Q(E_n)}{\mathbb{P}(E_n)} \le D_{\mathrm{KL}}(Q^{\otimes n}\|P^{\otimes n}) = nD_{\mathrm{KL}}(Q\|P),$$

which implies

$$\frac{1}{n}\log\frac{1}{\mathbb{P}(E_n)} \le (1+o(1))D_{\mathrm{KL}}(Q\|P).$$

Taking the infimum over such Q yields $\limsup \frac{1}{n}\log\frac{1}{\mathbb{P}(E_n)} \le \inf_{Q:\mathbb{E}_Q[X]>v} D_{\mathrm{KL}}(Q\|P)$.

Lower bound. Let $\tilde{P}_{X^n} := P_{X^n|E_n}$ (the conditional law given E_n). Then $\tilde{P}_{X^n}$ has mean $> v$ and

$$\frac{1}{n}\log\frac{1}{\mathbb{P}(E_n)} = \frac{1}{n}D_{\mathrm{KL}}(\tilde{P}_{X^n}\|P^{\otimes n}).$$

Moreover,

$$D_{\mathrm{KL}}(\tilde{P}_{X^n}\|P^{\otimes n}) = \sum_{i=1}^{n}\mathbb{E}_{\tilde{P}}\big[D_{\mathrm{KL}}(\tilde{P}_{X_i|X^{i-1}}\|P)\big] \ge \sum_{i=1}^{n} D_{\mathrm{KL}}\big(\mathbb{E}_{\tilde{P}}[\tilde{P}_{X_i|X^{i-1}}]\big\|P\big)$$
$$\ge nD_{\mathrm{KL}}\Big(\frac{1}{n}\sum_{i=1}^{n}\tilde{P}_{X_i}\,\Big\|\,P\Big),$$

by convexity of KL. Writing $\bar{P} := \frac{1}{n}\sum_{i=1}^n \tilde{P}_{X_i}$, we have $\mathbb{E}_{\bar{P}}[X] = \mathbb{E}_{\tilde{P}}[\frac{1}{n}\sum_i X_i] > v$. Therefore

$$\frac{1}{n} D_{\mathrm{KL}}(\tilde{P}_{X^n} \| P^{\otimes n}) \geq \inf_{Q:\mathbb{E}_Q[X]>v} D_{\mathrm{KL}}(Q\|P),$$

as desired.

4.4 Simple hypothesis testing and Neyman–Pearson

4.4.1 Setup

Simple hypothesis testing:

$$H_0 : X \sim P, \qquad H_1 : X \sim Q.$$

For a test $T = T(X) \in \{0,1\}$ (possibly randomized), define

$$\alpha := P(T=0) \quad (1 - \text{Type I error}), \qquad \beta := Q(T=0) \quad (\text{Type II error}).$$

Definition 4.14. Let $R(P,Q)$ denote the set of all achievable points $(\alpha,\beta) \in [0,1]^2$ when T ranges over all possible tests.

Basic properties.

1. $R(P,Q)$ is convex (randomized combination of two tests).
2. $(\alpha,\alpha) \in R(P,Q)$ for all $\alpha \in [0,1]$ (take $T \sim \mathrm{Bern}(1-\alpha)$ independent of X).
3. $(\alpha,\beta) \in R(P,Q) \iff (1-\alpha, 1-\beta) \in R(P,Q)$ (replace T by $1-T$).

4.4.2 Neyman–Pearson lemma

Likelihood ratio tests (LRT) attain the lower boundary of $R(P,Q)$. Fix a threshold $\tau \in \mathbb{R}$ and define

$$T^*(x) = \begin{cases} 0, & \log \frac{P(x)}{Q(x)} > \tau, \\ \in \{0,1\}, & \log \frac{P(x)}{Q(x)} = \tau \quad \text{(randomized)}, \\ 1, & \log \frac{P(x)}{Q(x)} < \tau. \end{cases}$$

Then for any other test T,

$$\alpha(T) \geq \alpha(T^*) \quad \Longrightarrow \quad \beta(T) \geq \beta(T^*).$$

Proof sketch. $\alpha(T) \geq \alpha(T^*)$ implies $\mathbb{E}_P[T - T^*] \leq 0$. Moreover, distinguishing the cases $\frac{dQ}{dP} \gtrless e^{-\tau}$ yields

$$\mathbb{E}_P\big[(\tfrac{dQ}{dP} - e^{-\tau})(T - T^*)\big] \leq 0 \quad \Rightarrow \quad \mathbb{E}_P\big[\tfrac{dQ}{dP}(T - T^*)\big] \leq 0.$$

But $\mathbb{E}_P[\frac{dQ}{dP}(T-T^*)] = \mathbb{E}_Q[T - T^*]$, so $\mathbb{E}_Q[T - T^*] \leq 0$, i.e. $\beta(T) \geq \beta(T^*)$. □

4.5 Asymptotics: Chernoff regime

Consider

$$H_0 : X^n \sim P^{\otimes n}, \qquad H_1 : X^n \sim Q^{\otimes n}, \qquad n \to \infty.$$

What are all possible values of (E_0, E_1) such that there exists tests T_n with

$$1 - \alpha(T_n) \le e^{-nE_0} \wedge 0.99 \quad \text{and} \quad \beta(T_n) \le e^{-nE_1} \wedge 0.99 \quad \text{asymptotically?}$$

In other words, what are the best tradeoffs between (E_0, E_1), the error exponents on Type I and Type II errors?

Theorem 4.15 ((E_0, E_1) tradeoff)**.** *Assume $P \ll Q$ and $Q \ll P$. The upper boundary of all achievable (E_0, E_1) pairs is given by*

$$E_0 = D_{\mathrm{KL}}(P_\lambda \| P), \qquad E_1 = D_{\mathrm{KL}}(P_\lambda \| Q), \qquad \lambda \in [0, 1],$$

where P_λ is the (normalized) geometric mixture

$$P_\lambda \propto P^{1-\lambda} Q^\lambda.$$

Corollary 4.16 (Chernoff information)**.**

$$\max_{(E_0,E_1)\ \textit{achievable}} \min\{E_0, E_1\} = - \inf_{\lambda \in (0,1)} \log \int (dP)^{1-\lambda} (dQ)^\lambda.$$

This quantity is denoted $C(P, Q)$ and is called the Chernoff information*.*

Remark 4.17 (Relation to Hellinger distance). Let $H^2(P, Q) := \int (\sqrt{dP} - \sqrt{dQ})^2$ (squared Hellinger distance), so that $\int \sqrt{dP\, dQ} = 1 - \frac{1}{2} H^2(P, Q)$. Then choosing $\lambda = \frac{1}{2}$ gives

$$-\log\Big(1 - \tfrac{1}{2} H^2(P, Q)\Big) \le C(P, Q) \le -2 \log\Big(1 - \tfrac{1}{2} H^2(P, Q)\Big).$$

One inequality uses

$$\int p^{1-\lambda} q^\lambda = \mathbb{E}_P\Big[\Big(\frac{q}{p}\Big)^\lambda\Big] \ge \Big(\mathbb{E}_P \sqrt{\frac{q}{p}}\Big)^{2\lambda} = \Big(\int \sqrt{pq}\Big)^{2\lambda} \quad \text{for } \lambda \ge \tfrac{1}{2},$$

and symmetrically for $\lambda \le \frac{1}{2}$.

4.5.1 Proof of the corollary and achievability (notes)

For

$$P_\lambda = \frac{P^{1-\lambda} Q^\lambda}{Z(\lambda)}, \qquad Z(\lambda) := \int (dP)^{1-\lambda} (dQ)^\lambda,$$

we have

$$D_{\mathrm{KL}}(P_\lambda \| P) = \mathbb{E}_{P_\lambda}\Big[\log \frac{dP_\lambda}{dP}\Big] = \mathbb{E}_{P_\lambda}\Big[\lambda \log \frac{dQ}{dP} - \log Z(\lambda)\Big],$$
$$D_{\mathrm{KL}}(P_\lambda \| Q) = \mathbb{E}_{P_\lambda}\Big[\log \frac{dP_\lambda}{dQ}\Big] = \mathbb{E}_{P_\lambda}\Big[(1-\lambda) \log \frac{dP}{dQ} - \log Z(\lambda)\Big].$$

Hence

$$D_{\mathrm{KL}}(P_\lambda \| P) - D_{\mathrm{KL}}(P_\lambda \| Q) = \mathbb{E}_{P_\lambda}\Big[\log \frac{dQ}{dP}\Big].$$

Let λ^* minimize the convex function $\lambda \mapsto \log Z(\lambda)$ on $[0,1]$. Then

$$0 = \frac{d}{d\lambda} \log Z(\lambda)\Big|_{\lambda=\lambda^*} = \mathbb{E}_{P_{\lambda^*}}\Big[\log \frac{dQ}{dP}\Big],$$

so $D_{\mathrm{KL}}(P_{\lambda^*} \| P) = D_{\mathrm{KL}}(P_{\lambda^*} \| Q)$ and

$$D_{\mathrm{KL}}(P_{\lambda^*} \| P) = -\log Z(\lambda^*) = -\inf_{\lambda \in (0,1)} \log Z(\lambda).$$

Achievability. A sufficient statistic is

$$L := \frac{1}{n} \sum_{i=1}^{n} L_i, \qquad L_i := \log \frac{P(X_i)}{Q(X_i)}.$$

A natural test is $T_n = \mathbb{1}\{L \le \gamma\}$ for some threshold $\gamma \in \mathbb{R}$. By large deviations,

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{P(L \le \gamma)} &= \psi_P^*(\gamma) = D_{\mathrm{KL}}(P^* \| P), \\
\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{Q(L > \gamma)} &= \psi_Q^*(\gamma) = D_{\mathrm{KL}}(Q^* \| Q),
\end{aligned}$$

where

$$\psi_P(\lambda) = \log \mathbb{E}_P e^{\lambda L_1} = \log \int p^{1+\lambda} q^{-\lambda} \qquad \text{(similarly for } \psi_Q\text{)},$$

and

$$\begin{aligned}
P^*(dx) &= \exp\big(\lambda_P^* \log \tfrac{p(x)}{q(x)} - \psi_P(\lambda_P^*)\big)\, P(dx), \qquad \mathbb{E}_{P^*}[L_1] = \gamma, \\
Q^*(dx) &= \exp\big(\lambda_Q^* \log \tfrac{p(x)}{q(x)} - \psi_Q(\lambda_Q^*)\big)\, Q(dx), \qquad \mathbb{E}_{Q^*}[L_1] = \gamma.
\end{aligned}$$

Since P^*, Q^* lie in the family $(P_\lambda)_{\lambda\in[0,1]}$, one concludes $P^* = Q^* = P_{\lambda^*}$ for an appropriate choice of γ. Thus one asymptotically achieves all pairs

$$(E_0, E_1) = (D_{\mathrm{KL}}(P_\lambda \| P), D_{\mathrm{KL}}(P_\lambda \| Q)), \qquad \lambda \in [0,1].$$

4.6 Converse: weak vs strong

Suppose some test T_n asymptotically attains

$$\alpha(T_n) \ge 1 - e^{-nE_0}, \qquad \beta(T_n) \le e^{-nE_1}.$$

Weak converse (by DPI).

$$D_{\mathrm{KL}}(\mathrm{Bern}(\alpha) \| \mathrm{Bern}(\beta)) \le n D_{\mathrm{KL}}(P \| Q), \qquad D_{\mathrm{KL}}(\mathrm{Bern}(\beta) \| \mathrm{Bern}(\alpha)) \le n D_{\mathrm{KL}}(Q \| P).$$

(These are insufficient to establish the tight (E_0, E_1) tradeoff.)

Strong converse (on the whole likelihood ratio). For all $\gamma > 0$,

$$\alpha - \gamma\beta \le P\Big(\sum_{i=1}^n \log\frac{p}{q}(X_i) > \log\gamma\Big),$$
$$\beta - \frac{\alpha}{\gamma} \le Q\Big(\sum_{i=1}^n \log\frac{p}{q}(X_i) < \log\gamma\Big).$$

Proof of the first inequality. Let $L := \sum_{i=1}^n \log\frac{p}{q}(X_i) = \log\frac{p^{\otimes n}}{q^{\otimes n}}(X^n)$. Then

$$\begin{aligned}\alpha - \gamma\beta &= P^{\otimes n}(T_n = 0) - \gamma Q^{\otimes n}(T_n = 0) = \mathbb{E}_{Q^{\otimes n}}\big[(e^L - \gamma)\,\mathbb{1}\{T_n = 0\}\big] \\ &\le \mathbb{E}_{Q^{\otimes n}}\big[(e^L - \gamma)\,\mathbb{1}\{T_n = 0,\ L > \log\gamma\}\big] \le \mathbb{E}_{Q^{\otimes n}}\big[e^L\,\mathbb{1}\{L > \log\gamma\}\big] = P^{\otimes n}(L > \log\gamma).\end{aligned}$$

The second inequality is similar. □

Returning to the converse, choose $\gamma = e^{n\theta}$. Then

$$1 - e^{-nE_0} - e^{-n(E_1-\theta)} \le \alpha - \gamma\beta \le P\Big(\frac{1}{n}\sum_{i=1}^n \log\frac{p}{q}(X_i) > \theta\Big),$$

so

$$\min\{E_0,\, E_1 - \theta\} \le \psi_P^*(\theta), \qquad \forall\theta.$$

If $E_0 \ge D_{\mathrm{KL}}(P_\lambda\|P) + \varepsilon$ and $E_1 \ge D_{\mathrm{KL}}(P_\lambda\|Q) + \varepsilon$, choose

$$\theta = D_{\mathrm{KL}}(P_\lambda\|Q) - D_{\mathrm{KL}}(P_\lambda\|P) = \mathbb{E}_{P_\lambda}\Big[\log\frac{p}{q}\Big],$$

then $\psi_P^*(\theta) = D_{\mathrm{KL}}(P_\lambda\|P)$ and we get a contradiction.

4.7 Special topic: Stein's regime

Stein's regime:

$$H_0 : X^n \sim P^{\otimes n}, \qquad H_1 : X^n \sim Q^{\otimes n}.$$

There exists a test T_n such that $\alpha(T_n) = 1 - \varepsilon$ and $\beta(T_n) = e^{-nE}$. What is the largest possible E_n^*?

From the Chernoff regime with $E_0 = 0$, we already know

$$E_n^* = D_{\mathrm{KL}}(P\|Q) + o(1) \qquad \text{(Stein's lemma)}.$$

4.7.1 Next-order term

Theorem 4.18.

$$E_n^* = D_{\mathrm{KL}}(P\|Q) - \sqrt{\frac{V(P\|Q)}{n}}\,\mathrm{erfc}^{-1}(\varepsilon) + o\Big(\frac{1}{\sqrt{n}}\Big),$$

where

$$\mathrm{erfc}(z) := \mathbb{P}(N(0,1) > z) = \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx, \qquad V(P\|Q) := \mathrm{Var}_P\Big(\log\frac{p}{q}\Big)\ (< \infty).$$

Proof sketch. Achievability. Consider

$$T_n = \mathbb{1}\Big\{\frac{1}{n}\sum_{i=1}^n \log\frac{p}{q}(X_i) \le \gamma\Big\}.$$

By CLT under P,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(\log\frac{p}{q}(X_i) - D_{\mathrm{KL}}(P\|Q)\Big) \xrightarrow{d} N\big(0, V(P\|Q)\big).$$

Thus choosing

$$\gamma = D_{\mathrm{KL}}(P\|Q) - \sqrt{\frac{V(P\|Q)}{n}}\,\mathrm{erfc}^{-1}(\varepsilon)$$

ensures $\alpha(T_n) \to 1-\varepsilon$. For $\beta(T_n)$, by Markov,

$$Q\Big(\sum_{i=1}^n \log\frac{p}{q}(X_i) > n\gamma\Big) \le e^{-n\gamma}\,\mathbb{E}_Q\Big[e^{\sum_{i=1}^n \log\frac{p}{q}(X_i)}\Big] = e^{-n\gamma}.$$

Converse. If $E_n \ge D_{\mathrm{KL}}(P\|Q) + \frac{c}{\sqrt{n}}$, then the strong converse gives a CLT-based bound implying $c \le -\sqrt{V(P\|Q)}\,\mathrm{erfc}^{-1}(\varepsilon)$ (up to a vanishing slack). □

Remark 4.19. Using Berry–Esseen bounds (under moment conditions), the $o(1/\sqrt{n})$ term can be improved to $O((\log n)/n)$.

4.7.2 Strong converse for channel coding (sketch)

Recall the standard channel coding setup:

- Message $m \sim \mathrm{Unif}(\{1,\dots,M\})$.
- Encoder maps $m \mapsto X^n \in \mathcal{X}^n$.
- Channel: $P_{Y|X}$; output $Y^n \in \mathcal{Y}^n$.
- Decoder outputs $\hat{m} \in \{1,\dots,M\}$.
- Error probability: $\mathbb{P}(\hat{m} \ne m) \le \varepsilon$.

Communications aim to maximize/minimize the rate

$$R := \frac{\log M}{n}.$$

In Lec 1, one uses Fano's inequality (DPI for KL) to prove the weak converse $R \le (1+o(1))C$ if $\varepsilon = o(1)$, where

$$C = \max_{P_X} I(X;Y) = \max_{P_X} I(P_X; P_{Y|X}).$$

Theorem 4.20 (Strong converse)**.** *For any fixed $\varepsilon < 1$,*

$$R \le (1+o(1))\,C.$$

Remark 4.21. This means the communication problem has a "sharp" threshold on the error probability. When $R > 1.001\,C$, then asymptotically one cannot achieve success probability 10^{-8}; when $R < 0.999\,C$, then asymptotically one can suddenly achieve success probability $1 - 10^{-8}$.

Idea. The communication problem is not binary hypothesis testing (it is a recovery problem), but one can reduce recovery to detection: if one can distinguish between different inputs, then one can also distinguish from the case where input and output are independent.

4.7.3 Reduction to hypothesis testing

Consider two scenarios (joint laws on $m, X^n, Y^n, \hat{m}$):

$$\begin{aligned} H_0: &\quad P_{m,X^n,Y^n,\hat{m}} = \frac{1}{M} P_{X^n|m} \, P_{Y^n|X^n} \, P_{\hat{m}|Y^n}, \\ H_1: &\quad Q_{m,X^n,Y^n,\hat{m}} = \frac{1}{M} P_{X^n|m} \, Q_{Y^n} \, P_{\hat{m}|Y^n}, \qquad ((m, X^n) \perp (Y^n, \hat{m})). \end{aligned}$$

Then $P(m = \hat{m}) \geq 1 - \varepsilon$ and $Q(m = \hat{m}) = \frac{1}{M}$. Moreover,

$$\frac{P_{m,X^n,Y^n,\hat{m}}}{Q_{m,X^n,Y^n,\hat{m}}} = \frac{P_{Y^n|X^n}}{Q_{Y^n}} = \prod_{i=1}^{n} \frac{P_{Y_i|X_i}}{Q_{Y_i}}.$$

Therefore, by the strong converse inequality,

$$1 - \varepsilon - \frac{\gamma}{M} \leq P\Big(\sum_{i=1}^{n} \log \frac{P_{Y_i|X_i}}{Q_{Y_i}} > \log \gamma \Big).$$

Technical difficulty. P_{X^n} is often not a product distribution.

Solution via types (finite $|\mathcal{X}|$). When $|\mathcal{X}| < \infty$, we can WLOG assume all codewords X^n have the same type P_0. Since there are at most $(n+1)^{|\mathcal{X}|-1}$ types, one can find a type class that changes the error probability to $\varepsilon + o(1)$ while the rate changes by at most $O(\frac{\log n}{n})$.

When X^n has type P_0 a.s., choose

$$Q_Y = \sum_{x \in \mathcal{X}} P_0(x) \, P_{Y|X=x}.$$

Then

$$\mathbb{E}\Big[\sum_{i=1}^{n} \log \frac{P_{Y_i|X_i}}{Q_{Y_i}} \Big] = n \, I(P_0; P_{Y|X}) \leq nC,$$

and

$$\operatorname{Var}\Big(\sum_{i=1}^{n} \log \frac{P_{Y_i|X_i}}{Q_{Y_i}} \Big) = n \, \mathbb{E}_{P_0} \Big[\operatorname{Var}\Big(\log \frac{P_{Y|X}}{Q_Y} \,\Big|\, X \Big) \Big] \leq n \operatorname{Var}\Big(\log \frac{P_{Y|X}}{Q_Y} \Big) = O(n).$$

Now choosing $\gamma = \frac{1-\varepsilon}{2} M$ and applying Chebyshev's inequality yields

$$\log \gamma \leq nC + O(\sqrt{n}) \quad \Rightarrow \quad R = \frac{\log M}{n} \leq C + O\Big(\frac{1}{\sqrt{n}} \Big).$$

4.7.4 Converse for finite blocklength (sketch)

Is there a next-order upper bound on R?

Theorem 4.22. *Suppose the capacity-achieving input distribution P_X^* is unique and $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Under regularity conditions,*

$$R \le C - \sqrt{\frac{V}{n}}\,\mathrm{erfc}^{-1}(\varepsilon) + o\Big(\frac{1}{\sqrt{n}}\Big), \qquad V := \mathrm{Var}\Big(\log \frac{P_{Y|X}}{P_Y^*}\Big),$$

where P_Y^* is the output distribution induced by P_X^*.

Proof sketch. Using the previous analysis and the uniqueness of P_X^*, one only needs to deal with input type $P_0 \approx P_X^*$. Then the result follows from Stein's regime as long as we can show

$$\mathbb{E}_{P_X^*}\Big[\mathrm{Var}\Big(\log \frac{P_{Y|X}}{P_Y^*} \,\Big|\, X\Big)\Big] = \mathrm{Var}\Big(\log \frac{P_{Y|X}}{P_Y^*}\Big) = V.$$

This follows from the lemma below. □

Lemma 4.23. *Any capacity-achieving input P_X^* satisfies*

$$D_{\mathrm{KL}}(P_{Y|X=x} \| P_Y^*) \le C, \qquad \forall x \in \mathcal{X},$$

and

$$D_{\mathrm{KL}}(P_{Y|X=x} \| P_Y^*) = C, \qquad \forall x \in \mathrm{supp}(P_X^*).$$

Proof. Consider the directional derivative of $I(P_X; P_{Y|X})$ at P_X^* in the direction $P_X - P_X^*$:

$$0 \ge \lim_{\epsilon \to 0^+} \frac{I(P_X^* + \epsilon(P_X - P_X^*); P_{Y|X}) - I(P_X^*; P_{Y|X})}{\epsilon} = (\mathbb{E}_{P_X} - \mathbb{E}_{P_X^*})\big[D_{\mathrm{KL}}(P_{Y|X} \| P_Y^*)\big].$$

Choosing $P_X = \delta_x$ yields $D_{\mathrm{KL}}(P_{Y|X=x} \| P_Y^*) \le C$ for all x. The second claim follows since $C = \mathbb{E}_{P_X^*}[D_{\mathrm{KL}}(P_{Y|X} \| P_Y^*)] \le C$, hence equality must hold for all $x \in \mathrm{supp}(P_X^*)$. □

Lecture 5: Functional (In)equalities

5.1 From Shannon-type inequalities to functional inequalities

Recall (Shannon-type inequalities). All entropy inequalities that can be derived using:

(1) *Monotonicity:* $H(X) \le H(X,Y)$.

(2) *Submodularity:* $I(X;Y \mid Z) \ge 0$.

This lecture covers some *non*-Shannon-type inequalities.

Definition 5.1 (Differential entropy)**.** For a random vector X with density f on $\mathbb{R}^d$, its *differential entropy* is

$$h(X) \coloneqq h(f) \coloneqq \int_{\mathbb{R}^d} -f(x)\log f(x)\,\mathrm{d}x.$$

Notes.

(1) $h(X) \in \mathbb{R} \cup \{\pm\infty\}$. In particular, it can be negative.

(2) For a scalar $a \neq 0$, $h(aX) = h(X) + \log|a|$ in dimension $d = 1$. More generally, if $X \in \mathbb{R}^d$, then $h(aX) = h(X) + d\log|a|$.

(3) The inequality $h(X) \le h(X,Y)$ no longer holds. However, it is still true that

$$I(X;Y) = h(X) + h(Y) - h(X,Y) \ge 0.$$

Example 5.2 (Gaussian differential entropy)**.** If $X \sim \mathcal{N}(\mu,\Sigma)$ on $\mathbb{R}^d$, then

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det\Sigma}} \exp\Big(-\tfrac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\Big),$$

so

$$\begin{aligned} h(X) &= \mathbb{E}_{X\sim f}\Big[\tfrac{1}{2}\log\big((2\pi)^d\det\Sigma\big) + \tfrac{1}{2}(X-\mu)^\top\Sigma^{-1}(X-\mu)\Big] \\ &= \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\Sigma. \end{aligned}$$

Easy fact (maximum entropy principle). If $\mathrm{Cov}(X) = \Sigma$, then $h(X) \le h(\mathcal{N}(0,\Sigma))$.

Proof. $0 \le D_{\mathrm{KL}}\big(P_X \,\|\, \mathcal{N}(\mathbb{E}X,\Sigma)\big) = -h(X) + h(\mathcal{N}(0,\Sigma))$ (check!). □

Theorem 5.3 (Entropy power inequality (EPI))**.** *For independent random vectors X, Y on $\mathbb{R}^d$,*

$$\exp\Big(\frac{2}{d}h(X+Y)\Big) \ge \exp\Big(\frac{2}{d}h(X)\Big) + \exp\Big(\frac{2}{d}h(Y)\Big).$$

Notes.

(1) Equality holds iff X, Y are Gaussian and $\Sigma_X = c\,\Sigma_Y$ for some $c > 0$.

(2) EPI shows that for given values of $h(X)$ and $h(Y)$, $h(X+Y)$ is minimized when X, Y are Gaussian.

5.2 Proof of EPI via Fisher information (Stam 1959)

We present the proof in Stam (1959).

5.2.1 A detour: Fisher information

Definition 5.4 (Fisher information of a location family)**.** For a real-valued random variable X with density f, the Fisher information is

$$J(X) \coloneqq \int_{\mathbb{R}} \frac{(f'(x))^2}{f(x)}\,\mathrm{d}x.$$

Recall (parametric Fisher information). In Lec 3, for $Y \sim P_\theta$ with density p_θ,

$$I(\theta) \coloneqq I^Y(\theta) \coloneqq \int \frac{\left(\partial_\theta p_\theta\right)^2}{p_\theta}\,\mathrm{d}x.$$

These are connected via $I^Y(\theta) = J(X)$ when $Y = \theta + X$.

Properties.

(1) $J(aX) = \frac{1}{a^2} J(X)$.

(2) **DPI.** $I^Y(\theta) \le I^X(\theta)$ if $\theta - X - Y$ is a Markov chain.

Proof sketch:

$$I^Y(\theta) = \lim_{\delta \to 0} \frac{1}{\delta^2} \chi^2\big(P_{Y|\theta+\delta} \,\|\, P_{Y|\theta}\big) \le \lim_{\delta \to 0} \frac{1}{\delta^2} \chi^2\big(P_{X|\theta+\delta} \,\|\, P_{X|\theta}\big) = I^X(\theta).$$

Theorem 5.5 (Stam)**.** *For independent X_1, X_2,*

$$\frac{1}{J(X_1+X_2)} \ge \frac{1}{J(X_1)} + \frac{1}{J(X_2)}.$$

Equivalently, for all $a, b > 0$,

$$(a+b)^2 J(X_1+X_2) \le a^2 J(X_1) + b^2 J(X_2).$$

Proof. Write $Y_1 = a\theta + X_1$, $Y_2 = b\theta + X_2$. Then

$$I^{Y_1}(\theta) = I^{Y_1/a}(\theta) = J(X_1/a) = a^2 J(X_1).$$

Therefore,

$$(a+b)^2 J(X_1+X_2) = I^{Y_1+Y_2}(\theta) \le I^{Y_1,Y_2}(\theta) = a^2 J(X_1) + b^2 J(X_2),$$

where the inequality is the data processing inequality. □

5.2.2 de Bruijn's identity

Theorem 5.6 (de Bruijn)**.** *Let $Z \sim \mathcal{N}(0,1)$ be independent of X. Then for $a > 0$,*

$$\frac{\mathrm{d}}{\mathrm{d}a} h(X + \sqrt{a}\, Z) = \frac{1}{2} J(X + \sqrt{a}\, Z).$$

Proof. Let $p_a = p * \mathcal{N}(0,a)$ be the density of $X + \sqrt{a}\, Z$. Then

$$\frac{\partial p_a}{\partial a} = \frac{1}{2} p_a'' \qquad (*) \tag{5.1}$$

($''$ denotes the second derivative).

To see (5.1), note that for any test function φ,

$$\begin{aligned} \frac{\partial}{\partial a}\mathbb{E}_{p_a}[\varphi] &= \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\big[\varphi(X + \sqrt{a + \Delta}\, Z) - \varphi(X + \sqrt{a}\, Z)\big] \\ &= \lim_{\Delta \to 0} \frac{1}{\Delta} \mathbb{E}\big[\varphi(X + \sqrt{a}\, Z + \sqrt{\Delta}\, Z') - \varphi(X + \sqrt{a}\, Z)\big], \end{aligned}$$

where Z' is an independent copy of Z. A Taylor expansion gives

$$\varphi(X + \sqrt{a}\, Z + \sqrt{\Delta}\, Z') - \varphi(X + \sqrt{a}\, Z) = \varphi'(X + \sqrt{a}\, Z)\sqrt{\Delta}\, Z' + \tfrac{1}{2}\varphi''(X + \sqrt{a}\, Z)\, \Delta\, (Z')^2 + o(\Delta).$$

Since $\mathbb{E}[Z'] = 0$ and $\mathbb{E}[(Z')^2] = 1$, we obtain

$$\frac{\partial}{\partial a}\mathbb{E}_{p_a}[\varphi] = \frac{1}{2}\mathbb{E}_{p_a}[\varphi''] = \frac{1}{2}\int \varphi\, p_a''\, \mathrm{d}x \qquad \text{(integration by parts)},$$

which is exactly (5.1).

Therefore,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}a} h(X + \sqrt{a}\, Z) &= -\int (1 + \log p_a)\, \frac{\partial p_a}{\partial a}\, \mathrm{d}x = -\frac{1}{2}\int (1 + \log p_a)\, p_a''\, \mathrm{d}x \\ &= \frac{1}{2}\int \frac{(p_a')^2}{p_a}\, \mathrm{d}x \qquad \text{(integration by parts)} \\ &= \frac{1}{2} J(X + \sqrt{a}\, Z). \end{aligned}$$

□

5.2.3 Proof of EPI in dimension $d = 1$

Step 1: smoothing. Let

$$X_\lambda = X * \mathcal{N}(0, f(\lambda)), \qquad Y_\lambda = Y * \mathcal{N}(0, g(\lambda)),$$

for some functions f, g (to be chosen later). Using de Bruijn's identity,

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\big[e^{2h(X_\lambda)}\big] = 2e^{2h(X_\lambda)}\, J(X_\lambda)\, f'(\lambda).$$

Step 2: a monotone ratio. A direct computation yields

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{e^{2h(X_\lambda)}+e^{2h(Y_\lambda)}}{e^{2h(X_\lambda+Y_\lambda)}}\right]=\frac{2}{e^{2h(X_\lambda+Y_\lambda)}}\Big(e^{2h(X_\lambda)}J(X_\lambda)f'(\lambda)+e^{2h(Y_\lambda)}J(Y_\lambda)g'(\lambda)$$
$$-\big(e^{2h(X_\lambda)}+e^{2h(Y_\lambda)}\big)J(X_\lambda+Y_\lambda)\big(f'(\lambda)+g'(\lambda)\big)\Big).$$

Step 3: choose f', g' . Choosing

$$f'(\lambda)=e^{2h(X_\lambda)},\qquad g'(\lambda)=e^{2h(Y_\lambda)},$$

the Stam inequality implies

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left[\frac{e^{2h(X_\lambda)}+e^{2h(Y_\lambda)}}{e^{2h(X_\lambda+Y_\lambda)}}\right]\geq 0,\qquad \forall\lambda>0.$$

As $\lambda\to\infty$, both X_λ and Y_λ are "more and more Gaussian", hence the ratio $\to 1$. Therefore, the ratio at $\lambda=0$ must be ≤ 1, which is exactly the EPI for $d=1$.

5.2.4 General $d\geq 2$ by induction

Let $X^d=(X_1,\dots,X_d)$ and similarly Y^d, with $X^d\perp\!\!\!\perp Y^d$. Write $X^{d-1}=(X_1,\dots,X_{d-1})$ and X_d for the last coordinate. Then

$$\begin{aligned}
h(X^d+Y^d)&=h(X^{d-1}+Y^{d-1})+h\big(X_d+Y_d\mid X^{d-1}+Y^{d-1}\big)\\
&\geq h(X^{d-1}+Y^{d-1})+h\big(X_d+Y_d\mid X^{d-1},Y^{d-1}\big)\qquad\text{(conditioning reduces entropy)}\\
&\geq\frac{d-1}{2}\log\left(e^{\frac{2}{d-1}h(X^{d-1})}+e^{\frac{2}{d-1}h(Y^{d-1})}\right)\qquad\text{(induction hypothesis)}\\
&\quad+\frac{1}{2}\mathbb{E}_{X^{d-1},Y^{d-1}}\log\left(e^{2h\left(X_d|X^{d-1}=x^{d-1}\right)}+e^{2h\left(Y_d|Y^{d-1}=y^{d-1}\right)}\right)\qquad(X\perp\!\!\!\perp Y)\\
&\geq\frac{d-1}{2}\log\left(e^{\frac{2}{d-1}h(X^{d-1})}+e^{\frac{2}{d-1}h(Y^{d-1})}\right)+\frac{1}{2}\log\left(e^{2h(X_d|X^{d-1})}+e^{2h(Y_d|Y^{d-1})}\right)\\
&\geq\frac{d}{2}\log\left(e^{\frac{2}{d}\left(h(X^{d-1})+h(X_d|X^{d-1})\right)}+e^{\frac{2}{d}\left(h(Y^{d-1})+h(Y_d|Y^{d-1})\right)}\right)\\
&\qquad\text{(convexity of }(x,y)\mapsto\log(e^x+e^y)\text{ again)}\\
&=\frac{d}{2}\log\left(e^{\frac{2}{d}h(X^d)}+e^{\frac{2}{d}h(Y^d)}\right),
\end{aligned}$$

which is the EPI in dimension d.

5.2.5 Example: the entropic CLT (Barron 1986)

Let $X_1,X_2,\dots$ be i.i.d. with $\mathbb{E}[X_1]=0$, $\mathrm{Var}(X_1)=1$, and $h(X_1)>-\infty$. Let

$$T_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i$$

be the standardized sum. Then by EPI,

$$\begin{aligned} h(T_{n+m}) &= h\left(\sqrt{\frac{m}{n+m}}\,\frac{1}{\sqrt{m}}\sum_{i=1}^{m} X_i + \sqrt{\frac{n}{n+m}}\,\frac{1}{\sqrt{n}}\sum_{i=m+1}^{m+n} X_i\right) \\ &\geq \frac{1}{2}\log\left(e^{2h\left(\sqrt{\frac{m}{n+m}}T_m\right)} + e^{2h\left(\sqrt{\frac{n}{n+m}}T_n\right)}\right) \\ &= \frac{1}{2}\log\left(\frac{m}{n+m}e^{2h(T_m)} + \frac{n}{n+m}e^{2h(T_n)}\right). \end{aligned}$$

In other words, the sequence $a_n \coloneqq ne^{2h(T_n)}$ is super-additive:

$$a_{n+m} \geq a_n + a_m, \qquad \forall n, m.$$

Moreover, since $\mathrm{Var}(T_n) = 1$, the maximum entropy principle implies

$$h(T_n) \leq \frac{1}{2}\log(2\pi e), \qquad \text{so that} \qquad \frac{a_n}{n} \leq 2\pi e.$$

Therefore a_n/n must have a limit, i.e. $h(T_n) \to h^*$, and

$$D_{\mathrm{KL}}\big(P_{T_n} \,\|\, \mathcal{N}(0,1)\big) = -h(T_n) + \frac{1}{2}\log(2\pi e) \to D^*.$$

Barron (1986) shows that $D^* = 0$, a result known as the *entropic CLT*.

5.3 Information and estimation in the Gaussian model: I–MMSE

Let X be a general random variable and

$$Y_\gamma = \sqrt{\gamma}\,X + Z, \qquad Z \sim \mathcal{N}(0,1) \text{ independent of } X,$$

where $\gamma > 0$ is the SNR parameter.

Theorem 5.7 (I–MMSE)**.**

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X; Y_\gamma) = \frac{1}{2}\mathbb{E}\big[(X - \mathbb{E}[X \mid Y_\gamma])^2\big] =: \frac{1}{2}\mathrm{mmse}(X \mid Y_\gamma).$$

Notes.

(1) Perhaps the most surprising part is that this is an *equality*.

(2)
$$\mathrm{mmse}(X \mid Y) = \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2\big] = \min_f \mathbb{E}\big[(X - f(Y))^2\big]$$
is called the minimum mean squared error for estimating X based on Y.

There are several proofs for the I–MMSE formula; the most generalizable one is via SDEs.

Theorem 5.8 (A more general result)**.** *If*

$$\mathrm{d}Y_t = X_t\,\mathrm{d}t + \mathrm{d}B_t, \qquad t \in [0, T],$$

then

$$I(X^T; Y^T) = \frac{1}{2}\int_0^T \mathbb{E}\Big[\big(X_t - \mathbb{E}[X_t \mid Y^t]\big)^2\Big]\,\mathrm{d}t.$$

How it implies I–MMSE. Take $X_t \equiv X$. Then Y_T is a sufficient statistic of Y^T for estimating X, i.e.

$$I(X^T; Y^T) = I(X; Y_T), \qquad \mathbb{E}[X \mid Y^T] = \mathbb{E}[X \mid Y_T].$$

Moreover,

$$\frac{Y_T}{\sqrt{T}} = \sqrt{T}\, X + \mathcal{N}(0,1),$$

so the SNR parameter is T.

5.3.1 Two lemmas from filtering theory

Lemma 5.9 (Lemma 1). *For* $\mathrm{d}Y_t = f(t)\ \mathrm{d}t + \mathrm{d}B_t$ *with* $f(t)$ *adapted to the filtration* $\mathcal{F}_t^Y$*,*

$$\log \frac{\mathrm{d}P_{Y^T}}{\mathrm{d}P_{B^T}}(\xi^T) = \int_0^T f(t)\ \mathrm{d}\xi_t - \frac{1}{2}\int_0^T f(t)^2\ \mathrm{d}t.$$

Remark 5.10 (Intuition). For $t > 0$ and small $\Delta > 0$, the conditional distribution of $\xi_{t+\Delta} - \xi_t \mid \xi^t$ is

$$\mathcal{N}\Big(\int_t^{t+\Delta} f(s)\ \mathrm{d}s,\, \Delta \Big) \text{ under } P_{Y^T}, \qquad \mathcal{N}(0, \Delta) \text{ under } P_{B^T}.$$

So the log-likelihood ratio is

$$\frac{1}{\Delta}\Big(\int_t^{t+\Delta} f(s)\ \mathrm{d}s \Big) (\xi_{t+\Delta} - \xi_t) - \frac{1}{2\Delta}\Big(\int_t^{t+\Delta} f(s)\ \mathrm{d}s \Big)^2 \approx f(t)(\xi_{t+\Delta} - \xi_t) - \frac{\Delta}{2} f(t)^2.$$

Summing up over a partition yields the stochastic integral representation. (Think: where did we use that f is adapted to $\mathcal{F}^Y$?)

Lemma 5.11 (Lemma 2). *For* $\mathrm{d}Y_t = X_t\ \mathrm{d}t + \mathrm{d}B_t$*, define*

$$\widetilde{B}_t \coloneqq Y_t - \int_0^t \mathbb{E}[X_s \mid Y^s]\ \mathrm{d}s.$$

Then $\widetilde{B}_t$ *is a Brownian motion adapted to* $\mathcal{F}^Y$*.*

(A major difference is that X_t *could be an unknown signal not adapted to* $\mathcal{F}^Y$*; however* $\mathbb{E}[X_t \mid Y^t]$ *is always adapted to* $\mathcal{F}^Y$*.)*

Proof. Clearly $\widetilde{B}_t$ is adapted to $\mathcal{F}^Y$. In addition,

$$\widetilde{B}_t = \int_0^t \big(X_s - \mathbb{E}[X_s \mid Y^s] \big)\ \mathrm{d}s + B_t$$

is an $\mathcal{F}^Y$-adapted martingale, satisfies $\widetilde{B}_0 = 0$, and has quadratic variation t. By Lévy's criterion, $\widetilde{B}_t$ is a Brownian motion.

(Think: B_t is a BM; but is it adapted to $\mathcal{F}^Y$?) □

5.3.2 Returning to the proof of the general identity

Returning to the proof:

$$I(X^T; Y^T) = \mathbb{E}_{P_{X^T,Y^T}}\left[\log \frac{P_{Y^T|X^T}}{P_{Y^T}} \right] = \mathbb{E}_{P_{X^T,Y^T}}\left[\log \frac{P_{Y^T|X^T}}{P_{B^T}} \right] - \mathbb{E}_{P_{X^T,Y^T}}\left[\log \frac{P_{Y^T}}{P_{B^T}} \right].$$

First term. Since X^T is given (conditioned), Lemma 1 gives

$$\mathbb{E}_{P_{X^T,Y^T}}\left[\log \frac{P_{Y^T|X^T}}{P_{B^T}}\right] = \mathbb{E}\left[\int_0^T X_t \,\mathrm{d}Y_t - \frac{1}{2}\int_0^T X_t^2 \,\mathrm{d}t\right].$$

Second term. Lemma 2 tells us that $\widetilde{B}_t = Y_t - \int_0^t \mathbb{E}[X_s \mid Y^s]\,\mathrm{d}s$ is an $\mathcal{F}^Y$-BM, so Lemma 1 again yields

$$\log \frac{P_{Y^T}}{P_{B^T}}(Y^T) = \log \frac{P_{Y^T}}{P_{\widetilde{B}^T}}(Y^T) = \int_0^T \mathbb{E}[X_t \mid Y^t]\,\mathrm{d}Y_t - \frac{1}{2}\int_0^T \mathbb{E}[X_t \mid Y^t]^2\,\mathrm{d}t.$$

Combine. Therefore,

$$\begin{aligned} I(X^T;Y^T) &= \mathbb{E}\left[\int_0^T \big(X_t - \mathbb{E}[X_t \mid Y^t]\big)\,\mathrm{d}Y_t + \frac{1}{2}\int_0^T \big(\mathbb{E}[X_t \mid Y^t]^2 - X_t^2\big)\,\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_0^T \Big(\big(X_t - \mathbb{E}[X_t \mid Y^t]\big)X_t + \frac{1}{2}\big(\mathbb{E}[X_t \mid Y^t]^2 - X_t^2\big)\Big)\,\mathrm{d}t\right] \\ &= \int_0^T \frac{1}{2}\mathbb{E}\Big[\big(X_t - \mathbb{E}[X_t \mid Y^t]\big)^2\Big]\,\mathrm{d}t. \end{aligned}$$

5.3.3 Why is I–MMSE useful in statistics?

Suppose we expect a problem to have a sharp phase transition at SNR $= \gamma^*$. We can try to show that

$$I(X;Y_\gamma) \geq \frac{A\gamma}{2}(1 - o(1)) \qquad \text{for all } \gamma \leq (1-\varepsilon)\gamma^* \quad \text{(see picture)},$$

where $A \coloneqq \mathrm{mmse}(\gamma = 0)$.

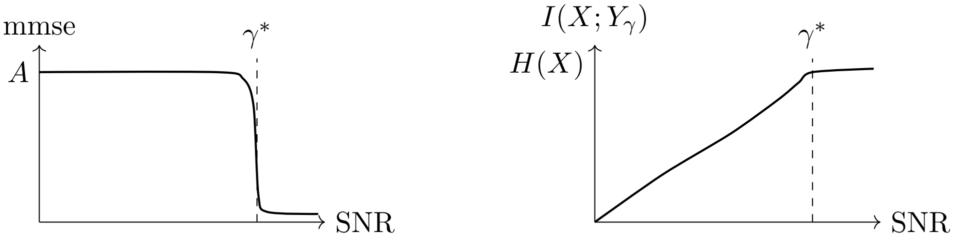


Figure 5.1: Heuristic phase transition: MMSE drops sharply around γ^*, while mutual information grows and saturates at $H(X)$.

In this case,

$$\frac{(1-\varepsilon)\gamma^*}{2}\,\mathrm{mmse}(0)(1-o(1)) \leq I\big(X;Y_{(1-\varepsilon)\gamma^*}\big) = \frac{1}{2}\int_0^{(1-\varepsilon)\gamma^*} \mathrm{mmse}(\gamma)\,\mathrm{d}\gamma.$$

Moreover, $\gamma \mapsto \mathrm{mmse}(\gamma)$ is non-increasing, so

$$\int_0^{(1-\varepsilon)\gamma^*} \mathrm{mmse}(\gamma)\,\mathrm{d}\gamma \leq (1-2\varepsilon)\gamma^*\,\mathrm{mmse}(0) + \varepsilon\gamma^*\,\mathrm{mmse}\big((1-2\varepsilon)\gamma^*\big).$$

Therefore

$$\mathrm{mmse}\big((1-2\varepsilon)\gamma^*\big) \geq (1-o(1))\mathrm{mmse}(0),$$

i.e. the MMSE does not really drop before $\gamma = \gamma^*$.

Comparison with Fano. At a high level, Fano's inequality shows that the estimation error is large when the information $I(X;Y)$ is small. Surprisingly, the I–MMSE formula shows that this is also the case if $I(X;Y)$ is "too large", and it is particularly good at showing sharp transitions and identifying the exact threshold.

5.3.4 An example: sparse mean estimation

Consider the sparse mean estimation problem:

$$Y \sim \mathcal{N}(\theta, 1), \qquad \theta \sim (1-p)\,\delta_0 + p\,\delta_\mu, \qquad p = o(1).$$

Theorem 5.12. *If* $\mu \le \sqrt{2(1-\varepsilon)\log\frac{1}{p}}$*, then*

$$\operatorname{mmse}(\theta \mid Y) \ge (1-o(1))\,\mathbb{E}[\theta^2] = (1-o(1))p\mu^2.$$

(In other words, the MMSE is essentially attained by the best estimator $\widehat{\theta} = p\mu$ *without seeing* Y*.)*

Proof sketch. Let $X \sim (1-p)\delta_0 + p\delta_1$ and set $\mu = \sqrt{\gamma}$. Then

$$Y \overset{d}{=} Y_\gamma = \sqrt{\gamma}\,X + \mathcal{N}(0,1).$$

The mutual information can be computed as

$$I(X;Y_\gamma) = \mathbb{E}\left[\log\frac{P_{Y_\gamma|X}}{P_{Y_\gamma}}\right] = \mathbb{E}\left[\log\frac{P_{Y_\gamma|X}}{Q_{Y_\gamma}}\right] - D_{\mathrm{KL}}\big(P_{Y_\gamma}\,\|\,Q_{Y_\gamma}\big) \quad \text{for any } Q.$$

Choose $Q_{Y_\gamma} = \mathcal{N}(p\sqrt{\gamma}, 1)$. Then

$$\mathbb{E}\left[\log\frac{P_{Y_\gamma|X}}{Q_{Y_\gamma}}\right] = \mathbb{E}\big[D_{\mathrm{KL}}(P_{Y_\gamma|X}\,\|\,Q_{Y_\gamma})\big] = \mathbb{E}\left[\frac{(\sqrt{\gamma}X - p\sqrt{\gamma})^2}{2}\right] = \frac{p(1-p)}{2}\,\gamma.$$

Moreover, $D_{\mathrm{KL}}(P_{Y_\gamma}\,\|\,Q_{Y_\gamma}) = o(p\gamma)$ after some algebra if $\gamma < 2(1-\varepsilon)\log\frac{1}{p}$. Hence

$$I(X;Y_\gamma) \ge \frac{p(1-p)}{2}\,\gamma\,(1-o(1)) \qquad \text{if } \gamma < 2(1-\varepsilon)\log\frac{1}{p}.$$

Now using the previous I–MMSE program proves that

$$\operatorname{mmse}(X \mid Y_\gamma) \ge (1-o(1))\operatorname{Var}(X) = (1-o(1))p \qquad \text{if } \gamma < 2(1-\varepsilon)\log\frac{1}{p}.$$

Therefore

$$\operatorname{mmse}(\theta \mid Y) = \gamma\operatorname{mmse}(X \mid Y_\gamma) \ge (1-o(1))p\gamma = (1-o(1))p\mu^2, \qquad \text{if } \mu < \sqrt{2(1-\varepsilon)\log\frac{1}{p}}.$$

5.3.5 Tensorization of I–MMSE

Theorem 5.13. *If $Y_\gamma = \sqrt{\gamma}\, X + \mathcal{N}(0, I_n)$, then*

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X; Y_\gamma) = \frac{1}{2}\mathbb{E}\big[\|X - \mathbb{E}[X \mid Y_\gamma]\|_2^2\big] =: \frac{1}{2}\mathrm{mmse}(X \mid Y_\gamma).$$

Proof. Consider the model where $Y_i = \sqrt{\gamma_i}\, X_i + \mathcal{N}(0,1)$ for possibly different $(\gamma_1, \ldots, \gamma_n)$. Then

$$\begin{aligned}\frac{\partial}{\partial \gamma_i} I(X; Y^n) &= \frac{\partial}{\partial \gamma_i} I(X_i; Y^n) + \frac{\partial}{\partial \gamma_i} I(X_{-i}; Y^n \mid X_i) \\ &= \frac{\partial}{\partial \gamma_i} I(X_i; Y_{-i}) + \frac{\partial}{\partial \gamma_i} I(X_i; Y_i \mid Y_{-i}),\end{aligned}$$

where the term $\partial_{\gamma_i} I(X_{-i}; Y^n \mid X_i)$ is zero since $\sqrt{\gamma_i} X_i$ can be subtracted from Y_i when X_i is known, and $\partial_{\gamma_i} I(X_i; Y_{-i}) = 0$ because Y_{-i} does not depend on γ_i. By the 1-D I–MMSE formula,

$$\frac{\partial}{\partial \gamma_i} I(X; Y^n) = \frac{1}{2}\mathrm{mmse}(X_i \mid Y^n).$$

Summing over i and then setting $\gamma_i \equiv \gamma$ gives

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} I(X; Y_\gamma) = \sum_{i=1}^{n} \frac{\partial}{\partial \gamma_i} I(X; Y_\gamma)\bigg|_{\gamma_i = \gamma} = \frac{1}{2}\mathrm{mmse}(X \mid Y_\gamma).$$

□

5.4 Area theorem for the BEC and sharp thresholds

5.4.1 Area theorem: a tensorization-flavored identity

Consider communication over a binary erasure channel (BEC)

$$Y = \begin{cases} X, & \text{w.p. } 1 - \varepsilon, \\ ?, & \text{w.p. } \varepsilon. \end{cases}$$

Let the input be

$$X^n \sim \mathrm{Unif}(\mathcal{C}) = \mathrm{Unif}\big(\{x_1^n, \ldots, x_M^n\}\big), \qquad M = e^{nR},$$

where $\mathcal{C}$ is the codebook. How to find a codebook such that

$$\frac{1}{n}\sum_{i=1}^{n} H(X_i \mid Y^n) \to 0 \qquad \text{when } R < C = 1 - \varepsilon?$$

(average bit error rate)

Definition 5.14 (EXIT function)**.** For $i \in [n]$, define

$$h_i(\varepsilon) := H(X_i \mid Y_{-i}), \qquad h(\varepsilon) := \frac{1}{n}\sum_{i=1}^{n} h_i(\varepsilon).$$

Lemma 5.15. *$H(X_i \mid Y^n) = \varepsilon\, h_i(\varepsilon)$.*

Proof.

$$\begin{aligned} H(X_i \mid Y^n) &= (1-\varepsilon)H(X_i \mid Y_{-i},\, Y_i \neq ?) + \varepsilon H(X_i \mid Y_{-i},\, Y_i = ?) \\ &= \varepsilon H(X_i \mid Y_{-i}) = \varepsilon\, h_i(\varepsilon). \end{aligned}$$

□

Remark 5.16. $h_i(\varepsilon)$ can be interpreted as the error probability of decoding X_i in the "non-trivial" scenario $Y_i = ?$.

Lemma 5.17.

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} H\big(X^n \mid Y(\varepsilon)^n\big) = n\, h(\varepsilon).$$

Proof. Think of n independent channels with possibly different erasure probabilities $(\varepsilon_1, \dots, \varepsilon_n)$. Then

$$\begin{aligned} \frac{\partial}{\partial \varepsilon_i} H(X^n \mid Y^n) &= \frac{\partial}{\partial \varepsilon_i} H(X_i \mid Y^n) + \frac{\partial}{\partial \varepsilon_i} H(X_{-i} \mid X_i, Y^n) \\ &= \frac{\partial}{\partial \varepsilon_i} H(X_i \mid Y^n) \qquad \text{since } H(X_{-i} \mid X_i, Y^n) = H(X_{-i} \mid X_i, Y_{-i}) \text{ (no dependence on } \varepsilon_i) \\ &= \frac{\partial}{\partial \varepsilon_i} \big(\varepsilon_i H(X_i \mid Y_{-i})\big) \qquad \text{(previous lemma)} \\ &= H(X_i \mid Y_{-i}). \end{aligned}$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon} H\big(X^n \mid Y(\varepsilon)^n\big) = \sum_{i=1}^{n} H(X_i \mid Y_{-i})\Big|_{\varepsilon_1 = \cdots = \varepsilon_n = \varepsilon} = n\, h(\varepsilon).$$

□

Theorem 5.18 (Area theorem for the BEC)**.**

$$\int_0^1 h(\varepsilon)\, \mathrm{d}\varepsilon = R.$$

Proof.

$$\begin{aligned} \int_0^1 h(\varepsilon)\, \mathrm{d}\varepsilon = \frac{1}{n} \int_0^1 \frac{\mathrm{d}}{\mathrm{d}\varepsilon} H\big(X^n \mid Y(\varepsilon)^n\big)\, \mathrm{d}\varepsilon &= \frac{H(X^n \mid Y(1)^n) - H(X^n \mid Y(0)^n)}{n} \\ = \frac{H(X^n)}{n} = R. \end{aligned}$$

□

What does the area theorem tell us? For a capacity-achieving code of rate $R = C$, it must hold that $h(\varepsilon) = o(1)$ when $\varepsilon < 1 - R$. However, since $h(\varepsilon) \le 1$ and $\int_0^1 h(\varepsilon)\, \mathrm{d}\varepsilon = R$, it must be the case that $h(\varepsilon) = 1$ for every $\varepsilon > 1 - R$, i.e. the code is really bad in the high-noise regime. Therefore any capacity-achieving code must have a sharp transition for the decoding error.

5.5 Special topic: symmetric linear codes achieve BEC capacity

Linear code. A code $\mathcal{C} = \{x_1^n, \dots, x_M^n\}$ is *linear* if it is a linear subspace of $\mathbb{F}_2^n$. (Encoding for linear codes is easy: just a matrix–vector product.)

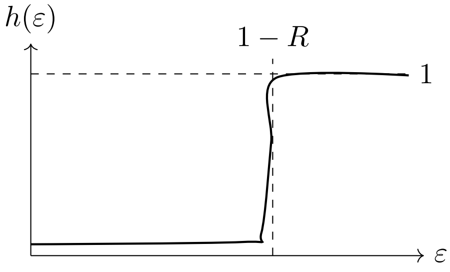


Figure 5.2: A capacity-achieving code forces a sharp transition in the EXIT curve.

Symmetry. For all $i \neq k$ and $j \neq \ell$, there exists a permutation $\pi \in S_n$ such that $\pi(i) = j$, $\pi(k) = \ell$, and

$$\pi\mathcal{C} = \mathcal{C} \qquad (\pi\mathcal{C} \text{ applies } \pi \text{ to all vectors in } \mathcal{C}).$$

Theorem 5.19. *For every symmetric linear code with $\frac{\log M}{n} \to R$, it attains the BEC capacity under the bit-MAP decoding*

$$\widehat{x}_i \coloneqq \arg\max_{x_i \in \{0,1\}} \mathbb{P}(x_i \mid y^n).$$

Remark 5.20. In the coding literature, this shows that the Reed–Muller code, which is symmetric and admits efficient encoding and decoding algorithms, is capacity-achieving.

5.5.1 Proof ingredient I: Boolean function sharp thresholds

Let $\Omega \subseteq \{0,1\}^n$. We call Ω:

(1) *Monotone:* if $x \in \Omega$ and $x \le x'$ (coordinate-wise), then $x' \in \Omega$.

(2) *Symmetric:* if for all $i, j \in [n]$, there exists $\pi \in S_n$ such that $\pi(i) = j$ and $\pi\Omega = \Omega$.

For $\varepsilon \in [0,1]$, define

$$p_\varepsilon(\Omega) \coloneqq \mathbb{P}\big(\mathrm{Bern}(\varepsilon)^{\otimes n} \in \Omega\big).$$

By monotonicity, $\varepsilon \mapsto p_\varepsilon(\Omega)$ is non-decreasing. For symmetry, we shall only need that all influence functions of Ω are the same, i.e. $I_1(\Omega) = \cdots = I_n(\Omega)$, where

$$I_i(\Omega) \coloneqq \mathbb{P}_\varepsilon\Big(x \in \{0,1\}^n : (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \notin \Omega \text{ and } (x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_n) \in \Omega\Big).$$

Let

$$\varepsilon(\delta) \coloneqq \max\{\varepsilon : p_\varepsilon(\Omega) \le \delta\}.$$

Theorem 5.21.

$$\varepsilon(1-\delta) - \varepsilon(\delta) = o(1), \qquad \forall\, \delta \in (0, 1/2).$$

(So $\varepsilon \mapsto p_\varepsilon(\Omega)$ has a sharp threshold.)

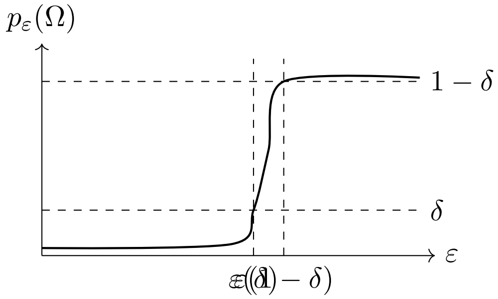


Figure 5.3: Sharp threshold phenomenon for a monotone, symmetric Boolean set Ω.

Proof sketch. A classical result shows that

$$\frac{\mathrm{d}}{\mathrm{d}\varepsilon}p_\varepsilon(\Omega) = \sum_{i=1}^{n} I_i(\Omega) = nI_1(\Omega) \qquad \text{(by symmetry)}.$$

It remains to show that $nI_1(\Omega) = \omega(1)$ whenever $p_\varepsilon(\Omega) \in [\delta, 1-\delta]$.
Classical Efron–Stein bound:

$$p_\varepsilon(\Omega)\big(1-p_\varepsilon(\Omega)\big) \le \sum_{i=1}^{n} I_i(\Omega)$$

only shows $nI_1(\Omega) = \Omega(1)$.
Key improvement (KKL theorem):

$$\frac{\log n}{n}\, p_\varepsilon(\Omega)\big(1-p_\varepsilon(\Omega)\big) \le \max\{I_1(\Omega), \dots, I_n(\Omega)\}$$

(essentially the log-Sobolev inequality on the hypercube) implies $nI_1(\Omega) = \Omega(\log n) = \omega(1)$.

5.5.2 Proof ingredient II: area theorem + sharp threshold ⇒ capacity

For a given linear code $\mathcal{C}$, define

$$\Omega_i := \Big\{\text{all erasure patterns } w \in \{0,1\}^{n-1} \text{ such that } w \odot x_{-i} \text{ fails to decode } x_i \text{ for some } x \in \mathcal{C}\Big\},$$

where 1 represents erasure and 0 represents non-erasure.

Since $\mathcal{C}$ is linear, WLOG assume that the transmitted codeword is $x = 0$, i.e.

$$\Omega_i = \Big\{w \in \{0,1\}^{n-1} :\ \exists\, x_{-i} \le w \text{ s.t. } (x_{-i}, 1) \in \mathcal{C}\Big\}.$$

Then:

(1) Ω_i is monotone (obvious).

(2) Ω_i is symmetric (follows from symmetry of $\mathcal{C}$).

(3) $p_\varepsilon(\Omega_i) = \mathbb{P}(Y_{-i} \text{ fails to decode } X_i) = h_i(\varepsilon)$.

(4) $h_i(\varepsilon) = h(\varepsilon)$ (symmetry of $\mathcal{C}$ again).

By the previous part, $\varepsilon \mapsto h(\varepsilon) = p_\varepsilon(\Omega_i)$ has a sharp threshold. In addition, $\int_0^1 h(\varepsilon)\,\mathrm{d}\varepsilon = R$ by the area theorem. This threshold can only be

$$\varepsilon^* = 1 - R,$$

i.e. the code is capacity-achieving.

Lecture 6: Statistical decision theory & classical asymptotics

6.1 Statistical decision theory

Definition 6.1 (Statistical model)**.** A *statistical model* is a family of distributions $(P_\theta)_{\theta\in\Theta}$.

- **Parametric:** $\dim(\Theta) < \infty$.
- **Nonparametric:** $\dim(\Theta) = \infty$.
- **Semiparametric:** $\Theta = \Theta_1 \times \Theta_2$ with $\dim(\Theta_1) < \infty$ and $\dim(\Theta_2) = \infty$.

Observation. We observe $X \sim P_\theta$ with an unknown $\theta \in \Theta$.

Decision rule / estimator. A (possibly randomized) decision rule is a map

$$\hat{\theta} : \mathcal{X} \to \mathcal{A},$$

where $\mathcal{A}$ is the *action space.*

Loss. A loss is a given function $L : \Theta \times \mathcal{A} \to \mathbb{R}_+$.

Risk (expected loss). The risk of an estimator $\hat{\theta}$ under L is

$$r(\hat{\theta}, \theta) = \mathbb{E}_{X\sim P_\theta}\big[L(\theta, \hat{\theta}(X))\big].$$

We often abbreviate $\mathbb{E}_{X\sim P_\theta}[\cdot]$ as $\mathbb{E}_\theta[\cdot]$.

Although originally proposed by Wald for statistical estimation, this framework is general enough to encapsulate many other scenarios.

Example 6.2 (Density estimation)**.** Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f$ be i.i.d. from an unknown density f. Then the parameter is $\theta = f$ and $P_\theta = f^{\otimes n}$. Different losses capture different goals, such as

$$\begin{aligned}
&\text{Density at a point:} && L_1(f, a) = |a - f(0)|,\\
&\text{Global estimation:} && L_2(f, a) = \int |f(x) - a(x)|^2 \,\mathrm{d}x,\\
&\text{Functional estimation:} && L_3(f, a) = \Big|a - \int h\big(f(x)\big)\,\mathrm{d}x\Big|.
\end{aligned}$$

Example 6.3 (Linear regression)**.** Let $X_1, \dots, X_n$ be either fixed or random design points, and let $P_{Y|X}$ satisfy

$$\mathbb{E}[Y \mid X] = \langle \theta, X \rangle .$$

Losses include

$$\text{Estimation error:} \qquad L_1(\theta, \hat{\theta}) = \left\| \hat{\theta} - \theta \right\|^2 ,$$

$$\text{Prediction error:} \qquad L_2(\theta, \hat{\theta}) = \mathbb{E}_{X \sim P_X} \left[\left(\langle \theta, X \rangle - \left\langle \hat{\theta}, X \right\rangle \right)^2 \right].$$

Example 6.4 (Learning theory)**.** Let $(X_1, Y_1), \dots, (X_n, Y_n) \sim P_{XY}$. A loss that captures *excess risk* w.r.t. a given function class $\mathcal{F}$ is

$$L(P_{XY}, \hat{f}) = \mathbb{E}_{P_{XY}} \left[(Y - \hat{f}(X))^2 \right] - \inf_{f \in \mathcal{F}} \mathbb{E}_{P_{XY}} \left[(Y - f(X))^2 \right].$$

Example 6.5 (Optimization)**. Parameter:** a function f to be minimized.
Action: a query strategy $x_{t+1} = \phi(x^t, y^t)$.
Observation: queries x^t and answers y^t (e.g. $y_t = f(x_t) + \xi_t$).
Loss:

$$L(f, x_{T+1}) = f(x_{T+1}) - \min f.$$

6.2 Comparison of estimators

For an estimator $\hat{\theta}$, recall that its risk $r(\hat{\theta}, \theta)$ is a function of θ. How to compare two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$?

Option I: (In)admissibility

$\hat{\theta}_2$ is *inferior* to $\hat{\theta}_1$ if

$$r(\hat{\theta}_2; \theta) \geq r(\hat{\theta}_1; \theta) \quad \text{for every } \theta \in \Theta,$$

and

$$r(\hat{\theta}_2; \theta) > r(\hat{\theta}_1; \theta) \quad \text{for some } \theta.$$

In this case, $\hat{\theta}_2$ is called *inadmissible.* However, admissibility is a weak notion: even the constant estimator $\hat{\theta} \equiv 0$ can be admissible.

Option II: Bayes risk

Given a probability distribution $\pi(\theta)$ on Θ, look at the weighted average

$$r_\pi(\hat{\theta}) = \int \pi(\theta) \, r(\hat{\theta}; \theta) \, \mathrm{d}\theta.$$

The distribution π is called the *prior.* The minimizer of $\hat{\theta} \mapsto r_\pi(\hat{\theta})$ is called the *Bayes estimator* under π.

Option III: Minimax risk

Look at the worst-case risk

$$r^*(\hat{\theta}) = \sup_\theta r(\hat{\theta}; \theta).$$

The minimizer of $\hat{\theta} \mapsto r^*(\hat{\theta})$ is called the *minimax estimator.*

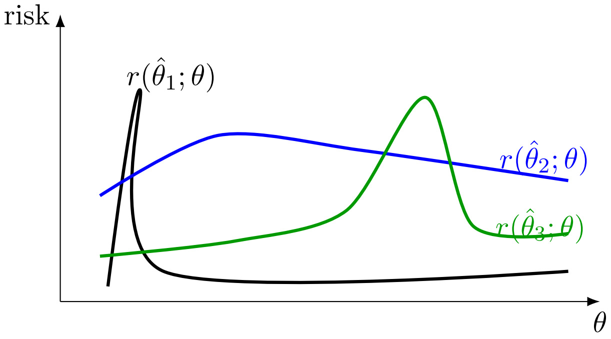


Figure 6.1: A schematic plot of several risk functions, matching the qualitative sketch in the notes.

6.3 Bayes risk vs minimax risk

Define the *Bayes risk*

$$r_\pi^* := \inf_{\hat\theta} r_\pi(\hat\theta) = \inf_{\hat\theta} \mathbb{E}_{\theta\sim\pi}\big[r(\hat\theta;\theta)\big],$$

and the *minimax risk*

$$r^* := \inf_{\hat\theta} r^*(\hat\theta) = \inf_{\hat\theta}\sup_{\theta} r(\hat\theta;\theta).$$

Theorem.

(1) $r^* \ge r_\pi^*$ for every prior π.

(2) Under regularity conditions (a minimax theorem),

$$r^* = \sup_\pi r_\pi^*.$$

The maximizer π^* is called the *least favorable prior.*

Proof. First, for any fixed estimator $\hat\theta$,

$$\sup_\theta r(\hat\theta;\theta) \ge \mathbb{E}_{\theta\sim\pi}[r(\hat\theta;\theta)] \qquad (\text{max } \ge \text{ average}),$$

hence taking $\inf_{\hat\theta}$ on both sides yields $r^* \ge r_\pi^*$.

For the other direction, recall that a *randomized* estimator can be viewed as a conditional distribution $p(\cdot \mid X)$ over actions. Then

$$\begin{aligned}\sup_\pi r_\pi^* &= \sup_\pi \inf_p \mathbb{E}_{\theta\sim\pi}\mathbb{E}_X\mathbb{E}_{a\sim p(\cdot|X)}\big[L(\theta,a)\big] && (\text{affine in both } \pi \text{ and } p)\\ &= \inf_p \sup_\pi \mathbb{E}_{\theta\sim\pi}\mathbb{E}_X\mathbb{E}_{a\sim p(\cdot|X)}\big[L(\theta,a)\big] && (\text{by Sion's minimax theorem})\\ &= \inf_p \sup_\theta \mathbb{E}_X\mathbb{E}_{a\sim p(\cdot|X)}\big[L(\theta,a)\big] = r^*.\end{aligned}$$

□

Posterior and Bayes estimator. Given a prior $\pi(\theta)$, it induces a joint distribution $\pi(\theta)p_\theta(x)$ on (Θ, X), which admits the posterior

$$\pi(\theta \mid x) \propto \pi(\theta)p_\theta(x).$$

The Bayes estimator is the barycenter of $\pi(\theta \mid X)$ under L, i.e.

$$\hat{\theta}_\pi(X) = \arg\min_a \mathbb{E}_{\theta\sim\pi(\cdot|X)}\big[L(\theta, a)\big].$$

Finding the Bayes estimator is often statistically easy (it is an expectation under the posterior), but can be computationally hard. Finding the minimax estimator can be statistically hard and is only feasible in a few examples. This motivates studying asymptotically minimax estimators (second part of the lecture) or rate-optimal results, namely to find $\hat{\theta}$ such that

$$r^*(\hat{\theta}) \le C\, r^* \qquad \text{for some constant } C.$$

6.4 Examples

6.4.1 Binomial model

Example 6.6 (Binomial)**.** Let $X \sim \mathrm{Bin}(n, \theta)$ and $L(\theta, a) = (\theta - a)^2$. To find the least favorable prior, try

$$\pi(\theta) \propto \theta^{b-1}(1-\theta)^{b-1} \qquad (\mathrm{Beta}(b, b)).$$

Then the posterior is

$$\begin{aligned}\pi(\theta \mid X) &\propto \pi(\theta)\,\theta^X(1-\theta)^{n-X} \\ &= \theta^{b+X-1}(1-\theta)^{b+n-X-1} \qquad (\mathrm{Beta}(b+X,\, b+n-X)).\end{aligned}$$

The Bayes estimator is

$$\hat{\theta}(X) = \mathbb{E}_\pi[\theta \mid X] = \frac{X+b}{n+2b}.$$

Its risk is

$$\begin{aligned}r(\hat{\theta}, \theta) &= \mathbb{E}_\theta\big[(\hat{\theta} - \theta)^2\big] = \mathrm{Bias}^2 + \mathrm{Var} \\ &= \Big(\frac{n\theta + b}{n+2b} - \theta\Big)^2 + \frac{n\theta(1-\theta)}{(n+2b)^2} \\ &= \frac{1}{(n+2b)^2}\Big[b^2 + (n - 4b^2)\,\theta(1-\theta)\Big].\end{aligned}$$

By choosing $b = \frac{\sqrt{n}}{2}$, we have $n - 4b^2 = 0$, hence

$$r(\hat{\theta}, \theta) \equiv \frac{1}{4(\sqrt{n}+1)^2}.$$

Therefore

$$\hat{\theta} = \frac{X + \frac{\sqrt{n}}{2}}{n + \sqrt{n}}$$

attains the worst-case risk $r^*(\hat{\theta}) = \frac{1}{4(\sqrt{n}+1)^2}$, and

$$r^* \le r^*(\hat{\theta}) = r_\pi(\hat{\theta}) = r_\pi^* \le r^* \quad \Longrightarrow \quad r^* = \frac{1}{4(\sqrt{n}+1)^2}.$$

6.4.2 Gaussian location model with bowl-shaped loss

Setup. Let $X \sim \mathcal{N}(\theta, I_d)$ and let

$$L(\theta, a) = \rho(\theta - a),$$

where $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is continuous and *bowl-shaped* (i.e. $\rho(x) = \rho(-x)$ and ρ is quasi-convex).

Example 6.7 (Gaussian location: minimax estimator)**. Claim.** $\hat{\theta} = X$ is the minimax estimator, with minimax risk

$$r^* = \mathbb{E}\big[\rho(Z)\big], \qquad Z \sim \mathcal{N}(0, I_d).$$

Proof (via a Gaussian prior). Try a prior $\pi = \mathcal{N}(0, \tau^2 I_d)$. Then

$$\pi(\theta \mid X) \propto \exp\Big(-\frac{\|\theta\|^2}{2\tau^2} - \frac{\|X-\theta\|^2}{2}\Big) = \mathbb{N}\Big(\frac{\tau^2}{1+\tau^2}X,\ \frac{\tau^2}{1+\tau^2}I_d\Big).$$

Hence

$$\begin{aligned} r^* \geq r^*_\pi &= \mathbb{E}_X\Big[\min_{a\in\mathbb{R}^d} \mathbb{E}_{\theta\sim\mathbb{N}\left(\frac{\tau^2}{1+\tau^2}X,\ \frac{\tau^2}{1+\tau^2}I_d\right)} \rho(\theta - a)\Big] \\ &= \mathbb{E}\Big[\rho\Big(\sqrt{\frac{\tau^2}{1+\tau^2}}\,Z\Big)\Big] \qquad \text{(by Anderson's lemma below).} \end{aligned}$$

Letting $\tau \to \infty$ gives $r^* \geq \mathbb{E}[\rho(Z)]$. On the other hand, the estimator $\hat{\theta} = X$ has constant risk

$$\mathbb{E}_\theta\big[\rho(X-\theta)\big] = \mathbb{E}[\rho(Z)],$$

so it achieves the lower bound and is minimax. □

Lemma 6.8 (Anderson)**.** *If $X \sim \mathcal{N}(0, \Sigma)$ and ρ is bowl-shaped, then*

$$\min_{a\in\mathbb{R}^d} \mathbb{E}\big[\rho(X+a)\big] = \mathbb{E}\big[\rho(X)\big].$$

Proof. Let $K_c = \{x : \rho(x) \leq c\}$. Since ρ is bowl-shaped, K_c is convex and $K_c = -K_c$. Then

$$\begin{aligned} \mathbb{E}[\rho(X+a)] &= \int_0^\infty \mathbb{P}\big(\rho(X+a) > c\big)\,\mathrm{d}c \\ &= \int_0^\infty \big(1 - \mathbb{P}(X + a \in K_c)\big)\,\mathrm{d}c \\ &\geq \int_0^\infty \big(1 - \mathbb{P}(X \in K_c)\big)\,\mathrm{d}c \qquad \text{(see the comparison below)} \\ &= \mathbb{E}[\rho(X)]. \end{aligned}$$

It remains to justify that $\mathbb{P}(X \in K_c) \geq \mathbb{P}(X \in K_c + a)$. Using convexity of K_c,

$$K_c = \tfrac{1}{2}(K_c + a) + \tfrac{1}{2}(K_c - a),$$

hence

$$\begin{aligned} \mathbb{P}(X \in K_c) &= \mathbb{P}\Big(X \in \tfrac{1}{2}(K_c + a) + \tfrac{1}{2}(K_c - a)\Big) \\ &\geq \sqrt{\mathbb{P}(X \in K_c + a)\,\mathbb{P}(X \in K_c - a)} \qquad (X \text{ has a log-concave distribution}) \\ &= \sqrt{\mathbb{P}(X \in K_c + a)\,\mathbb{P}(X \in -K_c - a)} \qquad (K_c = -K_c) \\ &= \mathbb{P}(X \in K_c + a) \qquad \text{(the distribution of } X \text{ is symmetric around 0).} \end{aligned}$$

This proves the claim. □

6.5 Hájek–Le Cam classical asymptotics

We now consider $X_1, \dots, X_n \sim P_\theta$ with $n \to \infty$.

6.5.1 Regular models: differentiable in quadratic mean (QMD)

Definition 6.9 (QMD)**.** A statistical model $(P_\theta)_{\theta\in\Theta}$ is called *differentiable in quadratic mean (QMD)* at θ if there exists a *score function* $s_\theta(x)$ such that

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \tfrac{1}{2} h^T s_\theta \sqrt{p_\theta}\right]^2 \mathrm{d}\mu = o(\|h\|^2),$$

where μ is any dominating measure for (P_θ) and $p_\theta = \frac{\mathrm{d}P_\theta}{\mathrm{d}\mu}$.

Remark 6.10. (1) When $h \mapsto \sqrt{p_{\theta+h}}$ is differentiable everywhere,

$$s_\theta(x) = \frac{2}{\sqrt{p_\theta(x)}} \frac{\partial}{\partial\theta} \sqrt{p_\theta(x)} = \frac{\partial_\theta p_\theta(x)}{p_\theta(x)} = \partial_\theta \log p_\theta(x).$$

(2) Since

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_\theta}\right)^2 \mathrm{d}\mu = H^2(P_{\theta+h}, P_\theta) \le 2,$$

QMD implies that the Fisher information

$$I(\theta) \coloneqq \mathbb{E}_\theta[s_\theta s_\theta^T]$$

exists.

6.6 Fisher's program and Hodges' estimator

Historically, a major goal of classical asymptotics was Fisher's program:

(1) The MLE $\hat{\theta}_n$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \mathbb{N}\big(0, I(\theta)^{-1}\big),$$

where $I(\theta)$ is the Fisher information matrix of $(P_\theta)_{\theta\in\Theta}$.

(2) For any other sequence of estimators (T_n) with

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \Sigma_\theta), \qquad \forall \theta \in \Theta,$$

we must have $\Sigma_\theta \succeq I(\theta)^{-1}$. (In other words, the MLE attains the asymptotically smallest variance.)

While (1) is true under mild regularity conditions, (2) is *not* true in full generality, as witnessed by Hodges' estimator (1951).

Example 6.11 (Hodges' estimator)**.** Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1)$ and let $\bar{X}_n$ be the sample mean. Define

$$\hat{\theta}_n = \begin{cases} \bar{X}_n, & |\bar{X}_n| \ge n^{-1/4}, \\ 0, & |\bar{X}_n| < n^{-1/4}. \end{cases}$$

It is easy to show that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \begin{cases} \mathcal{N}(0,1), & \theta \neq 0, \\ 0, & \theta = 0. \end{cases}$$

Therefore Fisher's statement (2) does not hold when $\theta = 0$.

Hodges' example shows that caution is needed when defining the "optimality" of the MLE or inverse Fisher information. It then took statisticians roughly 20 years to find the right definitions, through angles such as:

(1) Hodges' estimator is not *regular* (one restricts the class of estimators).

(2) The set of violations has Lebesgue measure 0 ("superefficiency" occurs rarely).

(3) The performance of Hodges' estimator is bad when $\theta \approx n^{-1/4}$ (a large asymptotic local risk).

6.7 A collection of asymptotic theorems

Convolution theorem. Let (P_θ) be QMD. If

$$\sqrt{n}(T_n - \psi(\theta)) \xrightarrow{\mathrm{d}} L_\theta \quad \text{under } P_\theta^{\otimes n},$$

and (T_n) is *regular* in the sense that

$$\sqrt{n}\Big(T_n - \psi\Big(\theta + \frac{h}{\sqrt{n}}\Big)\Big) \xrightarrow{\mathrm{d}} L_\theta \quad \text{under } P^{\otimes n}_{\theta+\frac{h}{\sqrt{n}}},\ \forall h \in \mathbb{R}^d,$$

then there exists a probability measure M_θ such that

$$L_\theta = \mathbb{N}\big(0, \nabla\psi(\theta)^T I(\theta)^{-1} \nabla\psi(\theta)\big) * M_\theta, \qquad \forall\theta.$$

Here $*$ denotes convolution: $(\mu * \nu)(A) = \int \mu(\,\mathrm{d}x)\,\nu(A - x)$.
(Convolution makes the distribution more "noisy".)

Almost everywhere convolution theorem. Under all of the above conditions *except* for regularity of (T_n), we still have

$$L_\theta = \mathbb{N}\big(0, \nabla\psi(\theta)^T I(\theta)^{-1} \nabla\psi(\theta)\big) * M_\theta$$

for Lebesgue-almost-every θ.

Local asymptotic minimax (LAM) theorem. For every continuous bowl-shaped loss ρ and any sequence of estimators (T_n),

$$\lim_{c\to\infty} \liminf_{n\to\infty} \sup_{\|h\|\le c} \mathbb{E}_{\theta+\frac{h}{\sqrt{n}}}\Big[\rho\Big(\sqrt{n}\big(T_n - \psi(\theta + \tfrac{h}{\sqrt{n}})\big)\Big)\Big] \ge \mathbb{E}[\rho(Z)],$$

where

$$Z \sim \mathbb{N}\big(0, \nabla\psi(\theta)^T I(\theta)^{-1} \nabla\psi(\theta)\big).$$

(This is a lower bound on the minimax risk of the local family $(P_{\theta+h/\sqrt{n}})_{\|h\|\le c}$ under the loss $L(\theta, a) = \rho(\sqrt{n}(a - \psi(\theta)))$.)

The proofs rely on the asymptotic equivalence between models $(P_{\theta+h/\sqrt{n}})_{\|h\|\le c}$ and the Gaussian shift model $(\mathcal{N}(h, I(\theta)^{-1}))_{\|h\|\le c}$; see the special topic at the end of this lecture.

6.8 A special case of LAM via Bayesian Cramér–Rao

6.8.1 Bayesian Cramér–Rao in one dimension (van Trees inequality)

Let $\theta \in [a, b]$, and let $\pi(\cdot)$ be a differentiable prior density on $[a, b]$ with $\pi(a) = \pi(b) = 0$ and

$$J(\pi) = \int_a^b \frac{\pi'(\theta)^2}{\pi(\theta)}\,\mathrm{d}\theta < \infty.$$

Then for any estimator $\hat\theta$,

$$\mathbb{E}_\pi \mathbb{E}_\theta\big[(\hat\theta - \theta)^2\big] \ge \frac{1}{\mathbb{E}_\pi[I(\theta)] + J(\pi)}.$$

(Compare with the usual Cramér–Rao bound $\mathbb{E}_\theta[(\hat\theta - \theta)^2] \ge 1/I(\theta)$ for unbiased $\hat\theta$.)

Proof. Consider

$$\begin{aligned}
\mathbb{E}_\pi \mathbb{E}_\theta\Big[(\hat\theta - \theta)\,\partial_\theta\big(\log \pi(\theta) p_\theta(X)\big)\Big] &= \int_{\mathcal{X}} \int_a^b (\hat\theta - \theta)\,\partial_\theta\big(\pi(\theta) p_\theta(x)\big)\,\mathrm{d}\theta\,\mu(\,\mathrm{d}x) \\
&= \int_{\mathcal{X}} \int_a^b \pi(\theta) p_\theta(x)\,\mathrm{d}\theta\,\mu(\,\mathrm{d}x) \qquad \text{(integration by parts)} \\
&= 1.
\end{aligned}$$

Then, by Cauchy–Schwarz,

$$1 \le \big(\mathbb{E}_\pi \mathbb{E}_\theta[(\hat\theta - \theta)^2]\big) \cdot \Big(\mathbb{E}_\pi \mathbb{E}_\theta\big[\partial_\theta \log(\pi(\theta) p_\theta(X))\big]^2\Big).$$

Next expand

$$\begin{aligned}
\mathbb{E}_\pi \mathbb{E}_\theta\big[\partial_\theta \log(\pi(\theta) p_\theta(X))\big]^2 = \mathbb{E}_\pi\Big[\Big(\frac{\pi'(\theta)}{\pi(\theta)}\Big)^2\Big] + \mathbb{E}_\pi \mathbb{E}_\theta\Big[\Big(\frac{\partial_\theta p_\theta(X)}{p_\theta(X)}\Big)^2\Big] \\
+ 2\mathbb{E}_\pi \mathbb{E}_\theta\Big[\frac{\pi'(\theta)}{\pi(\theta)} \cdot \frac{\partial_\theta p_\theta(X)}{p_\theta(X)}\Big].
\end{aligned}$$

The cross term equals 0 assuming

$$\int \mu(\,\mathrm{d}x)\,\partial_\theta p_\theta(x) = \partial_\theta \int \mu(\,\mathrm{d}x)\,p_\theta(x) = 0.$$

Therefore

$$\mathbb{E}_\pi \mathbb{E}_\theta\big[\partial_\theta \log(\pi(\theta) p_\theta(X))\big]^2 = J(\pi) + \mathbb{E}_\pi[I(\theta)],$$

and the inequality follows. □

6.8.2 Multivariate Bayesian Cramér–Rao

Statement. Let $\pi = \prod_{i=1}^d \pi_i$ be a differentiable prior density on $\prod_{i=1}^d [a_i, b_i]$ vanishing on the boundary, and let $J(\pi) = \operatorname{diag}(J(\pi_1), \ldots, J(\pi_d))$. Then for any estimator $\hat{\theta}$,

$$\mathbb{E}_\pi \mathbb{E}_\theta \left[\left\| \hat{\theta} - \theta \right\|^2 \right] \geq \operatorname{Tr}\Big(\big(\mathbb{E}_\pi [I(\theta)] + J(\pi) \big)^{-1} \Big).$$

Key step (as in the notes). Similar to the 1-D proof, one can show for each $k = 1, \ldots, d$ that

$$\mathbb{E}_\pi \mathbb{E}_\theta \Big[(\hat{\theta}_k - \theta_k) \, \nabla_\theta \log(\pi(\theta) p_\theta(X)) \Big] = e_k$$

(the k-th standard basis vector). Let

$$\Sigma = \mathbb{E}\Big[\nabla_\theta \log(\pi(\theta) p_\theta(X)) \, \nabla_\theta \log(\pi(\theta) p_\theta(X))^T \Big] = \mathbb{E}_\pi[I(\theta)] + J(\pi).$$

Then by Cauchy–Schwarz,

$$\mathbb{E}_\pi \mathbb{E}_\theta \big[(\hat{\theta}_k - \theta_k)^2 \big] \geq \sup_{u \neq 0} \frac{\langle u, e_k \rangle^2}{u^T \Sigma u} = (\Sigma^{-1})_{kk}.$$

6.8.3 Deriving LAM from BCR when $\psi(\theta) = \theta$ and $\rho(x) = \|x\|^2$

First, note that if

$$\pi(\theta) = \frac{2}{b-a} \cos^2 \Big(\frac{\pi}{2} \cdot \frac{2\theta - (a+b)}{b-a} \Big),$$

then $\pi(a) = \pi(b) = 0$, and

$$J(\pi) = \int_a^b \frac{8\pi^2}{(b-a)^3} \sin^2 \Big(\frac{\pi}{2} \cdot \frac{2\theta - (a+b)}{b-a} \Big) \, \mathrm{d}\theta = \frac{4\pi^2}{(b-a)^2}.$$

(Exercise: show that this choice of π minimizes the value of $J(\pi)$.)

Next, choosing the above prior on $[\theta_0 - \frac{c}{\sqrt{n}}, \theta_0 + \frac{c}{\sqrt{n}}]$, Bayesian Cramér–Rao gives

$$\begin{aligned} \inf_{\hat{\theta}} \sup_{\|h\|_\infty \leq c} \mathbb{E}_{\theta_0 + \frac{h}{\sqrt{n}}} \left[\left\| \hat{\theta} - (\theta_0 + \tfrac{h}{\sqrt{n}}) \right\|^2 \right] &\geq \inf_{\hat{\theta}} \; \mathbb{E}_\pi \mathbb{E}_{\theta_0 + \frac{h}{\sqrt{n}}} \left[\left\| \hat{\theta} - (\theta_0 + \tfrac{h}{\sqrt{n}}) \right\|^2 \right] \\ &\geq \operatorname{Tr}\Big(\big(n \, \mathbb{E}_\pi[I(\theta)] + \frac{n\pi^2}{c^2} I \big)^{-1} \Big) \qquad \text{(Fisher info. for } n \text{ samples is } nI(\theta)) \\ &= \frac{1 + o(1)}{n} \operatorname{Tr}(I(\theta_0)^{-1}) \qquad \text{as } n \to \infty \text{ and } c \to \infty, \end{aligned}$$

assuming that $\theta \mapsto I(\theta)$ is continuous at θ_0.

6.9 Applications of LAM

Since the global minimax risk is always lower bounded by the local minimax risk, LAM gives asymptotic lower bounds on r_n^*.

6.9.1 Binomial revisit

Revisit $X \sim \mathrm{Bin}(n,\theta)$. Then

$$\begin{aligned} r_n^* &= \inf_{\hat\theta} \sup_{\theta\in(0,1)} \mathbb{E}_\theta\big[(\hat\theta-\theta)^2\big] \\ &\ge \inf_{\hat\theta} \sup_{|h|\le c_n} \mathbb{E}_{\frac12+\frac{h}{\sqrt n}}\Big[\big(\hat\theta-(\tfrac12+\tfrac{h}{\sqrt n})\big)^2\Big] \qquad (c_n\to\infty \text{ as } n\to\infty) \\ &\ge \frac{1-o_n(1)}{nI(\frac12)} = \frac{1-o_n(1)}{4n}. \end{aligned}$$

This is consistent with the exact expression $r_n^* = \frac{1}{4(\sqrt n+1)^2}$.

6.9.2 Nonparametric entropy estimation

Let $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} f$, where f is a density on $[0,1]$. The target is to estimate the differential entropy

$$h(f) = \int_0^1 -f(x)\log f(x)\,\mathrm{d}x$$

under the squared loss.

Challenge. This is not a finite-dimensional model, so LAM does not directly apply.

Solution. Consider a one-parameter subfamily $(f_0+tg)_{|t|\le\varepsilon}$. Then

$$I(0) = \int_0^1 \frac{g(x)^2}{f_0(x)}\,\mathrm{d}x, \qquad \frac{\mathrm{d}}{\mathrm{d}t}h(f_0+tg)\Big|_{t=0} = -\int_0^1 \big(1+\log f_0(x)\big)g(x)\,\mathrm{d}x.$$

LAM applied to this subfamily at $t=0$ gives

$$\begin{aligned} r_n^* &\ge \frac{1-o_n(1)}{n}\Big(\int_0^1 \frac{g(x)^2}{f_0(x)}\,\mathrm{d}x\Big)^{-1}\Big(\int_0^1 \big(1+\log f_0(x)\big)g(x)\,\mathrm{d}x\Big)^2 \\ &=: \frac{1-o_n(1)}{n}\,V(f_0,g). \end{aligned}$$

We can maximize this lower bound w.r.t. g. Since $\int g = 0$ (because f_0+tg must remain a density), Cauchy–Schwarz gives

$$\begin{aligned} V(f_0,g) &= \Big(\int_0^1 \frac{g(x)^2}{f_0(x)}\,\mathrm{d}x\Big)^{-1}\Big(\int_0^1 \big(\log f_0(x)+h(f_0)\big)g(x)\,\mathrm{d}x\Big)^2 \\ &\le \int_0^1 f_0(x)\big(\log f_0(x)+h(f_0)\big)^2\,\mathrm{d}x \\ &= \int_0^1 f_0(x)\log^2 f_0(x)\,\mathrm{d}x - h(f_0)^2, \end{aligned}$$

where equality holds when $g(x) = f_0(x)\big(\log f_0(x)+h(f_0)\big)$.

Therefore,

$$r_n^* \ge \frac{1-o_n(1)}{n}\sup_{f_0}\Big(\int_0^1 f_0(x)\log^2 f_0(x)\,\mathrm{d}x - h(f_0)^2\Big).$$

6.9.3 Pros and cons for asymptotic theorems

- **Pro 1:** plug-and-play bound for essentially all statistical models.
- **Pro 2:** exact constant for the asymptotic risk.
- **Con 1:** bounds are asymptotic, assuming $n \to \infty$ while d is fixed.
- **Con 2:** bounds are for asymptotic variance, while for high-dimensional scenarios bias can be the dominating factor.

This motivates studying techniques for non-asymptotic lower bounds in the next few lectures.

6.10 Special topic: Le Cam's distance between statistical models

Ref: Liese and Miescke, *Statistical Decision Theory*, Springer (2008).

For two models $(P_\theta)_{\theta\in\Theta}$ and $(Q_\theta)_{\theta\in\Theta}$ with the same parameter set Θ, how do we compare their "strength"? Throughout this section we assume that Θ is a finite set.

Definition 6.12 (Deficiency)**.** A model $\mathcal{M} = (P_\theta)_{\theta\in\Theta}$ is called ε-deficient with respect to $\mathcal{N} = (Q_\theta)_{\theta\in\Theta}$ if

- for every finite decision space $\mathcal{A}$,
- for every bounded loss $L(\theta, a) \in [0, 1]$,
- for every (randomized) estimator $\hat{\theta}_{\mathcal{N}}$ under $\mathcal{N}$,

there exists an estimator $\hat{\theta}_{\mathcal{M}}$ under $\mathcal{M}$ such that

$$r(\hat{\theta}_{\mathcal{M}}; \theta) \le r(\hat{\theta}_{\mathcal{N}}; \theta) + \varepsilon, \qquad \forall \theta \in \Theta.$$

Theorem 6.13 (Randomization criterion)**.** *The following are equivalent:*

(1) $\mathcal{M}$ is ε-deficient w.r.t. $\mathcal{N}$.

(2) For every finite action set $\mathcal{A}$, bounded loss $L(\theta, a) \in [0, 1]$, and prior π on Θ, the Bayes risks satisfy

$$r^*_\pi(\mathcal{M}) \le r^*_\pi(\mathcal{N}) + \varepsilon.$$

(3) There exists a Markov kernel K from $\mathcal{X}$ to $\mathcal{Y}$ such that

$$\mathrm{TV}(KP_\theta, Q_\theta) \le \varepsilon, \qquad \forall \theta \in \Theta,$$

where $(KP_\theta)(y) = \sum_x P_\theta(x)\, K(y \mid x)$.

Proof. **(1)⇒(2).** Fix any finite action set $\mathcal{A}$, any bounded loss $L(\theta, a) \in [0, 1]$, and any prior π on Θ. Let $\hat{a}_{\mathcal{N}}$ be any (randomized) decision rule under $\mathcal{N}$. By ε-deficiency, there exists a decision rule $\hat{a}_{\mathcal{M}}$ under $\mathcal{M}$ such that

$$r(\hat{a}_{\mathcal{M}}; \theta) \le r(\hat{a}_{\mathcal{N}}; \theta) + \varepsilon, \qquad \forall \theta \in \Theta.$$

Averaging w.r.t. π gives

$$\int r(\hat{a}_{\mathcal{M}}; \theta)\, \pi(\,\mathrm{d}\theta) \le \int r(\hat{a}_{\mathcal{N}}; \theta)\, \pi(\,\mathrm{d}\theta) + \varepsilon.$$

Taking the infimum over $\hat{a}_{\mathcal{N}}$ yields $r^*_\pi(\mathcal{M}) \leq r^*_\pi(\mathcal{N}) + \varepsilon$.

(3)⇒(1). Assume there exists a kernel K from $\mathcal{X}$ to $\mathcal{Y}$ such that $\mathrm{TV}(KP_\theta, Q_\theta) \leq \varepsilon$ for all $\theta \in \Theta$. Given any decision rule $\hat{a}_{\mathcal{N}} : \mathcal{Y} \to \mathcal{A}$ under $\mathcal{N}$, define a decision rule under $\mathcal{M}$ by

$$X \sim P_\theta, \qquad Y \sim K(\cdot \mid X), \qquad \hat{a}_{\mathcal{M}} := \hat{a}_{\mathcal{N}}(Y).$$

Then for each θ,

$$\begin{aligned} r(\hat{a}_{\mathcal{M}}; \theta) - r(\hat{a}_{\mathcal{N}}; \theta) &= \mathbb{E}_{Y \sim KP_\theta}\big[L(\theta, \hat{a}_{\mathcal{N}}(Y))\big] - \mathbb{E}_{Y \sim Q_\theta}\big[L(\theta, \hat{a}_{\mathcal{N}}(Y))\big] \\ &\leq \mathrm{TV}(KP_\theta, Q_\theta) \leq \varepsilon, \end{aligned}$$

since $L(\theta, \hat{a}_{\mathcal{N}}(y)) \in [0,1]$. Hence $\mathcal{M}$ is ε-deficient w.r.t. $\mathcal{N}$.

(2)⇒(3). Let the action set be $\mathcal{A} = \mathcal{Y}$ and consider the (non-randomized) decision rule under $\mathcal{N}$ given by

$$\hat{a}_{\mathcal{N}}(y) = y.$$

Any randomized decision rule under $\mathcal{M}$ with action set $\mathcal{Y}$ can be identified with a kernel K from $\mathcal{X}$ to $\mathcal{Y}$. Condition (2) then implies (as in the notes)

$$\sup_{0 \leq L \leq 1} \sup_\pi \inf_K \mathbb{E}_{\theta \sim \pi}\Big[\mathbb{E}_{X \sim P_\theta} \mathbb{E}_{a \sim K(\cdot|X)} L(\theta, a) - \mathbb{E}_{a \sim Q_\theta} L(\theta, a)\Big] \leq \varepsilon, \tag{6.1}$$

where $0 \leq L \leq 1$ means $L(\theta, a) \in [0,1]$ for all (θ, a). The objective is linear in $K(\cdot \mid x)$ and in $\{\pi(\theta) L(\theta, a)\}_{\theta \in \Theta, a \in \mathcal{A}}$, so by a minimax theorem we can swap inf and sup:

$$\inf_K \sup_{0 \leq L \leq 1} \sup_\pi \mathbb{E}_{\theta \sim \pi}\Big[\mathbb{E}_{X \sim P_\theta} \mathbb{E}_{a \sim K(\cdot|X)} L(\theta, a) - \mathbb{E}_{a \sim Q_\theta} L(\theta, a)\Big] \leq \varepsilon. \tag{6.2}$$

For fixed K and L, the expression inside is linear in π, and since Θ is finite,

$$\sup_\pi \mathbb{E}_{\theta \sim \pi}[g(\theta)] = \max_{\theta \in \Theta} g(\theta).$$

Moreover, for each θ,

$$\sup_{0 \leq f \leq 1} \Big(\mathbb{E}_{a \sim KP_\theta}[f(a)] - \mathbb{E}_{a \sim Q_\theta}[f(a)]\Big) = \mathrm{TV}(KP_\theta, Q_\theta).$$

Therefore the inner supremum in (6.2) equals $\max_{\theta \in \Theta} \mathrm{TV}(KP_\theta, Q_\theta)$, so

$$\inf_K \max_{\theta \in \Theta} \mathrm{TV}(KP_\theta, Q_\theta) \leq \varepsilon.$$

Hence there exists a kernel K such that $\mathrm{TV}(KP_\theta, Q_\theta) \leq \varepsilon$ for all $\theta \in \Theta$, proving (3). □

Definition 6.14 (Le Cam's distance)**.** For finite models $\mathcal{M} = (P_\theta)_{\theta \in \Theta}$ and $\mathcal{N} = (Q_\theta)_{\theta \in \Theta}$, define Le Cam's distance as

$$\Delta(\mathcal{M}, \mathcal{N}) = \min\{\varepsilon : \mathcal{M} \text{ is } \varepsilon\text{-deficient to } \mathcal{N},\ \mathcal{N} \text{ is } \varepsilon\text{-deficient to } \mathcal{M}\}.$$

Example 6.15 (Sufficiency)**.** For $\mathcal{M} = (P_\theta)_{\theta \in \Theta}$ and a statistic $T = T(X)$, define the T-induced model $\mathcal{N} = (T_\# P_\theta)_{\theta \in \Theta}$. By the randomization criterion,

$$\begin{aligned} \Delta(\mathcal{M}, \mathcal{N}) = 0 &\iff \mathcal{M} \text{ and } \mathcal{N} \text{ are mutual randomizations} \\ &\iff (\theta \to X \to T) \text{ and } (\theta \to T \to X) \text{ are Markov chains} \quad \iff T \text{ is sufficient for } X. \end{aligned}$$

(Factorization theorem: T is sufficient $\iff p_\theta(x) = g(x) h(\theta, T)$ for some g, h.)

6.10.1 Standard model and a route to asymptotic equivalence

For a sequence of models $(\mathcal{M}_n)_{n\geq 1}$ and $(\mathcal{N}_n)_{n\geq 1}$, how to show asymptotic equivalence $\Delta(\mathcal{M}_n, \mathcal{N}_n) \to 0$ as $n \to \infty$?

Definition 6.16 (Standard model)**.** Let $\mathcal{M} = \{P_1, \ldots, P_m\}$ be a finite model and let

$$\bar{P} \coloneqq \frac{1}{m}\sum_{i=1}^{m} P_i.$$

Then

$$T(x) = \Big(\frac{P_1}{\bar{P}}(x), \ldots, \frac{P_m}{\bar{P}}(x)\Big)$$

is sufficient and lies on

$$\Delta_m \coloneqq \{u \in \mathbb{R}_+^m : \mathbb{1}^T u = m\}.$$

(Applying the factorization theorem to $P_i(x) = \bar{P}(x)T_i(x)$.)

Thus $\mathcal{M}$ is equivalent to the T-induced model $\mathcal{N} = \{\mu_1, \ldots, \mu_m\}$ with

$$\frac{\mu_i(\,\mathrm{d}T)}{\mu(\,\mathrm{d}T)} = T_i,$$

where μ is the distribution of T under $\bar{P}$, known as the *standard distribution.* Indeed,

$$\mathbb{E}_{\mu_i}[f(T)] = \mathbb{E}_{P_i}[f(T(X))] = \mathbb{E}_{\bar{P}}\Big[\frac{P_i}{\bar{P}} f(T(X))\Big] = \mathbb{E}_{\mu}[T_i f(T)].$$

Implication: standard model unifies all statistical models of size m to standard distributions μ on Δ_m.

Theorem 6.17. *If $\mu_n \xrightarrow{\mathrm{d}} \mu$, then $\Delta(\mathcal{M}_n, \mathcal{M}) \to 0$.*

Proof. By (2) in the randomization criterion, it suffices to check that

$$\sup_{\mathcal{A},\pi,L} \big|r_\pi^*(\mathcal{M}_n) - r_\pi^*(\mathcal{M})\big| \to 0.$$

In a standard model,

$$\begin{aligned} r_\pi^*(\mathcal{M}) &= \inf_{\hat{\theta}} \sum_{i=1}^{m} \pi_i\, \mathbb{E}_{\mu_i}\big[L(i, \hat{\theta}(T))\big] \\ &= \inf_{\hat{\theta}} \mathbb{E}_{\mu}\Big[\sum_{i=1}^{m} \pi_i T_i\, L(i, \hat{\theta}(T))\Big]. \end{aligned}$$

Let

$$C \coloneqq \mathrm{conv}\Big(\{(\pi_i L(i,a))_{i=1}^m : a \in \mathcal{A}\}\Big).$$

Then the inner infimum can be written as

$$\inf_{\hat{\theta}} \mathbb{E}_{\mu}\Big[\sum_{i=1}^{m} \pi_i T_i\, L(i, \hat{\theta}(T))\Big] = \mathbb{E}_{\mu}\Big[\inf_{c\in C} \langle c, T\rangle\Big].$$

Since $f(T) = \inf_{c\in C} \langle c, T\rangle$ is bounded by m and is 1-Lipschitz under $\|\cdot\|_1$,

$$\sup_{\mathcal{A},\pi,L} \big|r_\pi^*(\mathcal{M}_n) - r_\pi^*(\mathcal{M})\big| \leq \sup_{\substack{\|f\|_\infty \leq m,\\ |f(x)-f(y)|\leq \|x-y\|_1}} \big|\mathbb{E}_{\mu_n} f - \mathbb{E}_{\mu} f\big| \to 0.$$

(Here one can use that Dudley's metric metrizes weak convergence.) □

Now we're ready to present the main result.

6.10.2 Weak convergence of likelihood ratios ⇒ asymptotic equivalence

Theorem 6.18. *Let $\mathcal{M}_n = \{P_{1,n}, \dots, P_{m,n}\}$ for $n \geq 1$, and $\mathcal{M} = \{P_1, \dots, P_m\}$. Let*

$$L_n = \Big(\frac{P_{2,n}}{P_{1,n}}, \dots, \frac{P_{m,n}}{P_{1,n}}\Big), \qquad L = \Big(\frac{P_2}{P_1}, \dots, \frac{P_m}{P_1}\Big).$$

Suppose $\mathcal{M}$ is homogeneous, *i.e. P_i and P_j are mutually absolutely continuous. If*

$$\mathrm{Law}(L_n \mid P_{1,n}) \xrightarrow{\mathrm{d}} \mathrm{Law}(L \mid P_1),$$

then

$$\Delta(\mathcal{M}_n, \mathcal{M}) \to 0.$$

(In other words, weak convergence of likelihood ratios implies asymptotic equivalence.)

Proof. It suffices to show that the standard distributions $\mu_n \xrightarrow{\mathrm{d}} \mu$. Also note that $\mathrm{Law}(L_n \mid P_{1,n})$ is unchanged when moving to the standard model.

By compactness of $\Delta_m = \{u \in \mathbb{R}_+^m : \mathbb{1}^T u = m\}$ and Prokhorov's theorem, it suffices to show that if $\mu_{n_k} \xrightarrow{\mathrm{d}} \nu$ along some subsequence, then $\nu = \mu$.

For $s = (s_2, \dots, s_m)$ with $s_i > 0$ and $\sum_{i=2}^m s_i < 1$, define

$$f_s(L) = \prod_{i=2}^m L_i^{s_i},$$

which is a continuous function of L. Let $s_1 = 1 - \sum_{i=2}^m s_i \in (0,1)$. By Hölder's inequality,

$$\mathbb{E}_{P_1}\big[f_s(L)^{-1/s_1}\big] = \mathbb{E}_{P_1}\big[L_2^{-s_2/s_1} \cdots L_m^{-s_m/s_1}\big] \leq \prod_{i=2}^m \mathbb{E}_{P_1}[L_i^{-1}]^{s_i/s_1} \leq 1.$$

So the sequence of random variables $f_s(L_n)$ is uniformly integrable. Therefore, by weak convergence,

$$\mathbb{E}_\mu[T_1^{s_1} T_2^{s_2} \cdots T_m^{s_m}] = \mathbb{E}_{P_1}[f_s(L)] = \lim_{n\to\infty} \mathbb{E}_{P_{1,n}}[f_s(L_n)].$$

On the other hand, as $\mu_{n_k} \xrightarrow{\mathrm{d}} \nu$,

$$\mathbb{E}_{P_{1,n_k}}[f_s(L_{n_k})] = \mathbb{E}_{\mu_{n_k}}[T_1^{s_1} \cdots T_m^{s_m}] \to \mathbb{E}_\nu[T_1^{s_1} \cdots T_m^{s_m}].$$

Hence

$$\mathbb{E}_\mu[T_1^{s_1} \cdots T_m^{s_m}] = \mathbb{E}_\nu[T_1^{s_1} \cdots T_m^{s_m}], \qquad \forall s_i > 0, \sum_{i=1}^m s_i = 1.$$

By uniqueness results for moment generating functions, this implies that $\tilde{\mu} = \tilde{\nu}$, where $\tilde{\mu}$ is the restriction of μ to

$$\Delta_m^0 := \{x \in \mathbb{R}^m : x_i > 0,\ \mathbb{1}^T x = m\}, \qquad \text{i.e. } \tilde{\mu}(A) = \mu(A \cap \Delta_m^0).$$

Since $\mathcal{M}$ is homogeneous, $\tilde{\mu} = \mu$ and $\mu(\Delta_m^0) = 1$. Since ν is a probability measure, it follows that $\tilde{\nu} = \nu$. Therefore $\mu = \nu$. □

6.10.3 Local asymptotic normality via likelihood ratios

Finally, we show that if $(P_\theta)_{\theta\in\Theta}$ is QMD, then for any finite set I,

$$\mathcal{M}_n = \{P^{\otimes n}_{\theta_0+\frac{h}{\sqrt{n}}}\}_{h\in I}$$

is asymptotically equivalent to

$$\mathcal{M} = \{\mathbb{N}(h, I(\theta_0)^{-1})\}_{h\in I}.$$

This is called *local asymptotic normality.*

Proof (likelihood ratio expansion). Check the likelihood ratio. In the limiting Gaussian model, for $Z \sim \mathcal{N}(0, I(\theta_0)^{-1})$,

$$\log \frac{\mathcal{N}(h, I(\theta_0)^{-1})}{\mathcal{N}(0, I(\theta_0)^{-1})}(Z) = h^T I(\theta_0) Z - \tfrac{1}{2} h^T I(\theta_0) h,$$

with $I(\theta_0)Z \sim \mathcal{N}(0, I(\theta_0))$.

For the product model, define

$$W_{n,i} = 2\Big(\sqrt{\frac{p_{\theta_0+\frac{h}{\sqrt{n}}}(X_i)}{p_{\theta_0}(X_i)}} - 1\Big).$$

Then

$$\log \frac{p^{\otimes n}_{\theta_0+\frac{h}{\sqrt{n}}}}{p^{\otimes n}_{\theta_0}}(X^n) = 2\sum_{i=1}^n \log\Big(1 + \tfrac{1}{2}W_{n,i}\Big)$$
$$= \sum_{i=1}^n W_{n,i} - \tfrac{1}{4}\sum_{i=1}^n W_{n,i}^2 + \sum_{i=1}^n o(W_{n,i}^2).$$

By QMD,

$$\mathbb{E}_{P_{\theta_0}}\Big[\big(W_{n,i} - \tfrac{1}{\sqrt{n}} h^T s_{\theta_0}(X_i)\big)^2\Big] = o(1/n),$$

thus

$$\mathrm{Var}_{P_{\theta_0}}\Big(\sum_{i=1}^n W_{n,i} - \tfrac{1}{\sqrt{n}}\sum_{i=1}^n h^T s_{\theta_0}(X_i)\Big) \le n\cdot o(1/n) = o(1).$$

Also,

$$\mathbb{E}\sum_{i=1}^n W_{n,i} = -n\int \big(\sqrt{p_{\theta_0+\frac{h}{\sqrt{n}}}} - \sqrt{p_{\theta_0}}\big)^2 \,\mathrm{d}\mu \to -\tfrac{1}{4} h^T \mathbb{E}[s_{\theta_0} s_{\theta_0}^T] h = -\tfrac{1}{4} h^T I(\theta_0) h.$$

Moreover,

$$\sum_{i=1}^n W_{n,i}^2 = \sum_{i=1}^n \Big(\tfrac{1}{\sqrt{n}} h^T s_{\theta_0}(X_i)\Big)^2 + o_p(1)$$
$$= \frac{1}{n}\sum_{i=1}^n h^T s_{\theta_0}(X_i) s_{\theta_0}(X_i)^T h + o_p(1) \xrightarrow{\mathbb{P}} h^T I(\theta_0) h \qquad \text{(by the LLN)}.$$

Therefore,

$$\log \frac{p_{\theta_0+\frac{h}{\sqrt{n}}}^{\otimes n}}{p_{\theta_0}^{\otimes n}}(X^n) = h^T\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\theta_0}(X_i)\Big) - \tfrac{1}{2}h^T I(\theta_0)h + o_p(1),$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n s_{\theta_0}(X_i) \xrightarrow{\mathrm{d}} \mathcal{N}(0, I(\theta_0)) \qquad \text{(by the CLT)}.$$

□

Combining Anderson's lemma and the limiting Gaussian model above, and extending the previous definitions to general models by taking the supremum over all finite submodels, we arrive at the local asymptotic minimax theorem.

Lecture 7: Minimax lower bounds (Le Cam, Fano, Assouad)

7.1 Setup and the minimax risk

We consider a statistical model $\{P_\theta : \theta \in \Theta\}$. We observe $X \sim P_\theta$ and use an estimator $\hat{\theta} = \hat{\theta}(X)$. Let $L(\theta, a) \geq 0$ be a loss function.

The minimax risk is

$$r^\star := \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta\big[L(\theta, \hat{\theta}(X))\big].$$

- **Upper bound:** construct an estimator $\hat{\theta}$ and bound $\sup_\theta \mathbb{E}_\theta[L(\theta, \hat{\theta}(X))]$.
- **Lower bound:** show that no estimator can beat a certain rate.

In the previous lecture (LAN), we focused on asymptotic analysis and exact constants. In this lecture we focus on non-asymptotic minimax lower bounds, usually aiming for the optimal *rate*.

A high-level idea: for any prior π on Θ, the minimax risk dominates the Bayes risk,

$$r^\star \geq r_\pi^\star, \qquad r_\pi^\star := \inf_{\hat{\theta}} \mathbb{E}_{\theta \sim \pi} \mathbb{E}_\theta\big[L(\theta, \hat{\theta}(X))\big].$$

Finding a least favorable prior can be hard. Instead we use simple priors:

1. **Binary prior:** $\pi = \mathrm{Unif}\{\theta_0, \theta_1\}$ (Le Cam's two-point method).
2. **Multiple hypotheses:** $\pi = \mathrm{Unif}\{\theta_1, \dots, \theta_m\}$ (Fano, Assouad).

7.2 Le Cam's two-point method

Definition 7.1 (Total variation)**.** For two distributions P, Q on the same measurable space,

$$\mathrm{TV}(P, Q) := \sup_A |P(A) - Q(A)|.$$

If P, Q have densities p, q w.r.t. a common dominating measure,

$$\mathrm{TV}(P, Q) = \frac{1}{2} \int |p - q|.$$

Theorem 7.2 (Le Cam's two-point lower bound)**.** *Let* $\theta_0, \theta_1 \in \Theta$ *and suppose the* separation condition

$$\inf_a \big(L(\theta_0, a) + L(\theta_1, a)\big) \geq \Delta.$$

Then

$$r^\star \geq \inf_{\hat\theta} \frac{1}{2}\Big(\mathbb{E}_{\theta_0} L(\theta_0, \hat\theta(X)) + \mathbb{E}_{\theta_1} L(\theta_1, \hat\theta(X))\Big) \geq \frac{\Delta}{2}\big(1 - \mathrm{TV}(P_{\theta_0}, P_{\theta_1})\big).$$

Proof. Fix any estimator $\hat\theta = \hat\theta(X)$. Let p_0, p_1 be densities of $P_{\theta_0}, P_{\theta_1}$. Then

$$\begin{aligned}
\mathbb{E}_{\theta_0} L(\theta_0, \hat\theta(X)) + \mathbb{E}_{\theta_1} L(\theta_1, \hat\theta(X)) &= \int L(\theta_0, \hat\theta(x)) p_0(x)\, dx + \int L(\theta_1, \hat\theta(x)) p_1(x)\, dx \\
&\geq \int \inf_a \big(L(\theta_0, a) + L(\theta_1, a)\big) \min\{p_0(x), p_1(x)\}\, dx \\
&\geq \Delta \int \min\{p_0(x), p_1(x)\}\, dx.
\end{aligned}$$

Next,

$$\min\{p_0, p_1\} = \frac{1}{2}\big(p_0 + p_1 - |p_0 - p_1|\big),$$

so

$$\begin{aligned}
\int \min\{p_0, p_1\} &= \frac{1}{2}\int (p_0 + p_1) - \frac{1}{2}\int |p_0 - p_1| \\
&= 1 - \mathrm{TV}(P_{\theta_0}, P_{\theta_1}).
\end{aligned}$$

Combining and dividing by 2 gives the claim. □

7.2.1 A useful template

To lower bound $r^\star$ via Le Cam, pick θ_0, θ_1 such that

- **Separation:** $\inf_a(L(\theta_0, a) + L(\theta_1, a)) \geq \Delta$.
- **Indistinguishability:** $\mathrm{TV}(P_{\theta_0}, P_{\theta_1}) \leq 1 - \Omega(1)$.

Often the indistinguishability condition is shown via stronger (more tractable) bounds such as

1. $\mathrm{H}^2(P_{\theta_0}, P_{\theta_1}) \leq 2 - \Omega(1)$,
2. $D_{\mathrm{KL}}(P_{\theta_0} || P_{\theta_1}) = O(1)$ or $D_{\mathrm{KL}}(P_{\theta_1} || P_{\theta_0}) = O(1)$,
3. $\chi^2(P_{\theta_0} || P_{\theta_1}) = O(1)$ or $\chi^2(P_{\theta_1} || P_{\theta_0}) = O(1)$,
4. $I(\Theta; X) \leq \log 2 - \Omega(1)$ for $\Theta \sim \mathrm{Unif}\{\theta_0, \theta_1\}$ (exercise).

7.2.2 Example 1.1: normal mean estimation (one-dimensional)

Let $X \sim \mathcal{N}(\theta, \sigma^2)$ with unknown $\theta \in \mathbb{R}$ and known σ^2. Take squared loss $L(\theta, \hat\theta) = (\hat\theta - \theta)^2$. The minimax risk is

$$r^\star = \inf_{\hat\theta} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\big[(\hat\theta(X) - \theta)^2\big].$$

Upper bound. Choosing $\hat{\theta}(X) = X$ gives

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(X-\theta)^2] = \sigma^2,$$

so $r^\star \le \sigma^2$.

Lower bound (two points). Pick $\theta_0 = 0$ and $\theta_1 = \delta$. For squared loss,

$$\inf_a \big((a-\theta_0)^2 + (a-\theta_1)^2\big) = \frac{(\theta_1-\theta_0)^2}{2} = \frac{\delta^2}{2},$$

so $\Delta = \delta^2/2$.

For Gaussians with equal variance,

$$1 - \mathrm{TV}\big(\mathcal{N}(0,\sigma^2), \mathcal{N}(\delta,\sigma^2)\big) = 2\Big(1 - \Phi\big(|\delta|/(2\sigma)\big)\Big),$$

where Φ is the standard normal CDF. Therefore Le Cam yields

$$r^\star \ge \sup_{\delta \in \mathbb{R}} \frac{\Delta}{2}\big(1 - \mathrm{TV}\big) = \sup_{\delta \in \mathbb{R}} \frac{\delta^2}{4} \cdot 2\Big(1 - \Phi\big(|\delta|/(2\sigma)\big)\Big) = \sup_{\delta \in \mathbb{R}} \frac{\delta^2}{2}\Big(1 - \Phi\big(|\delta|/(2\sigma)\big)\Big) \approx 0.332\,\sigma^2.$$

(Compare with the exact value $r^\star = \sigma^2$ from Anderson's lemma in Lecture 6.)

7.2.3 Example 1.2: binomial model

Let $X \sim \mathrm{Bin}(n,\theta)$ with unknown $\theta \in [0,1]$. Target:

$$r^\star = \inf_{\hat{\theta}} \sup_{\theta \in [0,1]} \mathbb{E}_\theta\big[(\hat{\theta}(X)-\theta)^2\big].$$

Upper bound. Choose $\hat{\theta}(X) = X/n$. Then

$$\mathbb{E}_\theta\Big[\Big(\frac{X}{n} - \theta\Big)^2\Big] = \frac{\theta(1-\theta)}{n} \le \frac{1}{4n} = O(1/n).$$

Lower bound (two points). Apply the two-point method with

$$\theta_0 = \frac{1}{2}, \qquad \theta_1 = \frac{1}{2} + \frac{1}{2\sqrt{n}}.$$

For squared loss the separation parameter is

$$\Delta = \frac{1}{2}(\theta_1 - \theta_0)^2 = \frac{1}{2}\Big(\frac{1}{2\sqrt{n}}\Big)^2 = \Omega(1/n).$$

For indistinguishability, compute

$$\begin{aligned} D_{\mathrm{KL}}\big(\mathrm{Bin}(n,\theta_0)\,||\,\mathrm{Bin}(n,\theta_1)\big) &= n\,D_{\mathrm{KL}}\big(\mathrm{Bern}(\theta_0)\,||\,\mathrm{Bern}(\theta_1)\big) \\ &= \frac{n}{2}\Big((1+\tfrac{1}{\sqrt{n}})\log(1+\tfrac{1}{\sqrt{n}}) + (1-\tfrac{1}{\sqrt{n}})\log(1-\tfrac{1}{\sqrt{n}})\Big) \\ &= \frac{n}{2} \cdot O(1/n) = O(1). \end{aligned}$$

Thus Le Cam implies $r^\star = \Omega(1/n)$, matching the $O(1/n)$ upper bound.

7.2.4 Example 1.3: functional estimation (entropy estimation)

Let $X = (X_1, \dots, X_n)$ be i.i.d. draws from an unknown pmf $P = (p_1, \dots, p_k)$ on $[k]$. Consider the loss

$$L(P, a) = |a - H(P)|,$$

where $H(P) = -\sum_{i=1}^{k} p_i \log p_i$ is the entropy.

Known sharp result. (Jiao et al. 2015; Wu and Yang 2016)

$$r^\star := \inf_{\hat{\theta}} \sup_{P} \mathbb{E}_P\big[|\hat{\theta} - H(P)|\big] \asymp \frac{k}{n \log n} + \frac{\log k}{\sqrt{n}} \qquad \text{(in particular when } k \lesssim n \log n).$$

A simpler $\Omega((\log k)/\sqrt{n})$ lower bound via two points. Since

$$D_{\mathrm{KL}}(P^{\otimes n} || Q^{\otimes n}) = n\, D_{\mathrm{KL}}(P || Q),$$

the two-point method motivates the optimization problem

$$\max\ |H(P_0) - H(P_1)| \quad \text{s.t.} \quad D_{\mathrm{KL}}(P_0 || P_1) \le \frac{c}{n}.$$

Try

$$P_0 = \Big(\frac{1}{2}, \underbrace{\frac{1}{2(k-1)}, \dots, \frac{1}{2(k-1)}}_{k-1 \text{ times}}\Big),$$

$$P_1 = \Big(\frac{1-\varepsilon}{2}, \underbrace{\frac{1+\varepsilon}{2(k-1)}, \dots, \frac{1+\varepsilon}{2(k-1)}}_{k-1 \text{ times}}\Big), \qquad \varepsilon \in (0, 1/2).$$

KL computation. Only two types of coordinates appear, so

$$\begin{aligned} D_{\mathrm{KL}}(P_0 || P_1) &= \frac{1}{2} \log\Big(\frac{1/2}{(1-\varepsilon)/2}\Big) + (k-1) \cdot \frac{1}{2(k-1)} \log\Big(\frac{1/(2(k-1))}{(1+\varepsilon)/(2(k-1))}\Big) \\ &= \frac{1}{2} \log\Big(\frac{1}{1-\varepsilon}\Big) + \frac{1}{2} \log\Big(\frac{1}{1+\varepsilon}\Big) \\ &= \frac{1}{2} \log\Big(\frac{1}{1-\varepsilon^2}\Big) = O(\varepsilon^2). \end{aligned}$$

Thus $D_{\mathrm{KL}}(P_0 || P_1) = O(1/n)$ if $\varepsilon = O(1/\sqrt{n})$.

Entropy difference. Compute

$$\begin{aligned} H(P_0) &= -\frac{1}{2} \log\Big(\frac{1}{2}\Big) - (k-1) \cdot \frac{1}{2(k-1)} \log\Big(\frac{1}{2(k-1)}\Big) \\ &= \frac{1}{2} \log 2 + \frac{1}{2} \log(2(k-1)). \end{aligned}$$

Also

$$\begin{aligned} H(P_1) &= -\frac{1-\varepsilon}{2} \log\Big(\frac{1-\varepsilon}{2}\Big) - (k-1) \cdot \frac{1+\varepsilon}{2(k-1)} \log\Big(\frac{1+\varepsilon}{2(k-1)}\Big) \\ &= \frac{1-\varepsilon}{2} \log\Big(\frac{2}{1-\varepsilon}\Big) + \frac{1+\varepsilon}{2} \log\Big(\frac{2(k-1)}{1+\varepsilon}\Big). \end{aligned}$$

Therefore

$$\begin{aligned}|H(P_0) - H(P_1)| &= \Big|\frac{1}{2}\log 2 + \frac{1}{2}\log(2(k-1)) - \frac{1-\varepsilon}{2}\log\Big(\frac{2}{1-\varepsilon}\Big) - \frac{1+\varepsilon}{2}\log\Big(\frac{2(k-1)}{1+\varepsilon}\Big)\Big| \\ &\gtrsim \varepsilon \log k.\end{aligned}$$

Choosing $\varepsilon \asymp 1/\sqrt{n}$ yields

$$r^\star = \Omega\Big(\frac{\log k}{\sqrt{n}}\Big).$$

(The other term $\Omega\big(k/(n\log n)\big)$ requires a more involved two-point construction; this is the topic of Lecture 8.)

7.2.5 Example 1.4: two-armed bandit (Gaussian rewards)

Let $\theta = (\mu_1, \mu_2) \in [0,1]^2$. For $t \in [T]$, the learner pulls an arm $\pi_t \in \{1,2\}$ based on past history (π^{t-1}, r^{t-1}), and observes reward

$$r_t \sim \mathcal{N}(\mu_{\pi_t}, 1).$$

The (expected) regret is

$$R_T(\pi) := T\max\{\mu_1,\mu_2\} - \sum_{t=1}^{T}\mu_{\pi_t}.$$

Let $\Delta := |\mu_1 - \mu_2|$ be the gap. We will show

$$r_T^\star := \inf_\pi \sup_{\mu_1,\mu_2 \colon |\mu_1-\mu_2|\ge\Delta} \mathbb{E}_{\mu_1,\mu_2}[R_T(\pi)] = \Omega\Big(\Big(\frac{1\vee\log(T\Delta^2)}{\Delta}\Big) \wedge T\Delta\Big).$$

In particular, choosing $\Delta \asymp 1/\sqrt{T}$ gives the usual lower bound $\Omega(\sqrt{T})$.

Proof. First, by the chain rule for KL divergence (exercise),

$$\begin{aligned}D_{\mathrm{KL}}(P_{\mu_1,\mu_2}||P_{\mu_1',\mu_2'}) &= \sum_{t=1}^{T}\mathbb{E}_{P_{\mu_1,\mu_2}}\Big[\frac{(\mu_1-\mu_1')^2}{2}\,\mathbb{1}\{\pi_t=1\} + \frac{(\mu_2-\mu_2')^2}{2}\,\mathbb{1}\{\pi_t=2\}\Big] \\ &= \frac{(\mu_1-\mu_1')^2}{2}\,\mathbb{E}[T_1] + \frac{(\mu_2-\mu_2')^2}{2}\,\mathbb{E}[T_2],\end{aligned}$$

where $T_i := \sum_{t=1}^{T}\mathbb{1}\{\pi_t = i\}$ is the total number of pulls of arm i.

Motivated by this, choose two points

$$(\mu_1,\mu_2) = (\Delta, 0), \qquad (\mu_1',\mu_2') = (\Delta, 2\Delta).$$

The separation parameter for regret is $T\Delta$ (the gap is Δ in either model). Moreover,

$$D_{\mathrm{KL}}(P_{\mu_1,\mu_2}||P_{\mu_1',\mu_2'}) = 2\Delta^2\,\mathbb{E}_1[T_2],$$

where $\mathbb{E}_1 := \mathbb{E}_{P_{\Delta,0}}$. Le Cam's two-point bound then gives

$$r_T^\star = \Omega\Big(T\Delta\,\exp\big(-2\Delta^2\mathbb{E}_1[T_2]\big)\Big),$$

using the inequality $(1 - \mathrm{TV}(P,Q)) \ge \frac{1}{2}\exp(-D_{\mathrm{KL}}(P||Q))$.

Note that $\mathbb{E}_1[T_2]$ depends on the policy π, and the above is useful only if $\mathbb{E}_1[T_2]$ is small. A different lower bound comes from evaluating the regret directly under $(\Delta, 0)$:

$$r_T^\star \ge \mathbb{E}_1[R_T(\pi)] = \Delta\, \mathbb{E}_1[T_2].$$

Combining,

$$\begin{aligned} r_T^\star &= \Omega\Big(\max\big\{\Delta\mathbb{E}_1[T_2],\ T\Delta \exp(-2\Delta^2\mathbb{E}_1[T_2])\big\}\Big) \\ &= \Omega\Big(\min_{t\in[0,T]} \max\big\{\Delta t,\ T\Delta \exp(-2\Delta^2 t)\big\}\Big) \\ &= \Omega\Big(\Big(\frac{1 \vee \log(T\Delta^2)}{\Delta}\Big) \wedge T\Delta\Big). \end{aligned}$$

□

7.2.6 Example 1.5: multi-armed bandit

Same observation model, but with K arms. Let $\theta = (\mu_1, \dots, \mu_K) \in [0,1]^K$ and

$$R_T(\pi) = T \max_{i\in[K]} \mu_i - \sum_{t=1}^T \mu_{\pi_t}.$$

We will show

$$r^\star := \inf_\pi \sup_\theta \mathbb{E}_\theta[R_T(\pi)] = \Omega(\sqrt{KT}).$$

(Interestingly, two points suffice for this example!)

Proof. Choose

$$\theta_1 = (\Delta, 0, 0, \dots, 0),$$

and for each $i = 2, \dots, K$ let

$$\theta_{2,i} = (\Delta, 0, \dots, 0, \underbrace{2\Delta}_{i\text{-th coordinate}}, 0, \dots, 0).$$

For each pair $(\theta_1, \theta_{2,i})$, the separation parameter for regret is always $T\Delta$. Moreover, for any policy π,

$$D_{\mathrm{KL}}(P_{\theta_1} || P_{\theta_{2,i}}) = 2\Delta^2\, \mathbb{E}_1[T_i], \qquad T_i = \sum_{t=1}^T \mathbb{1}\{\pi_t = i\}.$$

Key observation: since $\sum_{i=2}^K \mathbb{E}_1[T_i] \le T$, there must exist some i_0 such that $\mathbb{E}_1[T_{i_0}] \le T/(K-1)$. Applying the two-point argument to $(\theta_1, \theta_{2,i_0})$ and choosing $\Delta \asymp \sqrt{K/T}$ makes $D_{\mathrm{KL}}(P_{\theta_1} || P_{\theta_{2,i_0}}) = O(1)$. Therefore

$$r^\star = \Omega(T\Delta) = \Omega(\sqrt{KT}).$$

□

7.2.7 Why two points may fail in high dimensions

Consider normal mean estimation in n dimensions:

$$X \sim \mathcal{N}(\theta, \sigma^2 I_n), \qquad L(\theta, \hat{\theta}) = \left\| \hat{\theta} - \theta \right\|_2^2 .$$

The two-point method gives at best

$$r^\star \geq \sup_{\theta_0, \theta_1} \frac{\|\theta_0 - \theta_1\|_2^2}{2} \Big(1 - \Phi\Big(\frac{\|\theta_0 - \theta_1\|_2}{2\sigma} \Big) \Big) \lesssim \sigma^2 .$$

This does not capture the dependence on dimension n. (Recall that $r^\star = n\sigma^2$ by Anderson's lemma.)

At a high level: testing between two hypotheses does *not* capture the true difficulty of high-dimensional problems.

7.3 Testing multiple hypotheses

7.3.1 Challenges in high dimensions

- **Separation:** we may want different separation structures than a single Δ for a pair of points.
- **Indistinguishability:** in binary testing, $1 - \mathrm{TV}(P, Q)$ tightly controls the optimal testing error. In multiple-hypotheses problems, the analogous tight quantity is often not tractable, and we need further lower bounds.

7.3.2 Pairwise separation: Fano's inequality

Theorem 7.3 (Fano-type lower bound). *Let $\theta_1, \dots, \theta_m \in \Theta$ satisfy the pairwise separation condition*

$$\min_{i \neq j} \inf_a \big(L(\theta_i, a) + L(\theta_j, a) \big) \geq \Delta .$$

Let $\pi = \mathrm{Unif}\{\theta_1, \dots, \theta_m\}$, and let $\Theta \sim \pi$, $X \mid \Theta \sim P_\Theta$. Then

$$r_\pi^\star \geq \frac{\Delta}{2} \Big(1 - \frac{I(\Theta; X) + \log 2}{\log m} \Big) .$$

Before proving it, we establish a useful "golden formula" for mutual information.

Lemma 7.4 (Golden formula for mutual information). *For any pair (X, Y),*

$$I(X;Y) = \min_{Q_Y} D_{\mathrm{KL}}(P_{XY} || P_X Q_Y) = \min_{Q_Y} \mathbb{E}_{P_X} \big[D_{\mathrm{KL}}(P_{Y|X} || Q_Y) \big] .$$

Proof. Simply note that for any Q_Y,

$$I(X;Y) = D_{\mathrm{KL}}(P_{XY} || P_X Q_Y) - D_{\mathrm{KL}}(P_Y || Q_Y) .$$

Taking $Q_Y = P_Y$ shows the minimum equals $I(X;Y)$. □

Proof of Fano. Fix any estimator $\hat{\theta} = \hat{\theta}(X)$. Consider the indicator map

$$(\Theta, X) \mapsto Z := \mathbb{1}\big\{ L(\Theta, \hat{\theta}(X)) < \Delta/2 \big\} .$$

Under $P_{\Theta X}$, $Z \sim \mathrm{Bern}(p)$ with $p := \mathbb{P}\big(L(\Theta, \hat{\theta}(X)) < \Delta/2 \big)$. Under $P_\Theta P_X$ (independent Θ and X), define $q := \mathbb{P}_{\Theta \sim \pi, X \sim P_X}(Z = 1)$.

Step 1: bound q by the separation condition. Fix x and let $a = \hat{\theta}(x)$. Because for any $i \neq j$ we have $\inf_a(L(\theta_i, a) + L(\theta_j, a)) \geq \Delta$, there cannot be two distinct indices $i \neq j$ such that both $L(\theta_i, a) < \Delta/2$ and $L(\theta_j, a) < \Delta/2$. Hence for any fixed action a, at most one hypothesis can satisfy $L(\theta_i, a) < \Delta/2$. Since $\Theta \sim \mathrm{Unif}\{\theta_1, \dots, \theta_m\}$ is independent of X under $P_\Theta P_X$, this implies

$$q \leq \frac{1}{m}.$$

Step 2: data processing. By the data processing inequality,

$$D_{\mathrm{KL}}\big(\mathrm{Bern}(p) \,||\, \mathrm{Bern}(q)\big) \leq I(\Theta; X).$$

Since $q \leq 1/m$, one can bound the Bernoulli KL to obtain

$$p \leq \frac{I(\Theta; X) + \log 2}{\log m},$$

equivalently

$$\mathbb{P}\big(L(\Theta, \hat{\theta}(X)) \geq \Delta/2\big) \geq 1 - \frac{I(\Theta; X) + \log 2}{\log m}.$$

Step 3: Markov/thresholding. Finally,

$$\mathbb{E}[L(\Theta, \hat{\theta}(X))] \geq \frac{\Delta}{2}\, \mathbb{P}\big(L(\Theta, \hat{\theta}(X)) \geq \Delta/2\big),$$

so the claimed lower bound follows after taking the infimum over $\hat{\theta}$. □

7.3.3 Generalized Fano

The previous argument yields a more general statement.

Theorem 7.5 (Generalized Fano). *For any prior π on Θ and any $\Delta > 0$,*

$$r_\pi^\star \geq \Delta\Big(1 - \frac{I(\Theta; X) + \log 2}{\log(1/P_0)}\Big), \qquad \Theta \sim \pi,\ X \mid \Theta \sim P_\Theta,$$

where

$$P_0 := \sup_a \pi\big(L(\Theta, a) < \Delta\big)$$

is the small-ball probability.

The classical Fano inequality is a special case with $\pi = \mathrm{Unif}\{\theta_1, \dots, \theta_m\}$ and a pairwise separation condition.

7.3.4 Additive separation: Assouad's lemma

Theorem 7.6 (Assouad). *For a hypercube parameterization $u \in \{\pm 1\}^d$, associate $\theta_u \in \Theta$. Suppose*

$$\inf_a \big(L(\theta_u, a) + L(\theta_{u'}, a)\big) \geq \Delta \cdot \sum_{i=1}^{d} \mathbb{1}\{u_i \neq u'_i\}.$$

Let $\pi = \mathrm{Unif}\{\theta_u : u \in \{\pm 1\}^d\}$. *Then*

$$r^\star_\pi \geq \frac{\Delta}{4} \sum_{i=1}^{d} \big(1 - \mathrm{TV}(P_{i,+}, P_{i,-})\big),$$

where

$$P_{i,+} := \frac{1}{2^{d-1}} \sum_{u:u_i=+1} P_{\theta_u}, \qquad P_{i,-} := \frac{1}{2^{d-1}} \sum_{u:u_i=-1} P_{\theta_u}.$$

The following corollaries are often used.

Corollary 7.7 (Classical Assouad)**.** *Let u, u' be* neighbors *if they differ in exactly one coordinate. Then*

$$r^\star_\pi \geq \frac{d\Delta}{4}\Big(1 - \max_{u,u' \ \textit{neighbors}} \mathrm{TV}(P_{\theta_u}, P_{\theta_{u'}})\Big).$$

Corollary 7.8 (Averaged-neighbor version)**.** *Let $u \oplus i$ denote u with the i-th bit flipped. Then*

$$r^\star_\pi \geq \frac{d\Delta}{4}\Big(1 - \mathbb{E}_{u\sim \mathrm{Unif}(\{\pm1\}^d)}\mathbb{E}_{i\sim\mathrm{Unif}([d])}\mathrm{TV}(P_{\theta_u}, P_{\theta_{u\oplus i}})\Big).$$

Proof of Assouad. Fix any estimator $\hat\theta$. Construct an estimate $\hat u = (\hat u_1, \dots, \hat u_d) \in \{\pm1\}^d$ by

$$\hat u := \arg\min_{u \in \{\pm1\}^d} L(\theta_u, \hat\theta).$$

Then for any u,

$$\begin{aligned} L(\theta_u, \hat\theta) &\geq \frac{L(\theta_u,\hat\theta) + L(\theta_{\hat u}, \hat\theta)}{2} \\ &\geq \frac{\Delta}{2}\sum_{i=1}^{d} \mathbb{1}\{u_i \neq \hat u_i\}. \end{aligned}$$

Averaging over u uniformly,

$$\begin{aligned} \frac{1}{2^d}\sum_u \mathbb{E}_{\theta_u}[L(\theta_u,\hat\theta)] &\geq \frac{\Delta}{2}\sum_{i=1}^{d}\frac{1}{2^d}\sum_u \mathbb{P}_{\theta_u}(\hat u_i \neq u_i) \\ &= \frac{\Delta}{4}\sum_{i=1}^{d}\Big(P_{i,+}(\hat u_i \neq +1) + P_{i,-}(\hat u_i \neq -1)\Big) \\ &\geq \frac{\Delta}{4}\sum_{i=1}^{d}\big(1 - \mathrm{TV}(P_{i,+}, P_{i,-})\big), \end{aligned}$$

where the last step applies Le Cam's two-point bound to testing $P_{i,+}$ vs $P_{i,-}$. Taking the infimum over $\hat\theta$ yields the theorem. □

Remark 7.9. (Exercise.) Show that

$$\frac{1}{d}\sum_{i=1}^{d}\mathrm{TV}(P_{i,+}, P_{i,-}) = 1 - \Omega(1) \quad\Longleftrightarrow\quad \frac{1}{d}\sum_{i=1}^{d} I(U_i; X) = \log 2 - \Omega(1)$$

for $U \sim \mathrm{Unif}(\{\pm1\}^d)$. Also note that $I(U;X) \geq \sum_{i=1}^d I(U_i;X)$, so under a hypercube construction Assouad is no weaker than Fano.

7.4 High-dimensional examples

7.4.1 Example 2.1: normal mean model (high dimensions)

Let $X \sim \mathcal{N}(\theta, \sigma^2 I_n)$ and loss $L(\theta, \hat{\theta}) = \left\|\hat{\theta} - \theta\right\|_2^2$. We show

$$r^\star = \Omega(n\sigma^2).$$

Proof 1 (Fano via packing). Construct a subset $\Theta_0 \subset \{\pm\delta\}^n$ (with δ to be chosen) such that

- $m := |\Theta_0|$ is large enough,
- $\min_{\theta \neq \theta' \in \Theta_0} \|\theta - \theta'\|_2^2 \geq \delta^2 n/5$.

By the Gilbert–Varshamov bound below, we can choose $m = \exp(\Omega(n))$. Then for squared loss,

$$\min_{\theta \neq \theta' \in \Theta_0} \inf_a \left(\|\theta - a\|_2^2 + \left\|\theta' - a\right\|_2^2 \right) = \frac{1}{2} \min_{\theta \neq \theta' \in \Theta_0} \left\|\theta - \theta'\right\|_2^2 \geq \frac{\delta^2 n}{10} =: \Delta.$$

Using the golden formula for mutual information,

$$\begin{aligned} I(\Theta; X) &\leq \max_{\theta \in \Theta_0} D_{\mathrm{KL}}\big(\mathcal{N}(\theta, \sigma^2 I_n) \,||\, \mathcal{N}(0, \sigma^2 I_n)\big) \\ &= \max_{\theta \in \Theta_0} \frac{\|\theta\|_2^2}{2\sigma^2} = \frac{n\delta^2}{2\sigma^2}. \end{aligned}$$

Fano then gives

$$r^\star = \Omega\Big(\delta^2 n \Big(1 - \frac{n\delta^2/(2\sigma^2) + \log 2}{\Omega(n)}\Big)\Big).$$

Choosing $\delta \asymp \sigma$ yields $r^\star = \Omega(n\sigma^2)$.

7.4.2 Gilbert–Varshamov bound

Lemma 7.10 (Gilbert–Varshamov). *There exists a set $A \subset \{\pm 1\}^n$ such that*

$$\min_{u \neq u' \in A} \sum_{i=1}^n \mathbb{1}\{u_i \neq u_i'\} \geq d,$$

and

$$m := |A| \geq \frac{2^n}{\sum_{j=0}^{d-1} \binom{n}{j}} = 2^{n(1 - h_2(d/n)) + o(n)},$$

where $h_2(x) = x \log_2 \frac{1}{x} + (1 - x) \log_2 \frac{1}{1-x}$ is the binary entropy.

Proof. (Volume argument.) Fix $u \in \{\pm 1\}^n$. The number of $u' \in \{\pm 1\}^n$ within Hamming distance $d - 1$ of u equals

$$\Big|\Big\{u' \in \{\pm 1\}^n : \sum_{i=1}^n \mathbb{1}\{u_i \neq u_i'\} \leq d - 1\Big\}\Big| = \sum_{j=0}^{d-1} \binom{n}{j}.$$

So if we have selected fewer than $2^n / \sum_{j=0}^{d-1} \binom{n}{j}$ points, there must exist a new point at Hamming distance at least d from all selected ones. Greedily adding such points constructs a set A of the claimed size. □

Proof 2 (generalized Fano on the full hypercube). Let $\Theta \sim \mathrm{Unif}(\{\pm\delta\}^n)$. Then $I(\Theta; X) \leq n\delta^2/(2\sigma^2)$. Pick

$$\Delta = \frac{n\delta^2}{12}.$$

The small-ball probability satisfies

$$P_0 = \sup_a \pi(\|\Theta - a\|_2^2 < \Delta) \leq \sup_{\hat{\theta} \in \{\pm\delta\}^n} \pi(\left\|\Theta - \hat{\theta}\right\|_2^2 < 4\Delta),$$

where $\hat{\theta}$ can be taken as a nearest hypercube point to a. Now, $\left\|\Theta - \hat{\theta}\right\|_2^2 = 4\delta^2 \, d_H(\Theta, \hat{\theta})$, so the event $\left\|\Theta - \hat{\theta}\right\|_2^2 < 4\Delta = n\delta^2/3$ implies $d_H(\Theta, \hat{\theta}) < n/12$. Thus

$$P_0 \leq 2^{-n} \sum_{j=0}^{\lfloor n/12 \rfloor} \binom{n}{j} = 2^{-\Omega(n)} \qquad \text{(by Stirling)}.$$

Applying generalized Fano again gives $r^\star = \Omega(n\sigma^2)$ for $\delta \asymp \sigma$.

Proof 3 (Assouad). For $\delta > 0$, let $\theta_u = \delta u$ for $u \in \{\pm 1\}^n$. For neighbors u, u', we have $\|\theta_u - \theta_{u'}\|_2^2 = 4\delta^2$, hence

$$\inf_a \left(\|\theta_u - a\|_2^2 + \|\theta_{u'} - a\|_2^2 \right) = \frac{1}{2} \|\theta_u - \theta_{u'}\|_2^2 = 2\delta^2.$$

Also,

$$1 - \max_{u,u' \text{ neighbors}} \mathrm{TV}\big(\mathcal{N}(\theta_u, \sigma^2 I_n), \mathcal{N}(\theta_{u'}, \sigma^2 I_n)\big) = 2\big(1 - \Phi(\delta/\sigma)\big).$$

Choosing $\delta = \sigma$ and applying Assouad yields $r^\star = \Omega(n\sigma^2)$.

7.4.3 Example 2.2: learning theory (VC lower bounds)

Let $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{XY}$ i.i.d. with $Y \in \{0, 1\}$. Let $\mathcal{F}$ be a class of functions $X \to \{0, 1\}$ with VC dimension d. For a trained classifier $\hat{f}$ based on the sample, define the excess risk

$$\mathrm{ER}(\hat{f}) := P_{XY}(Y \neq \hat{f}(X)) - \min_{f \in \mathcal{F}} P_{XY}(Y \neq f(X)).$$

We will show that for $n \geq d$,

$$\inf_{\hat{f}} \sup_{P_{XY}} \mathbb{E}[\mathrm{ER}(\hat{f})] = \Omega\Big(\sqrt{\frac{d}{n}}\Big) \qquad \text{(agnostic setting)},$$

and

$$\inf_{\hat{f}} \sup_{P_{XY} : \exists f \in \mathcal{F},\, Y = f(X) \ P_{XY}\text{-a.s.}} \mathbb{E}[\mathrm{ER}(\hat{f})] = \Omega\Big(\frac{d}{n}\Big) \qquad \text{(realizable setting)}.$$

Recall (VC shattering). $\mathrm{VCdim}(\mathcal{F}) = d$ implies that there exist $x_1, \ldots, x_d \in \mathcal{X}$ such that for every $u \in \{\pm 1\}^d$ there exists $f_u \in \mathcal{F}$ with $f_u(x_i) = u_i$ for all $i \in [d]$. (Here we encode labels as ±1 for convenience.)

Agnostic case (Assouad). Fix $x_1, \dots, x_d$ and the functions $\{f_u\}_{u \in \{\pm 1\}^d}$. For each $u \in \{\pm 1\}^d$, construct $P_u = P_{XY,u}$ as follows:

- $X \sim \text{Unif}\{x_1, \dots, x_d\}$.
- $Y \mid X = x_i$ equals u_i with probability $\frac{1}{2} + \delta$ and equals $-u_i$ with probability $\frac{1}{2} - \delta$.

Separation. For all u,

$$\min_{f \in \mathcal{F}} P_u(f(X) \neq Y) = \frac{1}{2} - \delta.$$

For any f, writing $\text{ER}(f, P_u) = P_u(Y \neq f(X)) - (\frac{1}{2} - \delta)$, one can check that for all u, u',

$$\begin{aligned} \text{ER}(f, P_u) + \text{ER}(f, P_{u'}) &= P_u(Y \neq f(X)) + P_{u'}(Y \neq f(X)) - 2(\tfrac{1}{2} - \delta) \\ &\geq \sum_{i=1}^{d} \frac{1}{d}\Big(\mathbb{1}\{u_i \neq u_i'\} \cdot 1 + \mathbb{1}\{u_i = u_i'\} \cdot (1 - 2\delta)\Big) - (1 - 2\delta) \\ &= \frac{2\delta}{d} \sum_{i=1}^{d} \mathbb{1}\{u_i \neq u_i'\}. \end{aligned}$$

Thus Assouad holds with $\Delta = 2\delta/d$.

Indistinguishability. For neighbors u, u',

$$D_{\text{KL}}(P_u^{\otimes n} || P_{u'}^{\otimes n}) = n D_{\text{KL}}(P_u || P_{u'}) = n \cdot \frac{1}{d} D_{\text{KL}}\big(\text{Bern}(\tfrac{1}{2} + \delta) \,||\, \text{Bern}(\tfrac{1}{2} - \delta)\big) = O\Big(\frac{n\delta^2}{d}\Big).$$

Choosing $\delta \asymp \sqrt{d/n}$ makes this $O(1)$. Assouad then yields

$$\inf_{\hat{f}} \sup_{P_{XY}} \mathbb{E}[\text{ER}(\hat{f})] = \Omega(d\Delta) = \Omega\Big(\sqrt{\frac{d}{n}}\Big).$$

Realizable case (Assouad). Now define a different family $\{P_u\}$ where the Bayes error within $\mathcal{F}$ is zero. Let $u \in \{\pm 1\}^{d-1}$ (we vary labels only on $x_2, \dots, x_d$). Define P_u by

- $X = x_1$ w.p. $1 - (d-1)/n$, and $X = x_i$ w.p. $1/n$ for each $2 \leq i \leq d$.
- $Y \mid X = x_i$ equals the prescribed label (deterministic): $Y = u_i$ for $2 \leq i \leq d$ (and fix $Y = +1$ on x_1).

Clearly $\min_{f \in \mathcal{F}} P_u(f(X) \neq Y) = 0$ for all u. A similar analysis gives a separation parameter $\Delta \asymp 1/n$. For neighbors $u' = u \oplus i$ (flipping the label at some x_i, $i \geq 2$), we have

$$\text{TV}(P_u^{\otimes n}, P_{u'}^{\otimes n}) \leq \mathbb{P}(x_i \text{ appears in } X_1, \dots, X_n) = 1 - (1 - 1/n)^n = 1 - \Omega(1).$$

Therefore Assouad yields

$$\inf_{\hat{f}} \sup_{\text{realizable } P_{XY}} \mathbb{E}[\text{ER}(\hat{f})] = \Omega((d-1)\Delta) = \Omega\Big(\frac{d}{n}\Big).$$

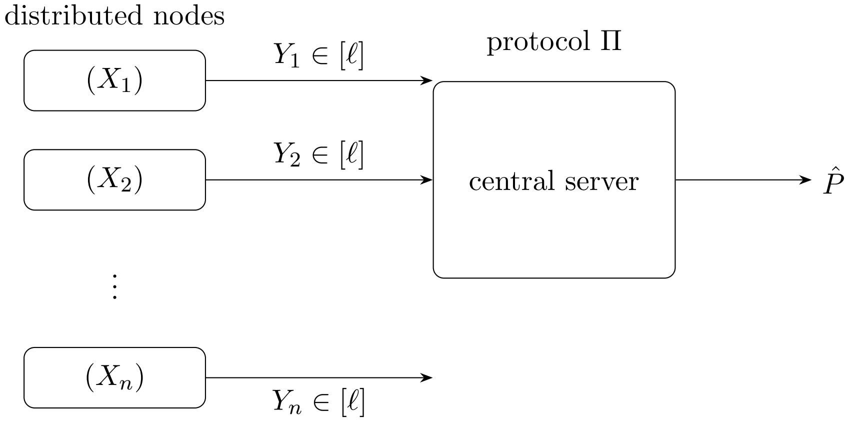
(See homework for further generalizations.)

7.5 Assouad in sequential settings: a communication lower bound

Assouad's lemma is also surprisingly flexible in sequential settings.

7.5.1 Example 2.3: distribution estimation under sequential communication protocols

Let $P = (p_1, \dots, p_k)$ be an unknown pmf on $[k]$. We observe $X_1, \dots, X_n \overset{\text{iid}}{\sim} P$. However, the estimator $\hat{P}$ must be formed via a sequential distributed protocol: node t sees X_t and sends a message $Y_t \in [\ell]$ to a central server, where Y_t can depend on (X_t, Y^{t-1}) (a protocol Π to be designed). Assume a communication constraint $\ell \le k$.



We will show

$$r^\star := \inf_{(\hat{P},\Pi)} \sup_P \mathbb{E}_P\big[\mathrm{TV}(P,\hat{P})\big] = \Omega\Big(\frac{k}{\sqrt{n\ell}}\Big) \qquad \text{(e.g. when } k \le \ell \text{ and } n \ge k^2/\ell).$$

Proof. Without loss of generality assume k is even. For $u \in \{\pm 1\}^{k/2}$, construct

$$P_u = \Big(\frac{1+\delta u_1}{k}, \frac{1-\delta u_1}{k}, \dots, \frac{1+\delta u_{k/2}}{k}, \frac{1-\delta u_{k/2}}{k}\Big), \qquad \delta \in (0, 1/2) \text{ (to be chosen)}.$$

It is easy to check that for the loss $L(P,\hat{P}) = \mathrm{TV}(P,\hat{P})$, this hypercube construction has separation parameter $\Delta = \Omega(\delta/k)$.

We use Corollary 2 of Assouad and upper bound the averaged-neighbor total variation distance between the message distributions. Let $u \oplus i$ denote u with the i-th bit flipped. Write $P_{Y^n|u}$ for the law of Y^n under P_u and protocol Π. Then

$$\begin{aligned}
\mathbb{E}_u\mathbb{E}_i \mathrm{TV}(P_{Y^n|u}, P_{Y^n|u\oplus i}) &\le \sqrt{\mathbb{E}_u\mathbb{E}_i\, \mathrm{TV}(P_{Y^n|u}, P_{Y^n|u\oplus i})^2} \qquad \text{(Jensen)}\\
&\le \sqrt{\mathbb{E}_u\mathbb{E}_i\, \mathrm{H}^2(P_{Y^n|u}, P_{Y^n|u\oplus i})} \qquad (\mathrm{TV} \le \mathrm{H})\\
&\le C\sqrt{\sum_{t=1}^n \mathbb{E}_u\mathbb{E}_i\mathbb{E}_{P_{Y^{t-1}|u}}\Big[\mathrm{H}^2\big(P_{Y_t|Y^{t-1},u}, P_{Y_t|Y^{t-1},u\oplus i}\big)\Big]},
\end{aligned}$$

where the last step uses Jayram's subadditivity of H^2 (Lecture 3).

Next, we upper bound the conditional Hellinger distance. Since $\mathrm{H}^2(P,Q) \le \chi^2(P||Q)$,

$$\mathbb{E}_i\Big[\mathrm{H}^2\big(P_{Y_t|Y^{t-1},u}, P_{Y_t|Y^{t-1},u\oplus i}\big)\Big] \le \mathbb{E}_i\Big[\chi^2\big(P_{Y_t|Y^{t-1},u\oplus i} \,||\, P_{Y_t|Y^{t-1},u}\big)\Big].$$

Write

$$P_{Y_t|Y^{t-1},u} = \sum_{x\in[k]} P_{Y_t|Y^{t-1},X_t=x}\, P_u(x).$$

Since $\delta \le 1/2$, we have $P_u(x) \ge (1-\delta)/k \ge 1/(2k)$ for all x, so

$$P_{Y_t|Y^{t-1},u}(y) \ge \frac{1}{2k}\sum_{x\in[k]} P_{Y_t|Y^{t-1},X_t=x}(y).$$

Therefore

$$\chi^2\big(P_{Y_t|Y^{t-1},u\oplus i}||P_{Y_t|Y^{t-1},u}\big) \le \sum_{y\in[\ell]} \frac{\big(P_{Y_t=y|Y^{t-1},u\oplus i} - P_{Y_t=y|Y^{t-1},u}\big)^2}{\frac{1}{2k}\sum_{x\in[k]} P_{Y_t=y|Y^{t-1},X_t=x}}.$$

Now note that $u\oplus i$ differs from u only on the pair $(2i-1, 2i)$, and the change in the pmf is $\pm 2\delta/k$ on those two coordinates. Thus

$$\begin{aligned} &P_{Y_t=y|Y^{t-1},u\oplus i} - P_{Y_t=y|Y^{t-1},u} \\ &\quad = \frac{2\delta}{k}\Big(P_{Y_t=y|Y^{t-1},X_t=2i-1} - P_{Y_t=y|Y^{t-1},X_t=2i}\Big). \end{aligned}$$

Plugging this into the previous display,

$$\begin{aligned} \chi^2\big(P_{Y_t|Y^{t-1},u\oplus i}||P_{Y_t|Y^{t-1},u}\big) &\le 2k\Big(\frac{2\delta}{k}\Big)^2 \sum_{y\in[\ell]} \frac{\big(P_{Y_t=y|Y^{t-1},X_t=2i-1} - P_{Y_t=y|Y^{t-1},X_t=2i}\big)^2}{\sum_{x\in[k]} P_{Y_t=y|Y^{t-1},X_t=x}} \\ &\le \frac{8\delta^2}{k}\sum_{y\in[\ell]} \frac{P_{Y_t=y|Y^{t-1},X_t=2i-1} + P_{Y_t=y|Y^{t-1},X_t=2i}}{\sum_{x\in[k]} P_{Y_t=y|Y^{t-1},X_t=x}}, \end{aligned}$$

where we used $(a-b)^2 \le a+b$. Averaging over $i \sim \mathrm{Unif}([k/2])$,

$$\begin{aligned} \mathbb{E}_i\chi^2\big(P_{Y_t|Y^{t-1},u\oplus i}||P_{Y_t|Y^{t-1},u}\big) &\le \frac{8\delta^2}{k}\sum_{y\in[\ell]} \frac{\mathbb{E}_i\big[P_{Y_t=y|Y^{t-1},X_t=2i-1} + P_{Y_t=y|Y^{t-1},X_t=2i}\big]}{\sum_{x\in[k]} P_{Y_t=y|Y^{t-1},X_t=x}} \\ &= \frac{8\delta^2}{k}\sum_{y\in[\ell]} \frac{\frac{2}{k}\sum_{x\in[k]} P_{Y_t=y|Y^{t-1},X_t=x}}{\sum_{x\in[k]} P_{Y_t=y|Y^{t-1},X_t=x}} \\ &= \frac{16\delta^2}{k^2}\sum_{y\in[\ell]} 1 = O\Big(\frac{\delta^2\ell}{k^2}\Big). \end{aligned}$$

Hence

$$\mathbb{E}_i\Big[\mathrm{H}^2\big(P_{Y_t|Y^{t-1},u}, P_{Y_t|Y^{t-1},u\oplus i}\big)\Big] \le O\Big(\frac{\delta^2\ell}{k^2}\Big).$$

Putting everything together,

$$\mathbb{E}_u\mathbb{E}_i\mathrm{TV}(P_{Y^n|u}, P_{Y^n|u\oplus i}) \le O\Big(\sqrt{\frac{n\delta^2\ell}{k^2}}\Big).$$

Assouad's lemma (Corollary 2) then yields

$$r^\star = \Omega\Big(\delta\Big(1 - O\Big(\sqrt{\frac{n\delta^2\ell}{k^2}}\Big)\Big)\Big).$$

Choosing $\delta = k/\sqrt{n\ell}$ (and ensuring $\delta < 1/2$) gives

$$r^\star = \Omega\Big(\frac{k}{\sqrt{n\ell}}\Big).$$

□

7.6 Special topic: interactive Le Cam and the DEC

7.6.1 Model for interactive decision making

We consider an interactive/sequential setting with

- an unknown true model $M^\star$ in a given model class $\mathcal{M}$ (e.g. the reward distributions of all arms),
- at each round $t = 1, \ldots, T$:
  1. learner chooses an action $a_t \in \mathcal{A}$;
  2. nature reveals reward $r_t \in [0, 1]$ and possibly an additional observation o_t, with

$$\mathbb{E}[r_t \mid a_t = a] = r^{M^\star}(a), \qquad (r_t, o_t) \sim M^\star(a_t).$$

- learner aims to minimize the regret

$$R_T := \sum_{t=1}^{T} \big(r_\star^{M^\star} - r^{M^\star}(a_t)\big), \qquad r_\star^M := \max_{a\in\mathcal{A}} r^M(a).$$

Example (multi-armed bandit). Take $\mathcal{A} = [K]$ and

$$M^\mu(a) = \mathrm{Bern}(\mu_a), \qquad r^{M^\mu}(a) = \mu_a, \qquad \mathcal{M} = \{M^\mu : \mu \in [0,1]^K\}.$$

Question. What is a general two-point lower bound for R_T?

Idea. Let

$$g^M(a) := r_\star^M - r^M(a)$$

denote the gap of action a under model M. A naive two-point lower bound might suggest

$$\inf_{\{a_t\}} \sup_{M^\star\in\mathcal{M}} \mathbb{E}[R_T] \gtrsim T \cdot \sup_{M_0, M_1 \in \mathcal{M}} \Big\{ \inf_{a\in\mathcal{A}} \big(g^{M_0}(a) + g^{M_1}(a)\big) : \ \mathrm{H}^2(M_0, M_1) \le c/T \Big\},$$

for small $c > 0$. However, several issues arise.

7.6.2 Challenges

1. The metric $\inf_a(g^{M_0}(a) + g^{M_1}(a))$ can be too pessimistic. If a policy uses an action distribution p under M_0, then $\mathbb{E}_{a\sim p}[g^{M_1}(a)]$ may be a better separation metric.

2. $\mathrm{H}^2(M_0, M_1)$ is not well-defined in the interactive setting, since the distribution of (r_t, o_t) depends on the chosen a_t. One should instead consider an averaged quantity such as $\mathbb{E}_{a\sim p}[\mathrm{H}^2(M_0(a), M_1(a))]$.

3. Where should we take the infimum over p (learner as the min player)?

   - $\sup_{M_0,M_1} \inf_p$ can be too small (same reason as item 1).
   - $\inf_p \sup_{M_0,M_1}$ can be too large, since the learner can adapt p sequentially.

7.6.3 Definition: constrained decision-to-estimation coefficient (DEC)

Definition 7.11 (DEC)**.** The constrained decision-to-estimation coefficient is defined as

$$\mathrm{dec}_\varepsilon(\mathcal{M}) := \sup_{\bar{M}} \inf_{p\in\Delta(\mathcal{A})} \sup_{M\in\mathcal{M}\cup\{\bar{M}\}} \Big\{\mathbb{E}_{a\sim p}[g^M(a)] : \mathbb{E}_{a\sim p}[\mathrm{H}^2(M(a), \bar{M}(a))] \le \varepsilon^2\Big\}.$$

- This is a sup–inf–sup structure: first choose a reference model $\bar{M}$, the learner chooses an action distribution p based on $\bar{M}$, then nature chooses an alternative model M.

- The separation condition is with respect to the average under p.

- The reference model $\bar{M}$ does not need to belong to $\mathcal{M}$.

7.6.4 Examples

Example 3.1 (two-armed bandit). For two-armed Bernoulli bandit $(\mathrm{Bern}(\mu_1), \mathrm{Bern}(\mu_2))$ with $|\mu_1 - \mu_2| \ge \Delta$, choose the reference

$$\bar{M} = (\mathrm{Bern}(\tfrac{1}{2} + \Delta), \mathrm{Bern}(\tfrac{1}{2})),$$

and consider alternatives

$$M \in \Big\{(\mathrm{Bern}(\tfrac{1}{2} + \Delta), \mathrm{Bern}(\tfrac{1}{2})),\ (\mathrm{Bern}(\tfrac{1}{2} + \Delta), \mathrm{Bern}(\tfrac{1}{2} + \Delta + \varepsilon))\Big\}.$$

A calculation gives

$$\mathrm{dec}_\varepsilon(\mathcal{M}) \ge \inf_{p_2\in[0,1]} \max\Big\{p_2\Delta,\ (1-p_2)\Big(\frac{\varepsilon}{p_2} - \Delta\Big) + \varepsilon\Big\} = \Omega\Big(\Delta \wedge \frac{\varepsilon^2}{\Delta}\Big).$$

Example 3.2 (multi-armed bandit). For $\mathcal{M} = \{\prod_{i=1}^K \mathrm{Bern}(\mu_i) : \mu_i \in [0,1]\}$, we may choose $\bar{M} = \prod_{i=1}^K \mathrm{Bern}(1/2)$. For any distribution p on $[K]$, pick $i_0 = \arg\min_i p_i$ and set $M(i_0) = \mathrm{Bern}(1/2 + \varepsilon\sqrt{K})$. This gives

$$\mathrm{dec}_\varepsilon(\mathcal{M}) = \Omega(\varepsilon\sqrt{K}), \qquad \text{when } \varepsilon\sqrt{K} = O(1).$$

7.6.5 DEC lower bound for regret

Theorem 7.12 (DEC lower bound). *There exist absolute constants $c, C > 0$ such that*

$$\inf_{\{a_t\}} \sup_{M^\star \in \mathcal{M}} \mathbb{E}_{M^\star}[R_T] = \Omega\Big(T\left(\mathrm{dec}_\varepsilon(\mathcal{M}) - C\varepsilon\right)_+\Big), \qquad \varepsilon = \sqrt{\frac{c}{T}}.$$

Specializing to the previous examples gives a lower bound $\Omega(1/\Delta)$ for Example 3.1 when $\Delta \gtrsim 1/\sqrt{T}$, and $\Omega(\sqrt{KT})$ for $T \geq K$.

7.6.6 Proof sketch (simpler case $\bar{M} \in \mathcal{M}$)

(Foster, Golowich, Han 2023). Let $\Delta := \mathrm{dec}_\varepsilon(\mathcal{M})$.

Let $p_t(\cdot \mid H^{t-1})$ denote the learner's action distribution at time t. Define the learner's average play under $\bar{M}$ by

$$p_{\bar{M}} := \mathbb{E}_{\bar{M}}\Big[\frac{1}{T}\sum_{t=1}^{T} p_t(\cdot \mid H^{t-1})\Big].$$

Let M be an inner maximizer under $p = p_{\bar{M}}$, and define the learner's average play under M by

$$p_M := \mathbb{E}_M\Big[\frac{1}{T}\sum_{t=1}^{T} p_t(\cdot \mid H^{t-1})\Big].$$

By definition of dec_ε, we have

$$\mathbb{E}_{a\sim p_{\bar{M}}}[g^M(a)] \geq \Delta, \qquad (1)$$
$$\mathbb{E}_{a\sim p_{\bar{M}}}[\mathrm{H}^2(M(a), \bar{M}(a))] \leq \varepsilon^2. \qquad (2)$$

By way of contradiction, assume that

$$\mathbb{E}_{a\sim p_M}[g^M(a)] \leq \Delta/100, \qquad (3)$$
$$\mathbb{E}_{a\sim p_{\bar{M}}}[g^{\bar{M}}(a)] \leq \Delta/100. \qquad (4)$$

We introduce two lemmas.

Lemma 7.13 (Lemma 1). *For $c > 0$ small enough (hence $\varepsilon = \sqrt{c/T}$ small enough),*

$$\mathrm{TV}(p_M, p_{\bar{M}}) \leq 0.1.$$

Proof. Let $P^M_{r^T,o^T}$ and $P^{\bar{M}}_{r^T,o^T}$ denote the law of the full interaction sequence. By data processing,

$$\mathrm{TV}(p_M, p_{\bar{M}})^2 \leq \mathrm{TV}(P^M_{r^T,o^T}, P^{\bar{M}}_{r^T,o^T})^2 \leq \mathrm{H}^2(P^M_{r^T,o^T}, P^{\bar{M}}_{r^T,o^T}).$$

Using subadditivity of H^2 in sequential models,

$$\begin{aligned}\mathrm{H}^2(P^M_{r^T,o^T}, P^{\bar{M}}_{r^T,o^T}) &\leq C\sum_{t=1}^{T}\mathbb{E}_{\bar{M}}\Big[\mathrm{H}^2\big(P^M_{r_t,o_t|H^{t-1}}, P^{\bar{M}}_{r_t,o_t|H^{t-1}}\big)\Big]\\ &= C\sum_{t=1}^{T}\mathbb{E}_{\bar{M}}\big[\mathrm{H}^2(M(a_t), \bar{M}(a_t))\big]\\ &= CT\,\mathbb{E}_{a\sim p_{\bar{M}}}[\mathrm{H}^2(M(a), \bar{M}(a))] \leq CT\varepsilon^2 \leq 0.1,\end{aligned}$$

where the last step used (2) and the choice $\varepsilon = \sqrt{c/T}$. □

Lemma 7.14 (Lemma 2).

$$\mathbb{E}_{a\sim p_{\bar{M}}}\left|r^{M}(a)-r^{\bar{M}}(a)\right|\le\varepsilon.$$

(This step critically uses that the rewards are observed.)

Proof. As $r_t\in[0,1]$, we have

$$\left|r^{M}(a)-r^{\bar{M}}(a)\right|\le\mathrm{TV}(M(a),\bar{M}(a))\le\mathrm{H}(M(a),\bar{M}(a)).$$

Taking expectation over $a\sim p_{\bar{M}}$ and using Jensen,

$$\mathbb{E}_{a\sim p_{\bar{M}}}\left|r^{M}(a)-r^{\bar{M}}(a)\right|\le\mathbb{E}_{a\sim p_{\bar{M}}}\mathrm{H}(M(a),\bar{M}(a))\le\sqrt{\mathbb{E}_{a\sim p_{\bar{M}}}\mathrm{H}^2(M(a),\bar{M}(a))}\le\varepsilon,$$

where the last step used (2). □

Contradiction argument. Let

$$A^{M}:=\{a:g^{M}(a)\le\Delta/10\}.$$

1. By Lemma 2 and (1),

   $$\begin{aligned}\Delta&\le r_\star^{M}-\mathbb{E}_{a\sim p_{\bar{M}}}[r^{M}(a)]\\&\le r_\star^{M}-\mathbb{E}_{a\sim p_{\bar{M}}}[r^{\bar{M}}(a)]+\varepsilon\\&=r_\star^{M}-r_\star^{\bar{M}}+\mathbb{E}_{a\sim p_{\bar{M}}}[g^{\bar{M}}(a)]+\varepsilon\\&\le r_\star^{M}-r_\star^{\bar{M}}+\Delta/100+\varepsilon\qquad\text{(by (4)).}\end{aligned}$$

   Hence

   $$r_\star^{M}-r_\star^{\bar{M}}\ge 99\Delta/100-\varepsilon.$$

2. By (3) and Markov's inequality,

   $$p_M(A^M)=\mathbb{P}_{a\sim p_M}(g^{M}(a)\le\Delta/10)\ge 9/10.$$

3. By item 2 and Lemma 1,

   $$p_{\bar{M}}(A^M)\ge 4/5.$$

4. By item 1 and item 3,

   $$\begin{aligned}\mathbb{E}_{a\sim p_{\bar{M}}}\left[(r^{M}(a)-r^{\bar{M}}(a))\mathbb{1}\{a\in A^M\}\right]&\ge(r_\star^{M}-\Delta/10-r_\star^{\bar{M}})\,p_{\bar{M}}(A^M)\\&\ge(89\Delta/100-\varepsilon)\cdot\frac{4}{5}.\end{aligned}$$

   However, Lemma 2 states that the left-hand side is at most ε. This is a contradiction when $\Delta>C\varepsilon$.

Therefore, at least one of (3) or (4) must fail, implying a regret lower bound of order $T(\Delta-C\varepsilon)_+$ (up to constants).

General $\bar{M}$ (Glasgow and Rakhlin 2023). For $\bar{M} \notin \mathcal{M}$, (4) is no longer a consequence of small regret. A stopping-time argument fixes this. Let ALG be the original learner's algorithm, and define ALG' as follows: $\mathsf{ALG}'_t = \mathsf{ALG}_t$ as long as

$$\sum_{s<t} g^{\bar{M}}(a_s) < \frac{\Delta T}{100},$$

and ALG' always pulls

$$a^\star = \arg\max_a r^{\bar{M}}(a)$$

otherwise. Now redefine $p_{\bar{M}}, M, p_M$ using ALG'. Then (1), (2), (4) and Lemmas 1–2 still hold.

Let $\tau > 0$ be the stopping time of

$$\sum_{t=1}^{\tau} g^{\bar{M}}(a_t) \geq \frac{\Delta T}{100}.$$

By Lemma 2 and Markov's inequality, with probability at least 0.9 under $P_{\bar{M}}^{\mathsf{ALG}'}$,

$$\frac{1}{T}\sum_{t=1}^{T} \left| r^M(a_t) - r^{\bar{M}}(a_t) \right| \leq 10\varepsilon.$$

On this event,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^M(a_t) &= \frac{1}{T}\sum_{t=1}^{T\wedge\tau} \left(r_\star^M - r^M(a_t)\right) \\
&\geq \frac{1}{T}\sum_{t=1}^{T\wedge\tau} \left(r_\star^M - r^{\bar{M}}(a_t)\right) - \frac{1}{T}\sum_{t=1}^{T} \left| r^M(a_t) - r^{\bar{M}}(a_t) \right| \\
&\geq \frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{\bar{M}}(a_t) + \frac{T\wedge\tau}{T}\left(r_\star^M - r_\star^{\bar{M}}\right) - 10\varepsilon \\
&\geq \frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{\bar{M}}(a_t) + \frac{T\wedge\tau}{T}\left(0.99\Delta - \varepsilon\right) - 10\varepsilon. \qquad (5)
\end{aligned}$$

Here the last step used item 1 above, i.e. $r_\star^M - r_\star^{\bar{M}} \geq 99\Delta/100 - \varepsilon$.

- If $\tau > T$, then by (5),

  $$\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^M(a_t) \geq 0.99\Delta - 11\varepsilon = \Omega(\Delta) \qquad \text{for } \Delta > C\varepsilon.$$

- If $\tau < T$, then by definition of τ,

  $$\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^{\bar{M}}(a_t) = \frac{1}{T}\sum_{t=1}^{\tau} g^{\bar{M}}(a_t) \geq \frac{\Delta}{100},$$

  so (5) yields

  $$\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^M(a_t) \geq \frac{\Delta}{100} - 10\varepsilon = \Omega(\Delta) \qquad \text{for } \Delta > C\varepsilon.$$

Therefore,

$$P_{\bar{M}}^{\mathsf{ALG}'}\Big(\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^M(a_t) = \Omega(\Delta)\Big) \geq 0.9$$

in both cases. Since $\mathrm{TV}(P_{\bar{M}}^{\mathsf{ALG}'}, P_{M}^{\mathsf{ALG}'}) \leq 0.1$ by Lemma 1, we get

$$P_{M}^{\mathsf{ALG}'}\Big(\frac{1}{T}\sum_{t=1}^{T\wedge\tau} g^M(a_t) = \Omega(\Delta)\Big) \geq 0.8.$$

Finally, since ALG' and ALG coincide up to time $T \wedge \tau$, this gives the claimed result.

Remark 7.15. In Glasgow and Rakhlin (2023), this stopping-time argument establishes a stronger high-probability statement: for any fixed $c_0 > 0$,

$$\inf_{\{a_t\}} \sup_{M^\star} \mathbb{P}_{M^\star}\Big(\frac{R(T)}{T} > \big((1-c_0)\,\mathrm{dec}_\varepsilon(\nu) - C\varepsilon\big)_+\Big) = \Omega(1), \qquad \varepsilon = \sqrt{\frac{c}{T}},$$

where c, C depend on c_0.

Lecture 8: Advanced Le Cam's Method

8.1 General hypothesis testing

We observe $X \sim P_\theta$, where $\theta \in \Theta$. We aim to test

$$H_0 : \ \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \ \theta \in \Theta_1.$$

Remark 8.1. *Simple* hypothesis: one of Θ_0, Θ_1 is a singleton. *Composite* hypothesis: Θ_0 and/or Θ_1 is a set.

In the composite setting, for a test $T : \mathcal{X} \to \{0, 1\}$,

$$\text{Type I error} = \sup_{\theta \in \Theta_0} P_\theta(T = 1), \qquad \text{Type II error} = \sup_{\theta \in \Theta_1} P_\theta(T = 0).$$

Theorem 8.2 (Composite testing via least favorable priors)**.**

$$\inf_T \Big(\sup_{\theta \in \Theta_0} P_\theta(T = 1) + \sup_{\theta \in \Theta_1} P_\theta(T = 0) \Big) = 1 - \inf_{\substack{\pi_0 \in \mathcal{P}(\Theta_0) \\ \pi_1 \in \mathcal{P}(\Theta_1)}} \mathrm{TV}\Big(\mathbb{E}_{\theta \sim \pi_0}[P_\theta], \ \mathbb{E}_{\theta \sim \pi_1}[P_\theta] \Big).$$

Remark 8.3. Last lecture, the basic Le Cam two-point method reduces estimation problems to hypothesis testing between two *simple* hypotheses. However, it can be helpful to let one or both hypotheses be *mixture distributions* (i.e. $\mathbb{E}_{\theta \sim \pi}[P_\theta]$) with a carefully chosen prior π.

8.2 Advanced Le Cam I: point vs. mixture

Theorem 8.4 (Point vs. mixture)**.** *Let $\theta_0 \in \Theta$ and $\Theta_1 \subset \Theta$. Assume there exists $\Delta > 0$ such that*

$$\inf_{\theta \in \Theta_1} \inf_a \big(L(\theta_0, a) + L(\theta, a) \big) \ \geq \ \Delta.$$

Then for any probability distribution π on Θ_1,

$$\inf_{\widehat{\theta}} \sup_{\theta \in \{\theta_0\} \cup \Theta_1} \mathbb{E}_\theta \big[L(\theta, \widehat{\theta}(X)) \big] \ \geq \ \frac{\Delta}{2} \Big(1 - \mathrm{TV}\big(P_{\theta_0}, \ \mathbb{E}_{\theta \sim \pi}[P_\theta] \big) \Big).$$

Proof. Consider the two-point prior $\frac{1}{2}(\delta_{\theta_0} + \pi)$. The Bayes risk lower bound follows from the same two-point argument as in basic Le Cam, with $P = P_{\theta_0}$ and $Q = \mathbb{E}_{\theta \sim \pi}[P_\theta]$. □

8.2.1 How to upper bound $\mathrm{TV}\big(P_{\theta_0}, \mathbb{E}_{\theta\sim\pi}[P_\theta]\big)$?

The "point vs. mixture" structure is only helpful when

$$\mathrm{TV}\big(P_{\theta_0}, \mathbb{E}_{\theta\sim\pi}[P_\theta]\big) \ll \inf_{\theta\in\Theta_1} \mathrm{TV}(P_{\theta_0}, P_\theta),$$

i.e. the mixture increases closeness.

A standard method is the Ingster–Suslina χ^2 method (a.k.a. the second-moment method), by upper bounding $\chi^2\big(\mathbb{E}_{\theta\sim\pi}[P_\theta] \,\|\, P_{\theta_0}\big)$.

Theorem 8.5 (χ^2-method)**.** *Assume P_θ has density p_θ w.r.t. a common dominating measure. Then*

$$\chi^2\big(\mathbb{E}_{\theta\sim\pi}[P_\theta] \;\|\; P_{\theta_0}\big) = \mathbb{E}_{\theta,\theta'\sim\pi}\Big[\int \frac{p_\theta\, p_{\theta'}}{p_{\theta_0}}\Big] - 1,$$

where $\theta' \sim \pi$ is an independent copy of θ.

Proof.

$$\begin{aligned}\chi^2\big(\mathbb{E}_{\theta\sim\pi}[P_\theta] \;\|\; P_{\theta_0}\big) + 1 &= \int \frac{\big(\mathbb{E}_{\theta\sim\pi}[p_\theta]\big)^2}{p_{\theta_0}} = \int \frac{\mathbb{E}_{\theta,\theta'\sim\pi}[p_\theta p_{\theta'}]}{p_{\theta_0}} \\ &= \mathbb{E}_{\theta,\theta'\sim\pi}\Big[\int \frac{p_\theta p_{\theta'}}{p_{\theta_0}}\Big],\end{aligned}$$

by Fubini. □

Corollary 8.6 (i.i.d. models)**.** *For i.i.d. observations,*

$$\chi^2\big(\mathbb{E}_{\theta\sim\pi}[P_\theta^{\otimes n}] \;\|\; P_{\theta_0}^{\otimes n}\big) = \mathbb{E}_{\theta,\theta'\sim\pi}\Big[\Big(\int \frac{p_\theta p_{\theta'}}{p_{\theta_0}}\Big)^n\Big] - 1.$$

Proof. Just check

$$\int \frac{p_\theta^{\otimes n} p_{\theta'}^{\otimes n}}{p_{\theta_0}^{\otimes n}} = \Big(\int \frac{p_\theta p_{\theta'}}{p_{\theta_0}}\Big)^n.$$

□

8.2.2 Example 1.1: Planted clique

Given an undirected graph G on n vertices, aim to test between

$$H_0:\ G \sim \mathcal{G}(n, \tfrac{1}{2}) \qquad \text{vs.} \qquad H_1:\ G \sim \mathcal{G}(n, \tfrac{1}{2}, k),$$

where under H_1 there exists an unknown $S \subseteq [n]$, $|S| = k$, such that

$$\mathbb{P}((i,j) \in E) = \begin{cases} 1, & i, j \in S, \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

Target. Find a constant C such that if

$$k < 2\log_2 n - 2\log_2\log_2 n + C,$$

then no test can reliably distinguish between H_0 and H_1.

Remark 8.7. Why is the mixture structure in H_1 important? Because for each *fixed* instance of H_1, the learner knows the set S and can look at whether $G[S]$ is a clique.

Proof. Let P be the law of $G \sim \mathcal{G}(n, \frac{1}{2})$. Let P_S be the law of G with a clique planted at S, and let S be uniform over $\binom{[n]}{k}$. Then

$$\int \frac{P_S P_{S'}}{P} = \sum_G \frac{P_S(G)\, P_{S'}(G)}{P(G)} = \sum_{(x_{ij})\in\{0,1\}^{\binom{n}{2}}} \frac{\mathbb{1}\{x_{ij}=1\ \forall i,j\in S\}\,\mathbb{1}\{x_{ij}=1\ \forall i,j\in S'\}\,(\frac{1}{2})^{2\binom{n}{2}-2\binom{k}{2}}}{(\frac{1}{2})^{\binom{n}{2}}}$$
$$= 2^{\binom{|S\cap S'|}{2}}.$$

Therefore

$$\begin{aligned}\chi^2\big(\mathbb{E}[P_S]\,\|\,P\big) &= \mathbb{E}_{S,S'}\Big[2^{\binom{|S\cap S'|}{2}}\Big] - 1\\ &= \sum_{r=0}^{k} 2^{\binom{r}{2}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} - 1 = o(1) \qquad \text{when } k < 2\log_2 n - 2\log_2\log_2 n + C,\end{aligned}$$

by algebra. □

8.2.3 Example 1.2: Uniformity testing

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P = (p_1, \ldots, p_k)$, aim to test

$$H_0:\ P = \mathrm{Unif}[k] \qquad \text{vs.} \qquad H_1:\ \mathrm{TV}(P, \mathrm{Unif}[k]) \geq \varepsilon.$$

Target. The sample complexity of a reliable uniformity test is

$$n = \Theta\Big(\frac{\sqrt{k}}{\varepsilon^2}\Big).$$

Remark 8.8. A naive two-point method does not succeed: if the learner knew the pattern of how P deviates from uniform, then $O(\varepsilon^{-2})$ samples would suffice.

Proof of the lower bound. WLOG assume k is even. Under H_0,

$$P = (1/k, \ldots, 1/k).$$

Under H_1, let

$$P_v = \Big(\frac{1-2\varepsilon v_1}{k}, \frac{1+2\varepsilon v_1}{k}, \ldots, \frac{1-2\varepsilon v_{k/2}}{k}, \frac{1+2\varepsilon v_{k/2}}{k}\Big), \qquad v = (v_1, \ldots, v_{k/2}) \sim \mathrm{Unif}(\{\pm 1\}^{k/2}).$$

Note that $\mathrm{TV}(P_v, \mathrm{Unif}[k]) = \varepsilon$ for all $v \in \{\pm 1\}^{k/2}$. Moreover,

$$\begin{aligned}\int \frac{P_v P_{v'}}{P} &= \sum_{x=1}^{k} \frac{P_v(x) P_{v'}(x)}{P(x)} = \sum_{i=1}^{k/2} \Big(\frac{(1-2\varepsilon v_i)(1-2\varepsilon v_i')}{k} + \frac{(1+2\varepsilon v_i)(1+2\varepsilon v_i')}{k}\Big)\\ &= 1 + \frac{8\varepsilon^2}{k} \sum_{i=1}^{k/2} v_i v_i'.\end{aligned}$$

Hence

$$\begin{aligned}\chi^2\big(\mathbb{E}[P_v^{\otimes n}] \,\|\, P^{\otimes n}\big) &= \mathbb{E}_{v,v'}\Big[\Big(1+\frac{8\varepsilon^2}{k}\sum_{i=1}^{k/2} v_i v_i'\Big)^n\Big]-1\\ &\le \mathbb{E}_{v,v'}\exp\Big(\frac{8n\varepsilon^2}{k}\sum_{i=1}^{k/2} v_i v_i'\Big)-1\\ &\le \exp\Big(\frac{1}{2}\Big(\frac{8n\varepsilon^2}{k}\Big)^2\cdot\frac{k}{2}\Big)-1=\exp\Big(\frac{16n^2\varepsilon^4}{k}\Big)-1.\end{aligned}$$

(The sum $\sum_{i=1}^{k/2} v_i v_i'$ is $k/2$-subGaussian.) Therefore, $\chi^2 = O(1)$ when $n = O(\sqrt{k}/\varepsilon^2)$. □

8.2.4 Example 1.3: Linear functional of sparse parameters

Let $X \sim \mathcal{N}(\mu, I_d)$ with $\|\mu\|_0 \le s$.

Target.

$$\inf_T \sup_{\|\mu\|_0\le s} \mathbb{E}_\mu\Big(T-\sum_{i=1}^d \mu_i\Big)^2 \;\gtrsim\; s^2\log\Big(1+\frac{d}{s^2}\Big).$$

Proof of lower bound. Let

$$H_0:\ \mu=0 \quad (\text{call it } P), \qquad H_1:\ \mu=\rho\,\mathbb{1}_S,\quad S\sim \mathrm{Unif}\Big(\binom{[d]}{s}\Big) \quad (\text{call it } \mathbb{E}[P_S]).$$

The separation condition is satisfied with $\Delta \asymp \rho^2 s^2$. Also,

$$\int \frac{P_S P_{S'}}{P} = \int \frac{\varphi(x-\rho\,\mathbb{1}_S)\,\varphi(x-\rho\,\mathbb{1}_{S'})}{\varphi(x)}\,\mathrm{d}x = e^{\rho^2\langle \mathbb{1}_S,\mathbb{1}_{S'}\rangle} = e^{\rho^2|S\cap S'|}.$$

To proceed, note that $|S\cap S'| \sim \mathrm{Hypergeometric}(d,s,s)$. By Hoeffding's lemma (stated next),

$$\chi^2(\mathbb{E}[P_S]\,\|\,P)+1 = \mathbb{E}\big[e^{\rho^2|S\cap S'|}\big] \le \mathbb{E}\big[e^{\rho^2 B(s,s/d)}\big] = \Big(1-\frac{s}{d}+\frac{s}{d}e^{\rho^2}\Big)^s = O(1)$$

when $\rho \asymp \sqrt{\log(1+d/s^2)}$. □

Lemma 8.9 (Hoeffding)**.** *Let $C=\{c_1,\dots,c_N\}\subset\mathbb{R}$ be a fixed population. Let $X_1,\dots,X_n$ be n draws from C* without *replacement, and $X_1^*,\dots,X_n^*$ be n draws from C* with *replacement. Then for any convex $f:\mathbb{R}\to\mathbb{R}$,*

$$\mathbb{E}\Big[f\Big(\sum_{i=1}^n X_i\Big)\Big] \;\le\; \mathbb{E}\Big[f\Big(\sum_{i=1}^n X_i^*\Big)\Big].$$

8.2.5 Example 1.4: Quadratic functional estimation

Let $X_1,\dots,X_n \overset{\text{i.i.d.}}{\sim} f$, where the density f is supported on $[0,1]^d$ and $\big\|f^{(s)}\big\|_\infty = O(1)$ for some integer s.

Target.

$$\inf_T \sup_f \mathbb{E}_f\Big|T-\int_{[0,1]^d} f(x)^2\,\mathrm{d}x\Big| \;\asymp\; n^{-\frac{4s}{4s+d}}+n^{-1/2}.$$

Proof of the lower bound. The parametric rate $\Omega(n^{-1/2})$ is trivial (by LAN or a simple two-point argument). For the $\Omega\big(n^{-\frac{4s}{4s+d}}\big)$ lower bound, let

$$H_0: \ f \equiv 1, \qquad H_1: \ f_v(x) = 1 + c\sum_{i=1}^{h^{-d}} v_i\, h^s\, g\Big(\frac{x-c_i}{h}\Big), \qquad v \sim \mathrm{Unif}(\{\pm 1\}^{h^{-d}}),$$

where $g(\cdot)$ is a smooth function on $[0,1]^d$ with $\int g = 0$. The cube $[0,1]^d$ is partitioned into h^{-d} subcubes with edge length h; c_i is the lower-left corner of the i-th subcube.

For a small absolute constant $c > 0$, one can verify $\left\|f_v^{(s)}\right\|_\infty = O(1)$ for all v, and

$$\int_{[0,1]^d} f_v(x)^2\,\mathrm{d}x = 1 + c^2\sum_{i=1}^{h^{-d}} h^{2s}\int_{[0,1]^d} g^2\Big(\frac{x-c_i}{h}\Big)\,\mathrm{d}x = 1 + c^2 h^{2s}\,\|g\|_2^2\,.$$

Thus the separation condition holds with $\Delta \asymp h^{2s}$.

For indistinguishability,

$$\int \frac{f_v f_{v'}}{f} = 1 + \int_{[0,1]^d} c^2\sum_{i=1}^{h^{-d}} v_i v_i'\, h^{2s}\, g^2\Big(\frac{x-c_i}{h}\Big)\,\mathrm{d}x = 1 + c^2\,\|g\|_2^2\, h^{2s+d}\sum_{i=1}^{h^{-d}} v_i v_i'.$$

Therefore

$$\begin{aligned}\chi^2\big(\mathbb{E}[f_v^{\otimes n}]\,\|\, f^{\otimes n}\big) + 1 &\le \mathbb{E}\exp\Big(nc^2\,\|g\|_2^2\, h^{2s+d}\sum_{i=1}^{h^{-d}} v_i v_i'\Big)\\ &\le \exp\Big(O\big(n^2 h^{4s+2d}\cdot h^{-d}\big)\Big) = \exp\big(O(n^2 h^{4s+d})\big) = O(1)\end{aligned}$$

when $h \asymp n^{-\frac{2}{4s+d}}$. □

8.3 Advanced Le Cam II: mixture vs. mixture

Theorem 8.10 (Mixture vs. mixture). *Fix any $\Theta_0 \subseteq \Theta$ and $\Theta_1 \subseteq \Theta$. Suppose*

$$\inf_{\theta_0\in\Theta_0,\ \theta_1\in\Theta_1}\ \inf_a \big(L(\theta_0,a) + L(\theta_1,a)\big) \ \ge\ \Delta.$$

Then for any probability distributions π_0 and π_1,

$$\inf_T\ \sup_{\theta\in\Theta_0\cup\Theta_1} \mathbb{E}_\theta[L(\theta, T(X))] \ \ge\ \frac{\Delta}{2}\Big(1 - \mathrm{TV}(\mathbb{E}_{\theta\sim\pi_0}[P_\theta],\ \mathbb{E}_{\theta\sim\pi_1}[P_\theta]) - \pi_0(\Theta_0^c) - \pi_1(\Theta_1^c)\Big).$$

Proof. The only new observation is that if $\widetilde{\pi}_0$ is the restriction of π_0 on Θ_0, then

$$\mathrm{TV}\big(\mathbb{E}_{\theta\sim\pi_0}[P_\theta],\ \mathbb{E}_{\theta\sim\widetilde{\pi}_0}[P_\theta]\big) \le \mathrm{TV}(\pi_0, \widetilde{\pi}_0) = \pi_0(\Theta_0^c).$$

(Similarly for π_1.) □

Challenge. What is a good way to upper bound $\mathrm{TV}(\mathbb{E}_{\theta\sim\pi_0}[P_\theta], \mathbb{E}_{\theta\sim\pi_1}[P_\theta])$ beyond trivial convexity arguments?

8.3.1 Orthogonal functions/polynomials

Suppose $(P_\theta)_{\theta\in[\theta_0-\varepsilon,\,\theta_0+\varepsilon]}$ is a 1-D family of distributions, with likelihood ratio expansion

$$\frac{P_{\theta_0+u}(x)}{P_{\theta_0}(x)} = \sum_{m=0}^{\infty} p_m(x;\theta_0)\,\frac{u^m}{m!}, \qquad \text{for } |u|\le\varepsilon.$$

Under some structural conditions, $\{p_m(x;\theta_0)\}_{m\ge0}$ are orthogonal under P_{θ_0}.

Lemma 8.11. *If*

$$\int \frac{P_{\theta_0+u}P_{\theta_0+v}}{P_{\theta_0}} \quad \textit{depends only on } (\theta_0,\, uv),$$

then

$$\mathbb{E}_{X\sim P_{\theta_0}}\big[p_m(X;\theta_0)p_n(X;\theta_0)\big] = 0 \qquad \forall\ m\ne n.$$

Proof.

$$\begin{aligned}\int \frac{P_{\theta_0+u}P_{\theta_0+v}}{P_{\theta_0}} &= \mathbb{E}_{X\sim P_{\theta_0}}\Big[\Big(\sum_{m=0}^{\infty} p_m(X;\theta_0)\frac{u^m}{m!}\Big)\Big(\sum_{n=0}^{\infty} p_n(X;\theta_0)\frac{v^n}{n!}\Big)\Big]\\ &= \sum_{m,n\ge0} \mathbb{E}_{X\sim P_{\theta_0}}\big[p_m(X;\theta_0)p_n(X;\theta_0)\big]\,\frac{u^m v^n}{m!\,n!}.\end{aligned}$$

Since this quantity depends on (u,v) only through uv, all coefficients with $m\ne n$ must be 0. □

8.3.2 Two important examples

Gaussian. For $P_0=\mathcal{N}(0,1)$,

$$\int \frac{P_uP_v}{P_0} = \exp(uv).$$

The corresponding $p_m(x;\theta_0=0)$ are the Hermite polynomials $H_m(x)$, with

$$\mathbb{E}_{X\sim\mathcal{N}(0,1)}[H_m(X)H_n(X)] = n!\,\mathbb{1}\{m=n\}.$$

Poisson. For $P_0=\mathrm{Poi}(\lambda)$,

$$\int \frac{P_{\lambda+u}P_{\lambda+v}}{P_\lambda} = \sum_{k=0}^{\infty} e^{-2\lambda-u-v}\,\frac{((\lambda+u)(\lambda+v))^k}{k!\,\lambda^k} = \exp\Big(\frac{uv}{\lambda}\Big).$$

The corresponding $p_m(x;\theta_0=\lambda)$ are the Poisson–Charlier polynomials $C_m(x;\lambda)$, with

$$\mathbb{E}_{X\sim\mathrm{Poi}(\lambda)}[C_m(X;\lambda)C_n(X;\lambda)] = \frac{n!}{\lambda^n}\,\mathbb{1}\{m=n\}.$$

8.3.3 Bounding TV and χ^2: methods of moments

Theorem 8.12 (Gaussian mixture)**.** *For $\mu\in\mathbb{R}$ and random variables U,V,*

$$\mathrm{TV}\big(\mathbb{E}[\mathcal{N}(\mu+U,1)],\ \mathbb{E}[\mathcal{N}(\mu+V,1)]\big) \le \frac{1}{2}\Big(\sum_{m=0}^{\infty}\frac{(\mathbb{E}[U^m]-\mathbb{E}[V^m])^2}{m!}\Big)^{1/2}.$$

If in addition $\mathbb{E}[V]=0$ and $\mathbb{E}[V^2]\le M^2$, then

$$\chi^2\big(\mathbb{E}[\mathcal{N}(\mu+U,1)]\ \|\ \mathbb{E}[\mathcal{N}(\mu+V,1)]\big) \le e^{M^2/2}\sum_{m=0}^{\infty}\frac{(\mathbb{E}[U^m]-\mathbb{E}[V^m])^2}{m!}.$$

Proof (for the TV **bound).** WLOG assume $\mu = 0$, and let $\Delta_m := \mathbb{E}[U^m] - \mathbb{E}[V^m]$. Let φ denote the $\mathcal{N}(0,1)$ density.

$$\begin{aligned}
\mathrm{TV}\big(\mathbb{E}[\mathcal{N}(U,1)], \mathbb{E}[\mathcal{N}(V,1)]\big) &= \frac{1}{2}\int_{\mathbb{R}} |\mathbb{E}_U[\varphi(x-U)] - \mathbb{E}_V[\varphi(x-V)]| \, \mathrm{d}x \\
&= \frac{1}{2}\int_{\mathbb{R}} \varphi(x) \left| \mathbb{E}_U\Big[\sum_{m=0}^{\infty} H_m(x)\frac{U^m}{m!}\Big] - \mathbb{E}_V\Big[\sum_{m=0}^{\infty} H_m(x)\frac{V^m}{m!}\Big] \right| \mathrm{d}x \\
&= \frac{1}{2}\mathbb{E}_{X\sim\mathcal{N}(0,1)} \left| \sum_{m=0}^{\infty} H_m(X)\frac{\Delta_m}{m!} \right| \\
&\le \frac{1}{2}\Big(\mathbb{E}_{X\sim\mathcal{N}(0,1)}\Big[\Big(\sum_{m=0}^{\infty} H_m(X)\frac{\Delta_m}{m!}\Big)^2\Big]\Big)^{1/2} \\
&= \frac{1}{2}\Big(\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!}\Big)^{1/2},
\end{aligned}$$

using Cauchy–Schwarz and the orthogonality $\mathbb{E}[H_m(X)H_n(X)] = n!\,\mathbb{1}\{m=n\}$.

Proof sketch (for the χ^2 bound). For the χ^2 upper bound, lower bound the denominator as

$$\mathbb{E}_{\theta\sim V}[\varphi(x-\theta)] = \varphi(x)\,\mathbb{E}_{\theta\sim V}\big[\exp(\theta x - \tfrac{\theta^2}{2})\big] \ge \varphi(x)\,\exp\Big(\mathbb{E}_{\theta\sim V}[\theta x - \tfrac{\theta^2}{2}]\Big) \ge \varphi(x)e^{-M^2/2},$$

and the rest is the same as the TV proof. □

Theorem 8.13 (Poisson mixture)**.** *For $\lambda > 0$ and random variables U, V supported on $[-\lambda, \infty)$,*

$$\mathrm{TV}\big(\mathbb{E}[\mathrm{Poi}(\lambda+U)],\ \mathbb{E}[\mathrm{Poi}(\lambda+V)]\big) \le \frac{1}{2}\Big(\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!\,\lambda^m}\Big)^{1/2}, \qquad \Delta_m := \mathbb{E}[U^m] - \mathbb{E}[V^m].$$

If in addition $\mathbb{E}[V] = 0$ and $|V| \le M$, then

$$\chi^2\big(\mathbb{E}[\mathrm{Poi}(\lambda+U)] \parallel \mathbb{E}[\mathrm{Poi}(\lambda+V)]\big) \le e^M \sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!\,\lambda^m}.$$

Proof. Exercise (the same argument as the Gaussian case, but using Poisson–Charlier polynomials). □

8.3.4 Example 2.1: Generalized uniformity testing

Given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P = (p_1, \ldots, p_k)$, aim to test

$$H_0:\ P = \mathrm{Unif}(S) \text{ for some } S \subseteq [k] \qquad \text{vs.} \qquad H_1:\ \min_{S\subseteq[k]} \mathrm{TV}(P, \mathrm{Unif}(S)) \ge \varepsilon/2.$$

Target. The sample complexity for a reliable test is

$$n = \Theta\Big(\frac{\sqrt{k}}{\varepsilon^2} + \frac{k^{2/3}}{\varepsilon^{4/3}}\Big).$$

Proof of lower bound.

(1) $n = \Omega(\sqrt{k}/\varepsilon^2)$ follows from uniformity testing (Example 1.2).

(2) For $n = \Omega(k^{2/3}/\varepsilon^{4/3})$, assume Poissonization, where the observations are

$$(N_1, \dots, N_k), \qquad N_i \overset{\text{ind}}{\sim} \text{Poi}(np_i).$$

Construct two product priors:

$$\text{under } H_0 : \ p_1, \dots, p_k \overset{\text{i.i.d.}}{\sim} \text{Law}(U), \qquad \text{under } H_1 : \ p_1, \dots, p_k \overset{\text{i.i.d.}}{\sim} \text{Law}(V),$$

where

$$U = \begin{cases} 0, & \text{w.p. } \dfrac{\varepsilon^2}{1+\varepsilon^2}, \\ \dfrac{1+\varepsilon^2}{k}, & \text{w.p. } \dfrac{1}{1+\varepsilon^2}, \end{cases} \qquad V = \begin{cases} \dfrac{1-\varepsilon}{k}, & \text{w.p. } \dfrac{1}{2}, \\ \dfrac{1+\varepsilon}{k}, & \text{w.p. } \dfrac{1}{2}. \end{cases}$$

Notes.

(1) Under H_0, $p_i \in \{0, (1+\varepsilon^2)/k\}$, so $(p_1, \dots, p_k)$ is generalized uniform.

(2) Under H_1, $(p_1, \dots, p_k)$ is $\Omega(\varepsilon)$-far from generalized uniform w.h.p.

(3) $\mathbb{E}[U] = \mathbb{E}[V] = 1/k$, so under both H_0 and H_1, $(p_1, \dots, p_k)$ is a pmf *in expectation.* (Additional arguments are needed to justify restricting to "approximate pmfs"; omitted here.)

(4) $\mathbb{E}[U^2] = \mathbb{E}[V^2] = (1+\varepsilon^2)/k^2$, and

$$|\mathbb{E}[(U - 1/k)^m] - \mathbb{E}[(V-1/k)^m]| \le \frac{2\varepsilon^2}{k^m}, \qquad m \ge 3.$$

Now by the Poisson mixture result,

$$\chi^2\big(\mathbb{E}[\text{Poi}(nU)] \parallel \mathbb{E}[\text{Poi}(nV)]\big) \le e^{n\varepsilon/k} \sum_{m=3}^{\infty} \frac{4\varepsilon^4\,(n/k)^{2m}}{m!\,(n/k)^m} = e^{n\varepsilon/k} \sum_{m=3}^{\infty} \frac{4\varepsilon^4\,(n/k)^m}{m!} = O\Big(\frac{n^3\varepsilon^4}{k^3}\Big).$$

Tensorization of χ^2 yields

$$\chi^2\Big(\mathbb{E}_U\big[\bigotimes_{i=1}^k \text{Poi}(np_i)\big] \Bigm\| \mathbb{E}_V\big[\bigotimes_{i=1}^k \text{Poi}(np_i)\big]\Big) + 1 \le \Big(1 + O\Big(\frac{n^3\varepsilon^4}{k^3}\Big)\Big)^k \le \exp\Big(O\Big(\frac{n^3\varepsilon^4}{k^2}\Big)\Big) = O(1)$$

if $n = O(k^{2/3}/\varepsilon^{4/3})$. □

Remark 8.14. This construction matches the first two moments of (U, V). Can we match more? No.

Lemma 8.15. *Let μ be a probability measure supported on $\{0, x_1, \dots, x_{k-1}\} \subset [0, \infty)$. Let ν be another probability measure supported on $[0,\infty)$ such that*

$$\mathbb{E}_\mu[X^m] = \mathbb{E}_\nu[X^m] \qquad \textit{for all } m = 0, 1, \dots, 2k-1.$$

Then $\mu = \nu$.

Proof.

$$0 = \mathbb{E}_\mu\big[X(X-x_1)^2\cdots(X-x_{k-1})^2\big] = \mathbb{E}_\nu\big[X(X-x_1)^2\cdots(X-x_{k-1})^2\big] \geq 0.$$

Hence $\operatorname{supp}(\nu) \subset \{0, x_1, \ldots, x_{k-1}\}$, which forces $\nu = \mu$. □

8.4 Example 2.2: ℓ_1-norm estimation

Let $X \sim \mathcal{N}(\theta, I_n)$ with $\|\theta\|_\infty \leq 1$.

Target.

$$\inf_T \sup_{\|\theta\|_\infty \leq 1} \mathbb{E}_\theta\big|T - \|\theta\|_1\big| \asymp n \cdot \frac{\log\log n}{\log n}.$$

Proof of lower bound (idea). Test between $H_0 : \|\theta\|_1 \leq \rho_0$ vs. $H_1 : \|\theta\|_1 \geq \rho_1$. Assign priors $\theta \sim \mu_0^{\otimes n}$ under H_0 and $\theta \sim \mu_1^{\otimes n}$ under H_1.

Desired properties.

(1) $\chi^2\big(\mu_0 * \mathcal{N}(0,1) \parallel \mu_1 * \mathcal{N}(0,1)\big) = O(1/n)$.

(2) $\mu_0^{\otimes n}(H_0^c) + \mu_1^{\otimes n}(H_1^c) = o(1)$.

(3) $\rho_1 - \rho_0 = \Omega\big(n \cdot \frac{\log\log n}{\log n}\big)$.

We design $(\rho_0, \rho_1, \mu_0, \mu_1)$ for these properties separately.

(1) Moment matching controls χ^2. If μ_0, μ_1 match the first K moments, then

$$\chi^2\big(\mu_0 * \mathcal{N}(0,1) \parallel \mu_1 * \mathcal{N}(0,1)\big) \leq O(1) \sum_{m=K+1}^{\infty} \frac{2^{m+1}}{m!} \leq \Big(\frac{O(1)}{K}\Big)^K.$$

(To make it $O(1/n)$, choose $K \asymp \frac{\log n}{\log\log n}$.)

(2) Choose thresholds using concentration. Choose

$$\rho_0 = n\,\mathbb{E}_{\mu_0}|\theta| + \omega(\sqrt{n}), \qquad \rho_1 = n\,\mathbb{E}_{\mu_1}|\theta| - \omega(\sqrt{n}).$$

Since under $\mu_0^{\otimes n}$, $\|\theta\|_1$ concentrates around $n\mathbb{E}_{\mu_0}|\theta|$ with fluctuations $O(\sqrt{n})$, Chebyshev gives $\mu_0^{\otimes n}(H_0^c), \mu_1^{\otimes n}(H_1^c) = o(1)$.

(3) Remaining optimization problem. It remains to solve

$$\begin{aligned} \max\ & \mathbb{E}_{\mu_1}|\theta| - \mathbb{E}_{\mu_0}|\theta| \\ \text{s.t. } & \mu_0, \mu_1 \text{ supported on } [-1,1], \\ & \mathbb{E}_{\mu_1}[\theta^m] = \mathbb{E}_{\mu_0}[\theta^m] \quad \text{for } 0 \leq m \leq K. \end{aligned}$$

There is a duality result between moment matching and best polynomial approximation.

8.4.1 Duality: moment matching vs. best polynomial approximation

Theorem 8.16. *Let $I \subset \mathbb{R}$ be compact and f continuous on I. Define*

$$V^* := \max\Big\{\mathbb{E}_\mu[f(X)] - \mathbb{E}_\nu[f(X)] : \operatorname{supp}(\mu), \operatorname{supp}(\nu) \subseteq I,\ \mathbb{E}_\mu[X^m] = \mathbb{E}_\nu[X^m]\ \forall m = 0, \ldots, K\Big\},$$

and

$$E^* := \inf_{P:\,\deg(P)\le K} \sup_{x\in I} |f(x) - P(x)|\,.$$

Then $V^* = 2E^*$.

Proof. *Step 1: $V^* \le 2E^*$.* Let P be any polynomial with $\deg(P) \le K$. If μ, ν match moments up to degree K, then $\mathbb{E}_\mu[P] = \mathbb{E}_\nu[P]$ and hence

$$\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f] = \mathbb{E}_\mu[f-P] - \mathbb{E}_\nu[f-P] \le \mathbb{E}_\mu|f-P| + \mathbb{E}_\nu|f-P| \le 2\sup_{x\in I}|f(x) - P(x)|\,.$$

Taking inf over P yields $V^* \le 2E^*$.

Step 2: $V^* \ge 2E^*$. Let $\mathcal{F} := \operatorname{span}\{1, x, \ldots, x^K, f(x)\}$. Define a linear functional L on $\mathcal{F}$ by

$$L(x^m) = 0\ (m = 0, \ldots, K), \qquad L(f) = E^*.$$

We claim $\|L\| = 1$, where

$$\|L\| := \sup_{h\in\mathcal{F},\ \|h\|_{L_\infty(I)}\le 1} |Lh|\,.$$

Let $P^*(x)$ be a best approximating polynomial of degree $\le K$ such that $\|f - P^*\|_{L_\infty(I)} = E^*$. Any $h \in \mathcal{F}$ can be written as $h = c(f - P^*) + P$ for some polynomial P of degree $\le K$. By definition of P^*, $\|h\|_{L_\infty(I)} \ge |c|\ E^*$. Thus

$$\frac{|Lh|}{\|h\|_{L_\infty(I)}} = \frac{|c|\ E^*}{\|h\|_{L_\infty(I)}} \le 1,$$

with equality for $P \equiv 0$, hence $\|L\| = 1$.

By Hahn–Banach, extend L to $C(I)$ with $\|L\| = 1$. By Riesz representation, there exists a signed measure μ on I such that

$$Lh = \int_I h \,\mathrm{d}\mu.$$

Let $\mu = \mu_+ - \mu_-$ be the Jordan decomposition. Since $L(1) = 0$, we have $\mu_+(I) = \mu_-(I)$. Since $\|L\| = 1$, we have $\mu_+(I) + \mu_-(I) = 1$. Hence $\mu_+(I) = \mu_-(I) = 1/2$. Also, $L(x^m) = 0$ implies

$$\int_I x^m \,\mathrm{d}\mu_+ = \int_I x^m \,\mathrm{d}\mu_- \qquad \text{for all } m = 0, \ldots, K.$$

Finally choose $\mu_1 = 2\mu_+$ and $\mu_0 = 2\mu_-$. Then

$$E^* = Lf = \mathbb{E}_{\mu_+}[f] - \mathbb{E}_{\mu_-}[f] = \frac{1}{2}\big(\mathbb{E}_{\mu_1}[f] - \mathbb{E}_{\mu_0}[f]\big),$$

so $\mathbb{E}_{\mu_1}[f] - \mathbb{E}_{\mu_0}[f] = 2E^*$ and therefore $V^* \ge 2E^*$. Combining both steps gives $V^* = 2E^*$. □

Remark 8.17. By approximation theory, the uniform approximation error of $|\theta|$ by $\operatorname{span}\{1, \theta, \ldots, \theta^K\}$ is $\Theta(1/K)$. So we get

$$\rho_1 - \rho_0 = \Omega(n/K) = \Omega\Big(n \cdot \frac{\log\log n}{\log n}\Big),$$

and combining (1)–(3) yields the target lower bound.

8.5 Special topic: dualizing Le Cam (Polyanskiy & Wu 2019)

8.5.1 Setting

Let $\theta_1, \dots, \theta_n \overset{\text{i.i.d.}}{\sim} \pi$ and $X_i \mid \theta_i \sim P_{\theta_i}$. Equivalently,

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{E}_{\theta\sim\pi}[P_\theta] =: \pi P.$$

Target: estimate a linear functional $T(\pi)$ and characterize

$$r^* = \inf_{\widehat{T}} \sup_{\pi\in\Pi} \mathbb{E}_\pi\big[(\widehat{T}(X_1, \dots, X_n) - T(\pi))^2\big].$$

Remark 8.18 (Related setting). If $(\theta_1, \dots, \theta_n)$ is an individual sequence and $X_i \mid \theta_i \sim P_{\theta_i}$, then the target is to estimate

$$T(\pi_\theta) = \frac{1}{n}\sum_{i=1}^n h(\theta_i), \qquad \text{where } \pi_\theta = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i},$$

which is linear in π_θ. This covers functional estimation such as ℓ_1-norm estimation in Example 2.2.

8.5.2 A modulus-of-continuity characterization

Definition 8.19 (χ^2-modulus of continuity)**.** For $t \geq 0$,

$$\delta_{\chi^2}(t) := \sup\Big\{ \big|T(\pi') - T(\pi)\big| : \chi^2(\pi' P \,\|\, \pi P) \leq t^2,\ \pi, \pi' \in \Pi \Big\}.$$

Theorem 8.20. *If T is linear and Π is convex, under regularity conditions,*

$$\frac{1}{7}\,\delta_{\chi^2}(1/\sqrt{n})^2 \ \leq\ r^* \ \leq\ \delta_{\chi^2}(1/\sqrt{n})^2.$$

Remark 8.21. (1) δ_{χ^2} is the best separation constant subject to the χ^2 indistinguishability constraint, and the lower bound $r^* \geq \frac{1}{7}\delta_{\chi^2}(1/\sqrt{n})^2$ follows from Le Cam's two-point method.

(2) The upper bound shows that for linear T, Le Cam's method can be *dualized* to obtain statistical upper bounds.

8.5.3 Proof of the upper bound

Try an estimator of the form

$$\widehat{T}(X_1, \dots, X_n) = \frac{1}{n}\sum_{i=1}^n g(X_i) \qquad \text{for some } g : \mathcal{X} \to \mathbb{R}.$$

By bias–variance analysis,

$$\sup_{\pi\in\Pi} \mathbb{E}_\pi\big[(\widehat{T} - T(\pi))^2\big] = \sup_{\pi\in\Pi} \Big\{ |T(\pi) - \pi P_g|^2 + \frac{1}{n}\operatorname{Var}_{\pi P}(g) \Big\},$$

where $\pi P_g := \mathbb{E}_{X\sim\pi P}[g(X)]$. Thus it suffices to show

$$\inf_g \sup_{\pi\in\Pi} \Big\{ |T(\pi) - \pi P_g| + \frac{1}{\sqrt{n}}\sqrt{\operatorname{Var}_{\pi P}(g)} \Big\} \ \leq\ \delta_{\chi^2}(1/\sqrt{n}). \tag{8.1}$$

Denote

$$L(\pi, g) := |T(\pi) - \pi P_g| + \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}(g)}.$$

To mitigate the non-concavity of the absolute value term in π, write

$$\begin{aligned} L(\pi, g) &\le \sup_{\pi'} \sup_{0\le\xi\le 2} \Big\{ \big(T(\pi) - \pi P_g\big) - \xi\big(T(\pi') - \pi' P_g\big) + \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}(g)} \Big\} \\ &= \sup_{\pi_2 \in \Pi_2} \Big\{ \big(T(\pi) - \pi P_g\big) - \big(T(\pi_2) - \pi_2 P_g\big) + \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}(g)} \Big\}, \end{aligned}$$

where

$$\Pi_2 := \{\xi\pi' : \pi' \in \Pi,\ 0 \le \xi \le 2\}.$$

The right-hand side is concave in (π, π_2) (thanks to linearity of T). Therefore, by a minimax theorem,

$$\begin{aligned} \inf_g \sup_{\pi\in\Pi} L(\pi, g) &\le \sup_{\pi\in\Pi,\ \pi_2\in\Pi_2} \inf_g \Big\{ \big(T(\pi) - \pi P_g\big) - \big(T(\pi_2) - \pi_2 P_g\big) + \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}(g)} \Big\} \\ &= \sup_{\pi,\pi'\in\Pi} \inf_g \Big\{ \big(T(\pi) - T(\pi')\big) + (\pi' - \pi)P_g + \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}(g)} \Big\}. \end{aligned}$$

Recall the dual representation

$$\chi^2(\pi' P \,\|\, \pi P) = \sup\Big\{ \big((\pi' - \pi)P_g\big)^2 : \mathrm{Var}_{\pi P}(g) \le 1 \Big\}.$$

If $\chi^2(\pi' P \,\|\, \pi P) > 1/n$, then there exists g_0 with $\mathrm{Var}_{\pi P}(g_0) \le 1$ and $(\pi' - \pi)P_{g_0} < -1/\sqrt{n}$. Choosing $g = cg_0$ and letting $c \to \infty$ gives

$$\inf_g \Big\{ (\pi' - \pi)P_g + \frac{1}{\sqrt{n}}\sqrt{\mathrm{Var}_{\pi P}(g)} \Big\} = -\infty.$$

On the other hand, if $\chi^2(\pi' P \,\|\, \pi P) \le 1/n$, then the infimum above is 0 (achieved by $g \equiv 0$). Hence

$$\inf_g \sup_{\pi\in\Pi} L(\pi, g) \le \sup_{\pi,\pi'\in\Pi} \Big\{ T(\pi) - T(\pi') : \chi^2(\pi' P \,\|\, \pi P) \le 1/n \Big\} = \delta_{\chi^2}(1/\sqrt{n}),$$

which proves (8.1). □

8.5.4 Example: Fisher's species problem

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p$ where p is supported on $\mathbb{N}$. Let $m = nr$, and hypothetically draw $X_1', \ldots, X_m' \overset{\text{i.i.d.}}{\sim} p$. Aim to estimate

$$U := \big|\{X_1', \ldots, X_m'\} \setminus \{X_1, \ldots, X_n\}\big| \qquad (\#\text{ of “new” species}).$$

Question. Characterize

$$r^* := \inf_{\widehat{U}} \sup_p \mathbb{E}_p\Big[\frac{1}{n^2}(\widehat{U} - U)^2\Big].$$

Answer.

$$r^* = \begin{cases} \Theta(1/n), & r \le 1, \\ \widetilde{\Theta}\big(n^{-\frac{2}{r+1}}\big), & r > 1. \end{cases}$$

8.5.5 Proof of the upper bound (sketch)

First we make some simplifications.

(1) **Poissonization.** The histograms $N_x = \sum_{i=1}^n \mathbb{1}(X_i = x) \sim \mathrm{Poi}(np_x)$ and $N'_x \sim \mathrm{Poi}(mp_x)$ are independent Poisson r.v.s.

(2) **Replace by expectation.** One can show $U \approx \mathbb{E}U$ w.h.p., so it is equivalent to estimate

$$\mathbb{E}[U] = \mathbb{E}\Big[\sum_x \mathbb{1}(N_x = 0,\ N'_x > 0)\Big] = \sum_x e^{-np_x}\,(1 - e^{-mp_x}).$$

(3) **Support size.** The support size of p is at most $O(n)$. In this case, let $\theta_x = np_x$ and let $\pi := \mathrm{Unif}(\{\theta_x\})$. Then $\frac{1}{n}\mathbb{E}[U]$ is equivalent to

$$\mathbb{E}_{\theta\sim\pi}[h(\theta)] = \mathbb{E}_{\theta\sim\pi}\big[e^{-\theta} - e^{-(1+r)\theta}\big], \qquad h(\theta) := e^{-\theta} - e^{-(1+r)\theta}.$$

By the previous result (dualizing Le Cam), it suffices to show that for $P = \mathrm{Poi}$,

$$\delta_{\chi^2}(1/\sqrt{n}) = \sup\Big\{\,|\mathbb{E}_{\pi'-\pi}[h(\theta)]| : \chi^2(\pi' P \,\|\, \pi P) \le \frac{1}{n}\Big\} \;\lesssim\; n^{-\min\{\frac{1}{2},\,\frac{1}{1+r}\}}.$$

Let $t = 1/\sqrt{n}$. Since $\chi^2 \le t^2$ implies $\mathrm{TV} \le t$, we have

$$\delta_{\chi^2}(t) \le \sup\Big\{\Big|\int h\,\mathrm{d}\Delta\Big| : \|\Delta\|_{\mathrm{TV}} \le 1,\ \|\Delta P\|_{\mathrm{TV}} \le t\Big\},$$

where $\Delta := \pi' - \pi$ is a signed measure. To upper bound this quantity we use complex analysis.

8.5.6 Complex analysis bound

Let

$$f_\Delta(z) := \int_{\mathbb{R}_+} e^{z\theta}\,\Delta(\,\mathrm{d}\theta) \qquad \text{(Laplace transform)}$$

and

$$f_{\Delta P}(z) := \sum_{m=0}^{\infty} z^m\,\Delta P(m) \qquad (z\text{-transform}).$$

Then

$$\int h\,\mathrm{d}\Delta = \int_{\mathbb{R}_+} (e^{-\theta} - e^{-(1+r)\theta})\,\Delta(\,\mathrm{d}\theta) = f_\Delta(-1) - f_\Delta(-1-r).$$

In addition,

$$\begin{aligned} f_{\Delta P}(z) &= \sum_{m=0}^{\infty} z^m \int e^{-\theta}\,\frac{\theta^m}{m!}\,\Delta(\,\mathrm{d}\theta) = \int e^{-\theta}\Big(\sum_{m=0}^{\infty}\frac{(z\theta)^m}{m!}\Big)\Delta(\,\mathrm{d}\theta) \\ &= \int e^{(z-1)\theta}\,\Delta(\,\mathrm{d}\theta) = f_\Delta(z-1). \end{aligned}$$

Finally,

$$|f_\Delta(z)| \le \int_{\mathbb{R}_+} |\Delta|\,(\,\mathrm{d}\theta) \le 2 \qquad \text{for } \Re(z) \le 0,$$

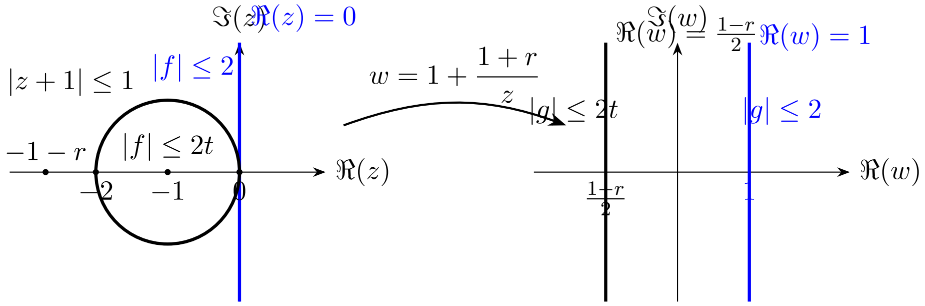


Figure 8.1: Conformal map used in the complex-analysis argument (Lecture 8, p. 16).

and

$$|f_{\Delta P}(z)| \le \sum_{m=0}^{\infty} |z|^m \, |\Delta P(m)| \le 2t \qquad \text{for } \ |z| \le 1.$$

Consequently,

$$\begin{aligned}\delta_{\chi^2}(t) &\le \sup_{\Delta} \Big\{ |f_\Delta(-1) - f_\Delta(-1-r)| : \|f_\Delta\|_{H^\infty(\Re z \le 0)} \le 2,\ \|f_\Delta\|_{H^\infty(D-1)} \le 2t \Big\} \\ &\le \sup_{f} \Big\{ |f(-1) - f(-1-r)| : \|f\|_{H^\infty(\Re z \le 0)} \le 2,\ \|f\|_{H^\infty(D-1)} \le 2t,\ f \text{ holomorphic on } \{z : \Re(z) \le 0\} \Big\},\end{aligned}$$

where $D - 1 := \{z : |z+1| \le 1\}$.

Case 1: $r \le 1$. Then $-1 - r \in D - 1$, so

$$|f(-1) - f(-1-r)| \le 4t.$$

Case 2: $r > 1$. Consider the Möbius transformation

$$w = \phi(z) := 1 + \frac{1+r}{z}, \qquad z = \phi^{-1}(w) = \frac{1+r}{w-1}.$$

Define $g(w) := f(\phi^{-1}(w))$, i.e. $f(z) = g(\phi(z))$. The map ϕ sends the imaginary axis $\Re(z) = 0$ to the vertical line $\Re(w) = 1$, and it sends the circle $|z+1| = 1$ to the vertical line $\Re(w) = \frac{1-r}{2}$; moreover $\phi(-1-r) = 0$ (see Figure 8.1). Thus g is holomorphic on the strip $\{w : \frac{1-r}{2} \le \Re(w) \le 1\}$ and

$$\|g\|_{H^\infty(\Re w = 1)} \le 2, \qquad \|g\|_{H^\infty(\Re w = \frac{1-r}{2})} \le 2t.$$

By Hadamard's three-line theorem (evaluated at $\Re(w) = 0$),

$$|f(-1-r)| = |g(0)| \le \|g\|_{H^\infty(\Re w = \frac{1-r}{2})}^{\frac{2}{1+r}} \, \|g\|_{H^\infty(\Re w = 1)}^{\frac{r-1}{1+r}} = O\big(t^{\frac{2}{1+r}}\big).$$

Therefore

$$|f(-1-r) - f(-1)| \le |f(-1-r)| + |f(-1)| = O\big(t^{\frac{2}{1+r}} + t\big) = O\big(t^{\frac{2}{1+r}}\big), \qquad \text{as } t = 1/\sqrt{n} \le 1.$$

Combining both cases yields $\delta_{\chi^2}(1/\sqrt{n}) \lesssim n^{-\min\{\frac{1}{2}, \frac{1}{1+r}\}}$, which gives the stated rates.

Lecture 9: Advanced Fano's Method

9.1 Covering and packing

Let $(\mathcal{X}, d)$ be a metric space and let $A \subseteq \mathcal{X}$ be compact.

Definition 9.1 (Covering / net)**.** A finite set $\{x_1, \dots, x_n\} \subseteq \mathcal{X}$ is an *ε-covering* (or *ε-net*) of A if

$$A \subseteq \bigcup_{i=1}^{n} B(x_i; \varepsilon), \qquad B(x; \varepsilon) := \{y \in \mathcal{X} : d(x, y) \le \varepsilon\}.$$

Definition 9.2 (Packing)**.** A finite set $\{a_1, \dots, a_m\} \subseteq A$ is an *ε-packing* of A if

$$\min_{i \ne j} d(a_i, a_j) > \varepsilon.$$

Definition 9.3 (Covering and packing numbers)**.**

$N(A, d, \varepsilon) := \min\{n : \exists\, \varepsilon\text{-covering of } A \text{ of size } n\}, \qquad M(A, d, \varepsilon) := \max\{m : \exists\, \varepsilon\text{-packing of } A \text{ of size } m\}.$

9.1.1 Basic relationship

Lemma 9.4. *For every $\varepsilon > 0$,*

$$M(A, d, 2\varepsilon) \le N(A, d, \varepsilon) \le M(A, d, \varepsilon).$$

In other words, up to a multiplicative factor 2 on ε, it is equivalent to consider covering or packing numbers.

Proof. **(1)** $M(A, d, 2\varepsilon) \le N(A, d, \varepsilon)$**.** Assume for contradiction that $M(A, d, 2\varepsilon) \ge N(A, d, \varepsilon) + 1$. Take an ε-covering $\{y_1, \dots, y_N\}$ of A with $N = N(A, d, \varepsilon)$. Also take a (2ε)-packing $\{x_1, \dots, x_{N+1}\}$ of A. By the pigeonhole principle, two points x, x' in this packing belong to the same ball $B(y; \varepsilon)$ of the covering. Hence

$$d(x, x') \le d(x, y) + d(x', y) \le 2\varepsilon,$$

contradicting that the set is a (2ε)-packing.

(2) $N(A, d, \varepsilon) \le M(A, d, \varepsilon)$**.** Let $\{a_1, \dots, a_m\}$ be a maximal ε-packing of A. We claim it is also an ε-covering of A. If not, then there exists $a \in A$ such that $d(a, a_i) > \varepsilon$ for all $i \in [m]$. But then $\{a_1, \dots, a_m, a\}$ is a larger ε-packing, contradicting maximality. □

9.1.2 Bounding covering/packing numbers: a volume bound

Let $\|\cdot\|$ be any norm on $\mathbb{R}^d$ and let

$$B := \{x \in \mathbb{R}^d : \|x\| \le 1\}$$

be its unit ball.

Lemma 9.5 (Volume bound)**.** *For every $\varepsilon > 0$,*

$$\left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)} \le N(A, \|\cdot\|, \varepsilon) \le M(A, \|\cdot\|, \varepsilon) \le \left(\frac{2}{\varepsilon}\right)^d \frac{\mathrm{vol}\big(A + (\varepsilon/2)B\big)}{\mathrm{vol}(B)}.$$

Here $A + (\varepsilon/2)B := \{a + (\varepsilon/2)b : a \in A,\ b \in B\}$ is the Minkowski sum.

Proof. **(Lower bound on N).** Let $\{x_1, \ldots, x_n\}$ be an ε-covering of A. Then

$$A \subseteq \bigcup_{i=1}^n B(x_i; \varepsilon) \quad \Longrightarrow \quad \mathrm{vol}(A) \le \sum_{i=1}^n \mathrm{vol}\big(B(x_i; \varepsilon)\big) = n\, \varepsilon^d\, \mathrm{vol}(B).$$

Hence

$$n \ge \left(\frac{1}{\varepsilon}\right)^d \frac{\mathrm{vol}(A)}{\mathrm{vol}(B)}.$$

Taking the minimum over all coverings gives the stated lower bound on $N(A, \|\cdot\|, \varepsilon)$.

(Upper bound on M). Let $\{a_1, \ldots, a_m\}$ be an ε-packing of A. Then the balls $B(a_i; \varepsilon/2)$ are disjoint, and

$$\bigcup_{i=1}^m B(a_i; \varepsilon/2) \subseteq A + (\varepsilon/2)B.$$

Therefore

$$\mathrm{vol}\big(A + (\varepsilon/2)B\big) \ge \sum_{i=1}^m \mathrm{vol}\big(B(a_i; \varepsilon/2)\big) = m \left(\frac{\varepsilon}{2}\right)^d \mathrm{vol}(B),$$

so

$$m \le \left(\frac{2}{\varepsilon}\right)^d \frac{\mathrm{vol}\big(A + (\varepsilon/2)B\big)}{\mathrm{vol}(B)}.$$

Taking the maximum over packings gives the claimed upper bound on $M(A, \|\cdot\|, \varepsilon)$. The middle inequality $N \le M$ is from the basic relationship. □

Example 1.1 (Unit ball). If $A = B = \{x : \|x\| \le 1\}$ is the unit ball under the same norm, then for all $0 < \varepsilon \le 1$,

$$\left(\frac{1}{\varepsilon}\right)^d \le N(A, \|\cdot\|, \varepsilon) \le \left(1 + \frac{2}{\varepsilon}\right)^d \le \left(\frac{3}{\varepsilon}\right)^d.$$

Example 1.2 (Gilbert–Varshamov bound). Let $A = \{0,1\}^d$ and let the Hamming distance be

$$d_H(x, x') := \sum_{i=1}^{d} \mathbb{1}\{x_i \neq x'_i\}.$$

Then for $1 \leq r \leq d-1$,

$$\frac{2^d}{\sum_{i=0}^{r} \binom{d}{i}} \leq M(A, d_H, r) \leq \frac{2^d}{\sum_{i=0}^{\lfloor r/2 \rfloor} \binom{d}{i}}.$$

If $r = pd$ and $d \to \infty$, then by Stirling approximation,

$$2^{d(1-h(p)+o(1))} \leq M\big(\{0,1\}^d, d_H, pd\big) \leq 2^{d(1-h(p/2)+o(1))},$$

where $h(p) = p\log_2(1/p) + (1-p)\log_2(1/(1-p))$ is the binary entropy.

9.1.3 Sudakov minoration

Define the *Gaussian width* of a set $A \subseteq \mathbb{R}^d$ by

$$\omega(A) := \mathbb{E} \sup_{a \in A} \langle a, Z \rangle, \qquad Z \sim \mathcal{N}(0, I_d).$$

Lemma 9.6 (Sudakov minoration). *There exists a universal constant $C > 0$ such that*

$$\omega(A) \geq C \sup_{\varepsilon > 0} \varepsilon \sqrt{\log M\big(A, \|\cdot\|_2, \varepsilon\big)}.$$

Two results used in the proof:

(1) **Slepian's lemma.** Let X, Y be centered Gaussian random vectors in $\mathbb{R}^d$. If $\mathbb{E}[(Y_i - Y_j)^2] \leq \mathbb{E}[(X_i - X_j)^2]$ for all $i, j \in [d]$, then $\mathbb{E}[\max_i Y_i] \leq \mathbb{E}[\max_i X_i]$.

(2) **Maximum of Gaussians.** If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, then

$$\mathbb{E}\big[\max_{1 \leq i \leq n} X_i\big] = (1 + o(1))\sqrt{2\log n}.$$

Proof of Sudakov minoration. Let $\{a_1, \ldots, a_m\}$ be an optimal ε-packing of A. Define

$$X_i := \langle a_i, Z \rangle \quad (Z \sim \mathcal{N}(0, I_d)), \qquad Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}\Big(0, \frac{\varepsilon^2}{2}\Big).$$

Then for all $i \neq j$,

$$\mathbb{E}[(Y_i - Y_j)^2] = \varepsilon^2 \leq \|a_i - a_j\|_2^2 = \mathbb{E}[(X_i - X_j)^2].$$

By Slepian's lemma and the definition of Gaussian width,

$$\omega(A) = \mathbb{E}\sup_{a \in A} \langle a, Z \rangle \geq \mathbb{E} \max_{1 \leq i \leq m} X_i \geq \mathbb{E} \max_{1 \leq i \leq m} Y_i.$$

Finally, using the "maximum of Gaussians" fact with standard deviation $\varepsilon/\sqrt{2}$,

$$\mathbb{E} \max_{1 \leq i \leq m} Y_i = \frac{\varepsilon}{\sqrt{2}}(1 + o(1))\sqrt{2\log m}.$$

Thus $\omega(A) \gtrsim \varepsilon\sqrt{\log m}$, and taking the supremum over $\varepsilon > 0$ yields the result. □

Example 1.3. When $A = B_1 := \{x \in \mathbb{R}^d : \|x\|_1 \le 1\}$, then

$$\omega(A) = \mathbb{E} \sup_{\|x\|_1 \le 1} \langle x, Z\rangle = \mathbb{E}\,\|Z\|_\infty \le \sqrt{2\log d}.$$

Hence Sudakov minoration implies

$$\log M\big(B_1, \|\cdot\|_2, \varepsilon\big) = O\Big(\frac{\log d}{\varepsilon^2}\Big).$$

In fact, one has the (nearly) sharp upper bound

$$\log M\big(B_1, \|\cdot\|_2, \varepsilon\big) \lesssim \begin{cases} d\Big(1 + \log \frac{1}{\varepsilon\sqrt{d}}\Big), & \varepsilon \le 1/\sqrt{d} \quad \text{(volume bound is tight)},\\ \dfrac{1+\log(\varepsilon^2 d)}{\varepsilon^2}, & 1/\sqrt{d} \ll \varepsilon \ll 1 \quad \text{(Sudakov nearly tight)}. \end{cases}$$

9.1.4 Maurey's empirical method

Let $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ be an inner product space and let $T \subset \mathcal{H}$ be a finite set.

Lemma 9.7 (Maurey's empirical method)**.** *Let $r := \inf_{y\in\mathcal{H}} \sup_{x\in T} \|x-y\|$ be the radius of T. Then for every $0 < \varepsilon \le r$,*

$$N\big(\mathrm{conv}(T), \|\cdot\|, \varepsilon\big) \le \binom{|T| + \lceil r^2/\varepsilon^2\rceil - 2}{\lceil r^2/\varepsilon^2\rceil - 1}.$$

Proof. Write $T = \{t_1, \dots, t_m\}$ and choose $c \in \mathcal{H}$ such that $r = \max_{i\in[m]} \|t_i - c\|$. Fix any $x \in \mathrm{conv}(T)$, so $x = \sum_{i=1}^m x_i t_i$ with $x_i \ge 0$ and $\sum_i x_i = 1$. Let Z be an $\mathcal{H}$-valued random variable with $\mathbb{P}(Z = t_i) = x_i$; then $x = \mathbb{E}[Z]$. Let $Z_1, \dots, Z_n$ be i.i.d. copies of Z and define

$$\bar{Z} := \frac{1}{n+1}\Big(c + \sum_{i=1}^n Z_i\Big).$$

Then

$$\begin{aligned} \mathbb{E}\,\big\|\bar{Z} - x\big\|^2 &= \frac{1}{(n+1)^2}\Big(\|c - x\|^2 + n\,\mathbb{E}\,\|Z - x\|^2\Big)\\ &\le \frac{1}{(n+1)^2}\Big(r^2 + n\,\mathbb{E}\,\|Z - c\|^2\Big) \le \frac{r^2}{n+1}. \end{aligned}$$

(Here $\|c - x\|^2 \le r^2$ by convexity, and $\mathbb{E}\,\|Z - x\|^2 \le \mathbb{E}\,\|Z - c\|^2 \le r^2$ since $\mathbb{E}[Z] = x$.)

Consequently, if $n = \lceil r^2/\varepsilon^2\rceil - 1$, there exists a realization of $\bar{Z}$ such that $\|x - \bar{Z}\| \le \varepsilon$. Moreover $\bar{Z}$ always belongs to the finite set

$$\Big\{\frac{1}{n+1}\Big(c + \sum_{i=1}^m n_i t_i\Big) : \ n_i \ge 0, \ \sum_{i=1}^m n_i = n\Big\},$$

whose cardinality is $\binom{n+m-1}{n}$ by stars-and-bars. Thus $\mathrm{conv}(T)$ can be ε-covered by at most $\binom{n+m-1}{n}$ points, yielding the claimed bound. □

Example 1.3 (continued). $B_1 = \mathrm{conv}\{\pm e_1, \dots, \pm e_d\}$ has radius 1. By Maurey's empirical method,

$$\log N(B_1, \|\cdot\|_2, \varepsilon) \le \log \binom{2d + \lceil 1/\varepsilon^2\rceil - 2}{\lceil 1/\varepsilon^2\rceil - 1} = O\Big(\frac{1+\log(\varepsilon^2 d)}{\varepsilon^2}\Big), \qquad 1/\sqrt{d} \ll \varepsilon \ll 1.$$

9.1.5 More results without proof

(1) For $0 < p \leq q \leq \infty$, let $B_p := \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$. Then

$$\log N(B_p, \|\cdot\|_q, \varepsilon) \asymp_{p,q} \begin{cases} \varepsilon^{-\frac{pq}{q-p}}\big(\log(d\,\varepsilon^{-\frac{pq}{q-p}}) + 1\big), & d^{1/q-1/p} \leq \varepsilon \leq 1, \\ d\Big(\log \frac{1}{d^{1/q-1/p}\,\varepsilon} + 1\Big), & \varepsilon < d^{1/q-1/p}. \end{cases}$$

(2) Let $N(A, B)$ be the smallest number of translates of B that cover A. There exist universal constants $\alpha, \beta > 0$ such that for any symmetric convex body A,

$$\frac{1}{\beta} \log N\Big(B_2, \frac{\varepsilon}{\alpha} A^\circ\Big) \leq \log N(A, \varepsilon B_2) \leq \beta \log N\big(B_2, \alpha\varepsilon A^\circ\big),$$

where

$$A^\circ := \Big\{y : \sup_{x \in A} \langle x, y\rangle \leq 1\Big\}$$

is the polar body of A.

(3) Let $\mathcal{H}^s := \{f \in C^s([0,1]) : \big\|f^{(s)}\big\|_\infty \leq 1\}$. Then for any $1 \leq p \leq \infty$,

$$\log N(\mathcal{H}^s, \|\cdot\|_p, \varepsilon) \asymp_p \varepsilon^{-1/s}.$$

(4) Let $\mathcal{F}_m := \{f : [0,1] \to [0,1] : f \text{ is non-decreasing}\}$. Then for any $1 \leq p < \infty$,

$$\log N(\mathcal{F}_m, \|\cdot\|_p, \varepsilon) \asymp_p \frac{1}{\varepsilon}.$$

(5) Let $\mathcal{F}_c := \{f : [0,1] \to [0,1] : f \text{ is convex}\}$. Then for any $1 \leq p < \infty$,

$$\log N(\mathcal{F}_c, \|\cdot\|_p, \varepsilon) \asymp_p \frac{1}{\sqrt{\varepsilon}}.$$

9.2 Global Fano's method

9.2.1 Recall the steps of Fano

1. Find a pairwise separated set $\{\theta_0, \ldots, \theta_m\} \subseteq \Theta$ such that for all $i \neq j$,

$$\min_{a \in \mathcal{A}} \big[L(\theta_i, a) + L(\theta_j, a)\big] \geq \delta.$$

2. Upper bound $I(\theta; X)$ (or more often $I(\theta; X^n)$) with $\theta \sim \mathrm{Unif}\{\theta_0, \ldots, \theta_m\}$ and $X \mid \theta \sim P_\theta$.

3. If $I(\theta; X) \leq \frac{1}{2} \log m$, then the minimax risk satisfies $r^* = \Omega(\delta)$.

9.2.2 Step 0: packing via a metric

If there is a metric $d(\theta, \theta')$ such that

$$\min_{a \in \mathcal{A}} \big[L(\theta, a) + L(\theta', a)\big] \geq h\big(d(\theta, \theta')\big)$$

for an increasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$, then a δ-packing $\{\theta_0, \ldots, \theta_m\}$ of Θ under d satisfies the separation condition with

$$\Delta = h(\delta).$$

9.2.3 KL covering

Definition 9.8 (KL covering number)**.** For a family $\mathcal{P}$ of distributions and $\varepsilon > 0$, let $N_{\mathrm{KL}}(\mathcal{P}, \varepsilon)$ be the smallest integer n such that there exist distributions $Q_1, \ldots, Q_n$ (not necessarily in $\mathcal{P}$) satisfying

$$\sup_{P \in \mathcal{P}} \min_{i \in [n]} D_{\mathrm{KL}}(P \| Q_i) \le \varepsilon^2.$$

(Note: D_{KL} is not a metric; Q_i appears in the second argument.)

Theorem 9.9 (Entropic upper bound of $I(\theta; X^n)$)**.** *Let* $\theta \sim \pi$ *with* $\mathrm{supp}(\pi) = \Theta_0$*, and let* $X^n \mid \theta \sim P_\theta^{\otimes n}$*. Then*

$$I(\theta; X^n) \le \inf_{\varepsilon > 0} \Big(n\varepsilon^2 + \log N_{\mathrm{KL}}\big((P_\theta)_{\theta \in \Theta_0}, \varepsilon\big) \Big).$$

Proof. Recall the "golden formula" (Lecture 7):

$$I(\theta; X^n) = \min_{Q_{X^n}} \mathbb{E}_{\theta \sim \pi}\big[D_{\mathrm{KL}}(P_\theta^{\otimes n} \| Q_{X^n})\big].$$

Let $Q_1, \ldots, Q_N$ be an ε-covering of $(P_\theta)_{\theta \in \Theta_0}$ under KL, where $N = N_{\mathrm{KL}}((P_\theta)_{\theta \in \Theta_0}, \varepsilon)$. Choose

$$Q_{X^n} := \frac{1}{N} \sum_{i=1}^N Q_i^{\otimes n}.$$

Then for any $\theta \in \Theta_0$,

$$\begin{aligned}
D_{\mathrm{KL}}\Big(P_\theta^{\otimes n} \Big\| \frac{1}{N} \sum_{i=1}^N Q_i^{\otimes n}\Big) &= \mathbb{E}_{P_\theta^{\otimes n}}\left[\log \frac{P_\theta^{\otimes n}}{\frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}}\right] \\
&\le \mathbb{E}_{P_\theta^{\otimes n}}\left[\min_{i \in [N]} \log \frac{P_\theta^{\otimes n}}{Q_i^{\otimes n}} + \log N\right] \qquad \text{(since } \textstyle\sum x_i \ge \max x_i\text{)} \\
&\le \min_{i \in [N]} \mathbb{E}_{P_\theta^{\otimes n}}\left[\log \frac{P_\theta^{\otimes n}}{Q_i^{\otimes n}}\right] + \log N \\
&= \min_{i \in [N]} n\, D_{\mathrm{KL}}(P_\theta \| Q_i) + \log N \\
&\le n\varepsilon^2 + \log N.
\end{aligned}$$

Taking expectation over $\theta \sim \pi$ and then infimum over $\varepsilon > 0$ yields the result. □

9.2.4 A diagram: global Fano

For hyperparameters $\Theta_0 \subseteq \Theta$ and $\varepsilon, \delta > 0$:

(1) Find a metric d and a function h so that $\min_{a \in \mathcal{A}}[L(\theta, a) + L(\theta', a)] \ge h(d(\theta, \theta'))$, then take a δ-packing of Θ_0 under d.

(2) Find an ε-covering of $(P_\theta)_{\theta \in \Theta_0}$ under KL.

(3) Apply Fano to obtain

$$r^* \gtrsim \frac{h(\delta)}{2} \left(1 - \frac{\log N_{\mathrm{KL}}\big((P_\theta)_{\theta \in \Theta_0}, \varepsilon\big) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, d, \delta)} \right).$$

Optimize over $(\Theta_0, \delta, \varepsilon)$.

9.2.5 Example 2.1 (Gaussian location model)

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, I_d)$ with unknown $\theta \in \mathbb{R}^d$.

Target.

$$\inf_{\hat\theta} \sup_{\theta} \mathbb{E}_\theta \left[\left\| \hat\theta - \theta \right\|_p \right] \gtrsim_p \begin{cases} \dfrac{d^{1/p}}{\sqrt{n}}, & 2 \le p < \infty, \\ \sqrt{\dfrac{\log d}{n}}, & p = \infty. \end{cases}$$

Proof of the lower bound. Choose $\Theta_0 = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le r\}$. Then for any $\varepsilon, \delta > 0$, global Fano gives

$$\begin{aligned} r^* &\gtrsim \delta \left(1 - \frac{\log N_{\mathrm{KL}}\big(\{\mathcal{N}(\theta, I_d)\}_{\theta \in \Theta_0}, \varepsilon\big) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, \|\cdot\|_p, \delta)} \right) \\ &= \delta \left(1 - \frac{\log N\big(\Theta_0, \|\cdot\|_2, \sqrt{2}\,\varepsilon\big) + n\varepsilon^2 + \log 2}{\log M(\Theta_0, \|\cdot\|_p, \delta)} \right), \end{aligned}$$

since $D_{\mathrm{KL}}(\mathcal{N}(\theta, I_d) \| \mathcal{N}(\theta', I_d)) = \frac{1}{2}\|\theta - \theta'\|_2^2$.

Choice of ε. Choose $\varepsilon = r/\sqrt{2}$, so that $\log N(\Theta_0, \|\cdot\|_2, \sqrt{2}\,\varepsilon) = \log N(\Theta_0, \|\cdot\|_2, r) = \log 1 = 0$.

Choice of δ/r. For $p \in (2, \infty)$ choose $\delta/r = d^{1/p - 1/2}$, so that $\log M(\Theta_0, \|\cdot\|_p, \delta) \gtrsim d$. For $p = \infty$ choose $\delta/r \asymp 1$, so that $\log M(\Theta_0, \|\cdot\|_\infty, \delta) \gtrsim \log d$.

Choice of r. Now we have

$$r^* \gtrsim \begin{cases} r\, d^{1/p - 1/2} \left(1 - \dfrac{C_1 n r^2 + \log 2}{C_2 d} \right), & p \in (2, \infty), \\ r \left(1 - \dfrac{C_1 n r^2 + \log 2}{C_2 \log d} \right), & p = \infty. \end{cases}$$

Thus choosing $r = \sqrt{d/n}$ for $p \in (2, \infty)$ and $r = \sqrt{(\log d)/n}$ for $p = \infty$ yields

$$r^* \gtrsim \begin{cases} \dfrac{d^{1/p}}{\sqrt{n}}, & 2 < p < \infty, \\ \sqrt{\dfrac{\log d}{n}}, & p = \infty. \end{cases}$$

9.2.6 Example 2.2 (Nonparametric density estimation)

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f$ on $[0,1]$ with $\left\| f^{(s)} \right\|_\infty \le 1$ (i.e., the function space is $\mathcal{H}^s$).

Target. For $p \in [1, \infty)$,

$$\inf_{\hat f} \sup_{f \in \mathcal{F}} \mathbb{E}_f \left[\left\| \hat f - f \right\|_p \right] \gtrsim n^{-\frac{s}{2s+1}}.$$

Proof of the lower bound. Consider $\mathcal{H}_0^s \subseteq \mathcal{H}^s$ with

$$\mathcal{H}_0^s := \{f \in \mathcal{H}^s : f \geq 1/2 \text{ on } [0,1]\}.$$

Then for $f, g \in \mathcal{H}_0^s$,

$$D_{\mathrm{KL}}(f\|g) \leq \chi^2(f\|g) \leq 2\left\|f-g\right\|_2^2.$$

Hence

$$N_{\mathrm{KL}}(\mathcal{H}_0^s, \varepsilon) \leq N\big(\mathcal{H}_0^s, \|\cdot\|_2, \varepsilon/\sqrt{2}\big) \leq N\big(\mathcal{H}^s, \|\cdot\|_2, \varepsilon/\sqrt{2}\big).$$

By global Fano, for any $\varepsilon, \delta > 0$,

$$\begin{aligned} r^* &\gtrsim \delta\left(1 - \frac{\log N_{\mathrm{KL}}(\mathcal{H}_0^s, \varepsilon) + n\varepsilon^2 + \log 2}{\log M(\mathcal{H}_0^s, \|\cdot\|_p, \delta)}\right) \\ &\gtrsim \delta\left(1 - \frac{C_1\varepsilon^{-1/s} + n\varepsilon^2 + \log 2}{C_2\delta^{-1/s}}\right), \end{aligned}$$

by the metric entropy bounds for $\mathcal{H}_0^s$. Choosing $\varepsilon \asymp \delta \asymp n^{-s/(2s+1)}$ gives $r^* = \Omega\big(n^{-s/(2s+1)}\big)$.

9.2.7 Example 2.3 (Isotonic regression)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_X$, where P_X (known or unknown) has a bounded density on $[0,1]$. Conditioned on X^n, let $Y_i \overset{\text{ind}}{\sim} \mathcal{N}(f(X_i), 1)$ with $f \in \mathcal{F}_m = \{f : [0,1] \to [0,1] : f \text{ is increasing}\}$.

Target. For all $p \in [1, \infty)$,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_m} \mathbb{E}_f\Big[\Big\|\hat{f} - f\Big\|_p\Big] \gtrsim_p n^{-1/3}.$$

Proof of the lower bound. Since P_X has a bounded density,

$$D_{\mathrm{KL}}(P_f\|P_{f'}) = \frac{1}{2}\left\|f - f'\right\|_{L^2(P_X)}^2 = O(1)\,\left\|f - f'\right\|_2^2.$$

Therefore

$$N_{\mathrm{KL}}\big((P_f)_{f\in\mathcal{F}_m}, \varepsilon\big) \leq N\Big(\mathcal{F}_m, \|\cdot\|_2, \frac{\varepsilon}{O(1)}\Big).$$

By global Fano,

$$\begin{aligned} r^* &\gtrsim \delta\left(1 - \frac{\log N\big(\mathcal{F}_m, \|\cdot\|_2, \varepsilon/O(1)\big) + n\varepsilon^2 + \log 2}{\log M(\mathcal{F}_m, \|\cdot\|_p, \delta)}\right) \\ &\gtrsim \delta\left(1 - \frac{c_1/\varepsilon + n\varepsilon^2 + \log 2}{1/\delta}\right), \end{aligned}$$

using $\log N(\mathcal{F}_m, \|\cdot\|_p, \varepsilon) \asymp_p 1/\varepsilon$. Choosing $\varepsilon \asymp n^{-1/3}$ and $\delta \asymp n^{-1/3}$ yields $r^* = \Omega(n^{-1/3})$.

9.2.8 Example 2.4 (Convex regression)

Same setting as Example 2.3, but with $\mathcal{F}_m$ replaced by

$$\mathcal{F}_c := \{f : [0,1] \to [0,1] : f \text{ is convex}\}.$$

Target. For $p \in [1, \infty)$,

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_c} \mathbb{E}_f \left[\left\| \hat{f} - f \right\|_p \right] \gtrsim_p n^{-2/5}.$$

Proof sketch. Similar to Example 2.3, now with $\log N(\mathcal{F}_c, \|\cdot\|_p, \varepsilon) \asymp_p 1/\sqrt{\varepsilon}$.

9.2.9 Example 2.5 (Sparse linear regression)

Let $Y \sim \mathcal{N}(X\theta, I_n)$ with fixed design $X \in \mathbb{R}^{n \times d}$, where all singular values of X are $O(\sqrt{n})$. The unknown parameter $\theta \in \mathbb{R}^d$ is sparse in the sense that

$$\|\theta\|_q \le R \qquad \text{for some } q \in (0,1).$$

Target. For small enough $R < f(n,d)$,

$$\inf_{\hat{\theta}} \sup_{\|\theta\|_q \le R} \mathbb{E}_\theta \left[\left\| \hat{\theta} - \theta \right\|_p \right] \ \gtrsim_{p,q} \ R^{q/p} \left(\frac{\log d}{n} \right)^{\frac{p-q}{2p}}.$$

Proof of the lower bound. **1. ℓ_p-packing of $B_q(R)$.** Let $B_q(R) := \{\theta \in \mathbb{R}^d : \|\theta\|_q \le R\}$. Then

$$\log M\big(B_q(R), \|\cdot\|_p, \delta\big) \ \gtrsim \ \left(\frac{R}{\delta}\right)^{\frac{pq}{p-q}} \log d \qquad \text{if } \ \delta \gg R\, d^{1/p - 1/q}.$$

2. KL covering of $\mathcal{P} := \{\mathcal{N}(X\theta, I_n) : \|\theta\|_q \le R\}$. For θ, θ',

$$D_{\mathrm{KL}}\big(\mathcal{N}(X\theta, I_n) \| \mathcal{N}(X\theta', I_n)\big) = \frac{1}{2} \left\| X(\theta - \theta') \right\|_2^2 = O(n) \ \left\| \theta - \theta' \right\|_2^2.$$

Hence

$$\begin{aligned} \log N_{\mathrm{KL}}(\mathcal{P}, \varepsilon) &\le \log N\Big(B_q(R), \|\cdot\|_2, \frac{\varepsilon}{O(\sqrt{n})}\Big) \\ &\lesssim \left(\frac{\sqrt{n}\, R}{\varepsilon}\right)^{\frac{2q}{2-q}} \log d \qquad \text{if } \ \varepsilon \gg R\sqrt{n}\, d^{1/2 - 1/q}. \end{aligned}$$

Now choose

$$\varepsilon \asymp n^{q/4} R^{q/2} (\log d)^{(2-q)/4}, \qquad \delta = R^{q/p} \left(\frac{\log d}{n}\right)^{\frac{p-q}{2p}}.$$

Then

$$\log M(\delta) \gtrsim R^q n^{q/2} (\log d)^{1-q/2}, \qquad \log N_{\mathrm{KL}}(\varepsilon) \lesssim \varepsilon^2 \lesssim R^q n^{q/2} (\log d)^{1-q/2},$$

and global Fano gives the stated lower bound.

9.3 Special topic: generalized Fano with χ^2-informativity

Since the proof of Fano is simply DPI, replacing KL by other f-divergences also leads to meaningful Bayes risk lower bounds.

Theorem 9.10 (Generalized Fano with χ^2-informativity). *For $\theta \sim \pi$, it holds that*

$$\mathbb{P}\big(L(\theta, X) \geq 0\big) \geq 1 - p_0 - \sqrt{p_0\, I_{\chi^2}(\theta; X)},$$

where

$$p_0 := \sup_a \pi\big(L(\theta, a) < 0\big) \quad \textit{is the small-ball probability,}$$

and

$$I_{\chi^2}(\theta; X) := \inf_{Q_X} \chi^2\big(P_{\theta X} \| \pi_\theta Q_X\big) = \inf_{Q_X} \mathbb{E}_{\theta\sim\pi}\Big[\chi^2\big(P_{X|\theta} \| Q_X\big)\Big]$$

is the χ^2-informativity.

Proof. Apply DPI to the mapping $(\theta, X) \mapsto \mathbb{1}\{L(\theta, X) \geq 0\}$:

$$P_{\theta X} \;\longrightarrow\; \mathbb{1}\{L(\theta, X) \geq 0\} \quad \text{and} \quad \pi_\theta Q_X \;\longrightarrow\; \mathbb{1}\{L(\theta, X) \geq 0\}.$$

This yields Bernoulli distributions $\mathrm{Bern}(\mathbb{P}(L(\theta, X) \geq 0))$ and $\mathrm{Bern}(\geq 1 - p_0)$. Hence

$$\begin{aligned}\chi^2(P_{\theta X} \| \pi_\theta Q_X) &\geq \chi^2\Big(\mathrm{Bern}(\mathbb{P}(L(\theta, X) \geq 0)) \;\Big\|\; \mathrm{Bern}(\geq 1 - p_0)\Big) \\ &\geq \frac{\big(\mathbb{P}(L(\theta, X) \geq 0) - (1 - p_0)\big)^2}{p_0(1 - p_0)} \qquad \text{if } \mathbb{P}(L(\theta, X) \geq 0) \leq 1 - p_0.\end{aligned}$$

Taking the infimum over Q_X and rearranging gives the stated inequality. □

Similarly, we have an entropic upper bound of $I_{\chi^2}(\theta; X)$ based on χ^2-covering.

Theorem 9.11 (Entropic upper bound for χ^2-informativity). *Let $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$ and suppose* $\mathrm{supp}(\pi) \subseteq \Theta$*. Then for $\theta \sim \pi$,*

$$I_{\chi^2}(\theta; X) + 1 \leq \inf_{\varepsilon>0} (1 + \varepsilon^2)\, N_{\chi^2}(\mathcal{P}, \varepsilon),$$

where

$$N_{\chi^2}(\mathcal{P}, \varepsilon) := \min\left\{ n : \min_{Q_1,\dots,Q_n} \sup_{P\in\mathcal{P}} \min_{i\in[n]} \chi^2(P \| Q_i) \leq \varepsilon^2 \right\}.$$

Proof. **Exercise.** □

9.3.1 Example 3.1 (Gaussian model with uniform prior)

Let $X \sim \mathcal{N}(\theta, I_d)$ with $\theta \sim \mathrm{Unif}(B_2(R))$ (denote this prior by π).

Target.

$$r_\pi^* := \inf_{\hat{\theta}} \mathbb{E}_{\theta\sim\pi}\Big[\Big\|\hat{\theta} - \theta\Big\|_2^2\Big] \;\gtrsim\; d \qquad \text{if } R = \Omega(\sqrt{d}).$$

Failure of mutual information. For $\Delta \in (0, R)$, the small-ball probability is

$$p_\Delta = \sup_a \pi\big(\|\theta - a\|_2^2 \le \Delta^2\big) = \left(\frac{\Delta}{R}\right)^d.$$

For mutual information, the entropic upper bound gives

$$\begin{aligned} I(\theta; X) &\le \inf_{\varepsilon>0}\Big(\log N(B_2(R), \|\cdot\|_2, \varepsilon) + \varepsilon^2\Big) \\ &\le \inf_{\varepsilon>0}\Big(d \log \frac{3R}{\varepsilon} + \varepsilon^2\Big) \;\sim\; d \log \frac{R}{\sqrt{d}}, \qquad R \gg \sqrt{d}. \end{aligned}$$

Therefore Fano gives

$$r_\pi^* \;\gtrsim\; \sup_{\Delta>0} \Delta^2 \left(1 - \frac{d \log \frac{R}{\sqrt{d}} + \log 2}{d \log \frac{R}{\Delta}}\right).$$

Usually one matches

$$d \log \frac{R}{\sqrt{d}} = (1-p)\, d \log \frac{R}{\Delta} \qquad \Longrightarrow \qquad \Delta = d^{\frac{1}{2(1-p)}} R^{-\frac{p}{1-p}}, \text{ for some constant } p > 0.$$

Thus

$$r_\pi^* = \Omega\Big(d^{\frac{1}{1-p}} R^{-\frac{2p}{1-p}}\Big) = \Omega\Big(d\,(d/R^2)^{\frac{p}{1-p}}\Big),$$

which is weaker than $\Omega(d)$.

Proof using χ^2-informativity. The entropic upper bound gives

$$I_{\chi^2}(\theta; X) + 1 \le \inf_{\varepsilon>0} (1+\varepsilon^2)\, N\Big(B_2(R), \|\cdot\|_2, \sqrt{\log(1+\varepsilon^2)}\Big),$$

since $\chi^2(\mathcal{N}(\theta, I_d) \| \mathcal{N}(\theta', I_d)) = e^{\|\theta-\theta'\|_2^2} - 1$. Using the crude covering bound $N(B_2(R), \|\cdot\|_2, \eta) \le (3R/\eta)^d$,

$$I_{\chi^2}(\theta; X) + 1 \le \inf_{\varepsilon>0} (1+\varepsilon^2) \left(\frac{3R}{\sqrt{\log(1+\varepsilon^2)}}\right)^d = \exp\Big(d \log \frac{O(1)R}{\sqrt{d}}\Big), \qquad R > C\sqrt{d},$$

by choosing $1 + \varepsilon^2 = e^d$.

Therefore, generalized Fano gives

$$r_\pi^* \;\gtrsim\; \sup_{\Delta>0} \Delta^2 \left(1 - \left(\frac{\Delta}{R}\right)^d - \sqrt{\left(\frac{\Delta}{R}\right)^d \cdot \exp\Big(d \log \frac{O(1)R}{\sqrt{d}}\Big)}\right).$$

The underbraced term can be made $\le 1/2$ by taking $\Delta = c\sqrt{d}$ for a small constant c. Hence $r_\pi^* = \Omega(d)$.

9.3.2 Example 3.2: ridge bandits

9.3.3 Setup and target

Model. $r_t \sim \mathcal{N}\big(f(\langle \theta^*, a_t \rangle), 1\big)$ for $\theta^* \sim \mathrm{Unif}(\mathbb{S}^{d-1})$. Here $f : [-1, 1] \to \mathbb{R}$ is a known increasing link function with $f(0) = 0$. Define

$$g(x) := \max\{|f(x)|, |f(-x)|\}.$$

Target statement. Define a recursive sequence with a large constant $C > 0$:

$$\varepsilon_1 = C\sqrt{\frac{\log(1/\delta)}{d}}, \qquad \varepsilon_{t+1}^2 = \varepsilon_t^2 + \frac{C}{d} g(\varepsilon_t)^2.$$

Then for any interactive learner,

$$\mathbb{P}\Big(|\langle \theta^*, a_s\rangle| \le \varepsilon_s \text{ for all } 1 \le s \le t\Big) \ge 1 - t\delta.$$

Remarks.

(1) The sequence $\{\varepsilon_t\}$ is a *pointwise upper bound* on the learning trajectory of any algorithm.

(2) The growth $\varepsilon_{t+1}^2 - \varepsilon_t^2$ increases with t: interactive learning becomes faster and faster.

9.3.4 Intuition: mutual information is not strong enough

Let $I_t = I(H_t; \theta^*) := I(a^t, r^t; \theta^*)$. Then

$$\begin{aligned} I_{t+1} - I_t &= I\big(\theta^*; r_{t+1} \mid H_t, a_{t+1}\big) \\ &\le \mathbb{E}\Big[D_{\mathrm{KL}}\big(\mathcal{N}(f(\langle \theta^*, a_{t+1}\rangle), 1) \| \mathcal{N}(0,1)\big)\Big] \qquad \text{(golden formula)} \\ &= \frac{1}{2}\mathbb{E}\big[f(\langle \theta^*, a_{t+1}\rangle)^2\big]. \end{aligned}$$

We aim to upper bound this information gain. A key observation is that

$$I(\theta^*; a_{t+1}) \le I(\theta^*; H_t) = I_t,$$

so a_{t+1} is "constrained" in information and we expect $\langle \theta^*, a_{t+1}\rangle$ to be small.

The tempting (but false) intuition is:

$$I(\theta^*; a) \le d\varepsilon^2 \quad \Longrightarrow \quad |\langle \theta^*, a\rangle| \le \varepsilon \text{ w.h.p.} \qquad (*)$$

If $(*)$ were true, we would get the recursion by the correspondence $I_t \lesssim d\varepsilon_t^2$.

However, mutual information is not strong enough to ensure $(*)$: Fano only gives

$$\mathbb{P}\big(|\langle \theta^*, a\rangle| \le \varepsilon\big) \ge 1 - \frac{I(\theta^*; a) + \log 2}{c\, d\varepsilon^2},$$

which is not small enough to apply a union bound!

9.3.5 Proof using χ^2-informativity

Let

$$E_t := \bigcap_{s=1}^{t} \Big\{ |\langle \theta^*, a_s\rangle| \le \varepsilon_s \Big\}.$$

Define a slight variant of χ^2-informativity:

$$I_{\chi^2}(X; Y \mid E) := \inf_{Q_Y} \chi^2\big(P_{XY|E} \| P_{X|E} Q_Y\big).$$

Then we can still get

$$\mathbb{P}\big(|\langle \theta^*, a\rangle| \le \varepsilon \mid E\big) \ge 1 - c_1 e^{-c_0 d\varepsilon^2} \sqrt{I_{\chi^2}(\theta^*; a \mid E) + 1}.$$

(For fixed a, $\mathbb{P}(|\langle \theta^*, a\rangle| \le \varepsilon) \le e^{-c_0 d\varepsilon^2}$.)

Key recursion. The crux is to establish:

$$I_{\chi^2}(\theta^*; H_t \mid E_t) + 1 \le \frac{e^{g(\varepsilon_t)^2}}{\mathbb{P}(E_t \mid E_{t-1})^2}\Big(I_{\chi^2}(\theta^*; H_{t-1} \mid E_{t-1}) + 1\Big). \qquad (*)$$

If $(*)$ holds, then

$$\begin{aligned} I_{\chi^2}(\theta^*; H_t \mid E_t) + 1 &\le \prod_{s=1}^{t} \frac{e^{g(\varepsilon_s)^2}}{\mathbb{P}(E_s \mid E_{s-1})^2} \\ &= \frac{1}{\mathbb{P}(E_t)^2} \exp\left(\sum_{s\le t} g(\varepsilon_s)^2\right). \end{aligned}$$

Therefore

$$\begin{aligned} \mathbb{P}(E_{t+1} \mid E_t) &\ge 1 - c_1 e^{-c_0 d \varepsilon_{t+1}^2} \sqrt{I_{\chi^2}(\theta^*; a_{t+1} \mid E_t) + 1} \\ &\ge 1 - c_1 e^{-c_0 d \varepsilon_{t+1}^2} \sqrt{I_{\chi^2}(\theta^*; H_t \mid E_t) + 1} \qquad \text{(DPI)} \\ &\ge 1 - \frac{c_1}{\mathbb{P}(E_t)} \exp\left(-c_0 d \varepsilon_{t+1}^2 + \frac{1}{2}\sum_{s\le t} g(\varepsilon_s)^2\right). \end{aligned}$$

The recursion ensures that the exponent is $\le -c_0 d\varepsilon_1^2 \le -c' \log(1/\delta)$. Consequently,

$$\mathbb{P}(E_{t+1}) = \mathbb{P}(E_t)\,\mathbb{P}(E_{t+1} \mid E_t) \ge \mathbb{P}(E_t) - \delta.$$

Iterating yields $\mathbb{P}(E_t) \ge 1 - t\delta$.

Proof of $(*)$.

$$\begin{aligned} &I_{\chi^2}(\theta^*; H_t \mid E_t) + 1 \\ &= \inf_{Q_{H_t}} \int \frac{\mathbb{P}(\theta^*, H_t \mid E_t)^2}{\pi(\theta^*)\, Q_{H_t}(H_t)}\, d\theta^*\, da^t\, dr^t \\ &\le \inf_{Q_{H_{t-1}}} \int \frac{\Big[\frac{\mathbb{1}(E_t)}{\mathbb{P}(E_t)} \pi(\theta^*) \prod_{s=1}^{t} p_s(a_s \mid H_{s-1})\, \varphi(r_s - f(\langle \theta^*, a_s\rangle))\Big]^2}{\pi(\theta^*)\, Q_{H_{t-1}}(H_{t-1})\, p_t(a_t \mid H_{t-1})\, \varphi(r_t)}\, d\theta^*\, da^t\, dr^t \\ &= \inf_{Q_{H_{t-1}}} \int \frac{\Big[\frac{\mathbb{1}(E_t)}{\mathbb{P}(E_t)} \pi(\theta^*) \prod_{s=1}^{t-1} p_s(a_s \mid H_{s-1})\, \varphi(r_s - f(\langle \theta^*, a_s\rangle))\Big]^2}{\pi(\theta^*)\, Q_{H_{t-1}}(H_{t-1})}\, p_t(a_t \mid H_{t-1})\, e^{f(\langle \theta^*, a_t\rangle)^2}\, d\theta^*\, da^t\, dr^{t-1}. \end{aligned}$$

Here we used that the last-step likelihood ratio contributes a factor $e^{f(\langle \theta^*, a_t\rangle)^2}$, and on E_t we have $f(\langle \theta^*, a_t\rangle)^2 \le g(\varepsilon_t)^2$. Also,

$$\frac{\mathbb{1}(E_t)}{\mathbb{P}(E_t)} \le \frac{\mathbb{1}(E_{t-1})}{\mathbb{P}(E_{t-1})} \cdot \frac{1}{\mathbb{P}(E_t \mid E_{t-1})}.$$

Therefore

$$I_{\chi^2}(\theta^*; H_t \mid E_t) + 1 \le \frac{e^{g(\varepsilon_t)^2}}{\mathbb{P}(E_t \mid E_{t-1})^2}\Big(I_{\chi^2}(\theta^*; H_{t-1} \mid E_{t-1}) + 1\Big),$$

which is $(*)$.

Lecture 10: Entropic upper bounds of density estimation

10.1 Setup and overview

Last lecture: use covering/packing to prove statistical lower bounds via Fano.

This lecture: they can also prove *upper* bounds.

Density estimation. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$, where $P \in \mathcal{P}$ is an unknown distribution. The target is: for a divergence/distance $D \in \{D_{\text{KL}}, \text{TV}, H^2\}$, construct an estimator $\widehat{P} = \widehat{P}(X^n)$ such that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{X^n\sim P^{\otimes n}}\big[D(P,\widehat{P})\big] \quad \text{is small.}$$

Overview of results

- **KL (Yang–Barron).** There exists $\widehat{P}$ such that
$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\big[D_{\text{KL}}(P\|\widehat{P})\big] \lesssim \inf_{\varepsilon>0}\Big(\varepsilon^2 + \frac{1}{n}\log N_{\text{KL}}(\mathcal{P},\varepsilon)\Big).$$
- **TV (Yatracos).** There exists $\widehat{P}$ such that
$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\big[\text{TV}^2(P,\widehat{P})\big] \lesssim \inf_{\varepsilon>0}\Big(\varepsilon^2 + \frac{1}{n}\log N_{\text{TV}}(\mathcal{P},\varepsilon)\Big).$$
- **Hellinger (Le Cam–Birgé).** There exists $\widehat{P}$ such that
$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\big[H^2(P,\widehat{P})\big] \lesssim \inf_{\varepsilon>0}\Big(\varepsilon^2 + \frac{1}{n}\log N_{H}(\mathcal{P},\varepsilon)\Big).$$

Examples.

1. For finite-dimensional models $\mathcal{P}$ with d parameters, usually
$$\log N_D(\mathcal{P},\varepsilon) \simeq d\log\frac{1}{\varepsilon} \qquad \text{(volume bound).}$$
In this case,
$$\inf_{\widehat{P}}\sup_{P\in\mathcal{P}} \mathbb{E}_P\big[D(P,\widehat{P})\big] \lesssim \inf_{\varepsilon>0}\Big(\varepsilon^2 + \frac{d}{n}\log\frac{1}{\varepsilon}\Big) \lesssim \frac{d\log n}{n},$$
usually optimal up to a $\log n$ factor.

2. For nonparametric classes $\mathcal{P}$ with

$$\log N_D(\mathcal{P}, \varepsilon) \simeq \varepsilon^{-d},$$

we have

$$\inf_{\widehat{P}} \sup_{P\in\mathcal{P}} \mathbb{E}_P\big[D(P, \widehat{P})\big] \;\lesssim\; \inf_{\varepsilon>0} \Big(\varepsilon^2 + \frac{1}{n\varepsilon^d}\Big) \;\lesssim\; n^{-\frac{2}{2+d}}.$$

10.2 Yang–Barron: progressive mixing / online-to-batch conversion

10.2.1 An "online" guarantee

Similar to global Fano, let $P_1, \ldots, P_N$ be an ε-covering of $\mathcal{P}$, i.e.

$$\sup_{P\in\mathcal{P}} \min_{i\in[N]} D_{\mathrm{KL}}(P\|P_i) \le \varepsilon^2, \qquad [N] \coloneqq \{1, 2, \ldots, N\}.$$

Let $Q_{X^{n+1}}$ be the *average product distribution*:

$$Q_{X^{n+1}} \;\coloneqq\; \frac{1}{N}\sum_{i=1}^{N} P_i^{\otimes(n+1)}.$$

Lemma 10.1.

$$\sup_{P\in\mathcal{P}} D_{\mathrm{KL}}\big(P^{\otimes(n+1)} \,\|\, Q_{X^{n+1}}\big) \;\le\; (n+1)\varepsilon^2 + \log N.$$

Proof. Similar to global Fano: for any $P \in \mathcal{P}$,

$$\begin{aligned} D_{\mathrm{KL}}\big(P^{\otimes(n+1)} \,\|\, Q_{X^{n+1}}\big) &= \mathbb{E}_{X^{n+1}\sim P^{\otimes(n+1)}}\left[\log \frac{P^{\otimes(n+1)}(X^{n+1})}{\frac{1}{N}\sum_{i=1}^{N} P_i^{\otimes(n+1)}(X^{n+1})}\right] \\ &\le \mathbb{E}_{X^{n+1}\sim P^{\otimes(n+1)}}\left[\min_{i\in[N]} \log \frac{P^{\otimes(n+1)}(X^{n+1})}{P_i^{\otimes(n+1)}(X^{n+1})} + \log N\right] \\ &\le \min_{i\in[N]} D_{\mathrm{KL}}\big(P^{\otimes(n+1)} \,\|\, P_i^{\otimes(n+1)}\big) + \log N \\ &\le (n+1)\varepsilon^2 + \log N. \end{aligned}$$

□

This is called an *online* guarantee as it concerns the density estimation performance for joint distributions of $X_1, \ldots, X_{n+1} \overset{\text{iid}}{\sim} P$.

10.2.2 Online-to-batch conversion

Given $Q_{X^{n+1}}$, we can define

$$\widehat{P}(x) \;\coloneqq\; \frac{1}{n+1}\sum_{t=0}^{n} Q_{X_{t+1}=x|X^t}.$$

Note that $\widehat{P}$ is a well-defined estimator and depends on X^n. Expanding out the definition of $Q_{X^{n+1}}$ gives the *progressive mixing* form

$$\widehat{P}(x) = \frac{1}{n+1}\sum_{t=0}^{n} \frac{\frac{1}{N}\sum_{i=1}^{N}\left(\prod_{s\le t} P_i(X_s)\right)P_i(x)}{\frac{1}{N}\sum_{i=1}^{N}\prod_{s\le t} P_i(X_s)} \in \mathrm{conv}(\mathcal{P}).$$

The Yang–Barron result follows from the next lemma.

Lemma 10.2.

$$\mathbb{E}_P\big[D_{D_{\mathrm{KL}}}(P\|\widehat{P})\big] \;\le\; \frac{1}{n+1}D_{D_{\mathrm{KL}}}\big(P^{\otimes(n+1)} \,\|\, Q_{X^{n+1}}\big).$$

Proof.

$$\begin{aligned}\mathbb{E}_P\big[D_{D_{\mathrm{KL}}}(P\|\widehat{P})\big] &= \mathbb{E}_P\left[D_{D_{\mathrm{KL}}}\left(P\,\middle\|\,\frac{1}{n+1}\sum_{t=0}^{n}Q_{X_{n+1}|X^t}\right)\right]\\ &\le \frac{1}{n+1}\sum_{t=0}^{n}\mathbb{E}_P\big[D_{D_{\mathrm{KL}}}(P\|Q_{X_{n+1}|X^t})\big] \qquad \text{(convexity)}\\ &= \frac{1}{n+1}D_{D_{\mathrm{KL}}}\big(P^{\otimes(n+1)} \,\|\, Q_{X^{n+1}}\big) \qquad \text{(chain rule).}\end{aligned}$$

□

Remark 10.3.

1. This online-to-batch conversion provides a general paradigm for converting "redundancy" bounds to prediction risk bounds, even beyond i.i.d. data (see more next lecture).
2. The Yang–Barron estimator is often *improper* (i.e. $\widehat{P} \in \mathrm{conv}(\mathcal{P})$ but often $\widehat{P} \notin \mathcal{P}$), and computationally hard to obtain.

10.3 Yatracos: minimum distance estimator for TV

The TV density estimation result is a corollary of the following general result in the robust case.

Theorem 10.4. *Let $X_1,\dots,X_n \overset{\text{iid}}{\sim} P$, and let $Q_1,\dots,Q_N$ be arbitrary candidate distributions. Then there exists an estimator $\widehat{P}$ such that*

$$\mathrm{TV}(P,\widehat{P}) \le 3\min_{i\in[N]}\mathrm{TV}(P,Q_i) + \varepsilon_n, \qquad \textit{with} \quad \mathbb{E}[\varepsilon_n^2] = O\Big(\frac{\log N}{n}\Big).$$

10.3.1 Proof via a minimum-distance estimator

We prove the theorem using a minimum-distance estimator:

$$\widehat{P} \;=\; \operatorname*{arg\,min}_{Q\in\{Q_1,\dots,Q_N\}} \widetilde{\mathrm{TV}}(P_n,Q),$$

where

$$P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

is the empirical distribution, and $\widetilde{\mathrm{TV}}$ is a pseudo-distance.

(What if $\widetilde{\mathrm{TV}} = \mathrm{TV}$? If $Q_1, \dots, Q_N$ are all continuous distributions, since P_n is discrete, $\mathrm{TV}(P_n, Q_i) = 1$ for all i, so it's not useful.)

Let us defer the choice of $\widetilde{\mathrm{TV}}$ and proceed to the analysis. Let

$$Q^* = \operatorname*{arg\,min}_{Q \in \{Q_1, \dots, Q_N\}} \mathrm{TV}(P, Q).$$

Then

$$\begin{aligned} \mathrm{TV}(\widehat{P}, P) &\le \mathrm{TV}(\widehat{P}, Q^*) + \mathrm{TV}(Q^*, P) \\ &\overset{\text{hope}}{\le} \widetilde{\mathrm{TV}}(\widehat{P}, Q^*) + \mathrm{TV}(Q^*, P) \\ &\le \widetilde{\mathrm{TV}}(\widehat{P}, P_n) + \widetilde{\mathrm{TV}}(P_n, Q^*) + \mathrm{TV}(Q^*, P) \\ &\le 2\widetilde{\mathrm{TV}}(P_n, Q^*) + \mathrm{TV}(Q^*, P) \qquad (\text{definition of } \widehat{P}) \\ &\le 2\widetilde{\mathrm{TV}}(P_n, P) + 2\widetilde{\mathrm{TV}}(P, Q^*) + \mathrm{TV}(P, Q^*) \\ &\overset{\text{hope}}{\le} 2\widetilde{\mathrm{TV}}(P_n, P) + 3\mathrm{TV}(P, Q^*). \end{aligned}$$

To make the analysis go through, we need:

1. $\widetilde{\mathrm{TV}}(P, Q) \le \mathrm{TV}(P, Q)$ for all P, Q.
2. $\widetilde{\mathrm{TV}}(Q_i, Q_j) = \mathrm{TV}(Q_i, Q_j)$ for all $i, j \in [N]$.
3. $\mathbb{E}\big[\widetilde{\mathrm{TV}}(P_n, P)^2\big]$ is small.

Motivated by (1)+(2), define

$$\widetilde{\mathrm{TV}}(P, Q) \;\coloneqq\; \sup_{A \in \mathcal{A}} \big|P(A) - Q(A)\big|,$$

where $\mathcal{A} = \{A_{ij} : i, j \in [N]\}$ with

$$A_{ij} \;\coloneqq\; \{x : q_i(x) \ge q_j(x)\}.$$

Verification of (1)–(3).

1. (1) is immediate since $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$.
2. (2) is also true since

$$\mathrm{TV}(Q_i, Q_j) = \big|Q_i(A_{ij}) - Q_j(A_{ij})\big| \le \widetilde{\mathrm{TV}}(Q_i, Q_j).$$

3. (3) Note that $|\mathcal{A}| \le \binom{N}{2}$, and for fixed A,

$$\mathbb{P}\big(\,|P(A) - P_n(A)| > \varepsilon\,\big) \le 2\exp(-2n\varepsilon^2) \qquad (\text{Hoeffding}).$$

Therefore, a union bound over A gives

$$\mathbb{P}\big(\widetilde{\mathrm{TV}}(P, P_n) > \varepsilon\big) \le 2N^2 \exp(-2n\varepsilon^2).$$

Consequently,

$$\begin{aligned}\mathbb{E}\big[\widetilde{\mathrm{TV}}^2(P,P_n)\big] &= \int_0^\infty \mathbb{P}\big(\widetilde{\mathrm{TV}}^2(P,P_n)\geq r\big)\ \mathrm{d}r\\ &\leq \int_0^\infty \min\{1,\ 2N^2e^{-2nr}\}\ \mathrm{d}r\\ &\leq \frac{\log(2N^2)}{2n} + \int_{\log(2N^2)/(2n)}^\infty 2N^2e^{-2nr}\ \mathrm{d}r\\ &= O\Big(\frac{\log N}{n}\Big).\end{aligned}$$

Remark 10.5.

1. The Yatracos estimator is *proper*, i.e. $\widehat{P}\in\mathcal{P}$.
2. The above proof also yields a high-probability guarantee on $\mathrm{TV}(\widehat{P},P)$.
3. It is known that the constant 3 is not improvable if the estimator is required to be proper.
4. There are some recent interests in computationally efficient versions of Yatracos.

10.4 Le Cam–Birgé: pairwise comparison

10.4.1 Composite hypothesis testing

- H_0: $X_1,\dots,X_n\sim P$ with $P\in\mathcal{P}$.
- H_1: $X_1,\dots,X_n\sim Q$ with $Q\in\mathcal{Q}$.
- Test: $T=T(X^n)\in\{0,1\}$.
- Type-I error: $\sup\limits_{P\in\mathcal{P}} P^{\otimes n}(T=1)$.
- Type-II error: $\sup\limits_{Q\in\mathcal{Q}} Q^{\otimes n}(T=0)$.

10.4.2 A testing lemma in terms of Hellinger distance

We use the convention

$$H^2(P,Q) := \int(\sqrt{p}-\sqrt{q})^2\ \mathrm{d}\mu, \qquad H(P,Q) := \sqrt{H^2(P,Q)},$$

so that

$$1-\frac{H^2(P,Q)}{2} = 1-\frac{1}{2}\int(\sqrt{p}-\sqrt{q})^2\ \mathrm{d}\mu = \int\sqrt{pq}\ \mathrm{d}\mu.$$

Lemma 10.6.

$$\inf_T\Big(\sup_{P\in\mathcal{P}}P^{\otimes n}(T=1)+\sup_{Q\in\mathcal{Q}}Q^{\otimes n}(T=0)\Big)\ \leq\ \exp\Big(-\frac{n}{2}H^2\big(\mathrm{conv}(\mathcal{P}),\mathrm{conv}(\mathcal{Q})\big)\Big),$$

where

$$H^2\big(\mathrm{conv}(\mathcal{P}),\mathrm{conv}(\mathcal{Q})\big)\ :=\ \inf_{P\in\mathrm{conv}(\mathcal{P})}\ \inf_{Q\in\mathrm{conv}(\mathcal{Q})} H^2(P,Q).$$

Proof. In Lecture 8 we know that

$$\text{LHS} = 1 - \mathrm{TV}\big(\mathrm{conv}(\mathcal{P}^{\otimes n}), \mathrm{conv}(\mathcal{Q}^{\otimes n})\big),$$

where $\mathcal{P}^{\otimes n} := \{P^{\otimes n} : P \in \mathcal{P}\}$ and $\mathrm{TV}(\mathcal{P}, \mathcal{Q}) := \inf_{P \in \mathcal{P}, Q \in \mathcal{Q}} \mathrm{TV}(P, Q)$. Moreover $\mathrm{TV}(P, Q) \geq \frac{1}{2}H^2(P, Q)$, hence

$$\begin{aligned}
\text{LHS} &\leq 1 - \frac{1}{2}H^2\big(\mathrm{conv}(\mathcal{P}^{\otimes n}), \mathrm{conv}(\mathcal{Q}^{\otimes n})\big) \\
&\leq \Big(1 - \frac{1}{2}H^2\big(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q})\big)\Big)^n \qquad \text{(next lemma)} \\
&\leq \exp\Big(-\frac{n}{2}H^2\big(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q})\big)\Big).
\end{aligned}$$

□

Lemma 10.7.

$$1 - \frac{1}{2}H^2\Big(\mathrm{conv}\Big(\bigotimes_{i=1}^n \mathcal{P}_i\Big),\ \mathrm{conv}\Big(\bigotimes_{i=1}^n \mathcal{Q}_i\Big)\Big) \ \leq\ \prod_{i=1}^n \Big(1 - \frac{1}{2}H^2\big(\mathrm{conv}(\mathcal{P}_i), \mathrm{conv}(\mathcal{Q}_i)\big)\Big).$$

Proof. Suffices to prove the case $n = 2$. Note that

$$1 - \frac{1}{2}H^2(P, Q) = \int \sqrt{pq}.$$

Any $P_{XY} \in \mathrm{conv}(\mathcal{P}_1 \otimes \mathcal{P}_2)$ can be written as $P_{XY} = \mathbb{E}_Z\big[P_{X|Z}P_{Y|Z}\big]$ with $P_{X|Z} \in \mathcal{P}_1$ and $P_{Y|Z} \in \mathcal{P}_2$. Then

$$\begin{aligned}
1 - \frac{1}{2}H^2(P_{XY}, Q_{XY}) &= \int \sqrt{p_{XY}q_{XY}} \\
&= \int_x \sqrt{p_X q_X} \int_y \sqrt{p_{Y|X}q_{Y|X}} \\
&= \int_x \sqrt{p_X q_X}\ \Big(1 - \frac{1}{2}H^2(P_{Y|X}, Q_{Y|X})\Big) \\
&\leq \int_x \sqrt{p_X q_X}\ \Big(1 - \frac{1}{2}H^2\big(\mathrm{conv}(\mathcal{P}_2), \mathrm{conv}(\mathcal{Q}_2)\big)\Big) \\
&\qquad \text{since } \mathbb{E}_{Z|X}[P_{Y|Z}] \in \mathrm{conv}(\mathcal{P}_2) \\
&\leq \Big(1 - \frac{1}{2}H^2\big(\mathrm{conv}(\mathcal{P}_1), \mathrm{conv}(\mathcal{Q}_1)\big)\Big)\Big(1 - \frac{1}{2}H^2\big(\mathrm{conv}(\mathcal{P}_2), \mathrm{conv}(\mathcal{Q}_2)\big)\Big) \\
&\qquad \text{since } \mathbb{E}_Z[P_{X|Z}] \in \mathrm{conv}(\mathcal{P}_1).
\end{aligned}$$

This proves the $n = 2$ case; the general n follows by induction. □

Remark 10.8. The same proof holds for all Rényi divergences

$$D_\alpha = \frac{1}{\alpha - 1}\log \int p^\alpha q^{1-\alpha}.$$

10.4.3 A corollary for testing two Hellinger balls

This lemma will be applied in the following setting:

$$H_0 : \ X_1, \dots, X_n \sim P, \quad P \in B_H(P_0, \varepsilon) \coloneqq \{P : H^2(P, P_0) \le \varepsilon^2\},$$

$$H_1 : \ X_1, \dots, X_n \sim Q, \quad Q \in B_H(Q_0, \varepsilon).$$

Corollary 10.9. *If $H(P_0, Q_0) \ge 4\varepsilon$, then*

$$\inf_T \Big(\sup_{P \in B_H(P_0,\varepsilon)} P^{\otimes n}(T = 1) + \sup_{Q \in B_H(Q_0,\varepsilon)} Q^{\otimes n}(T = 0) \Big) \ \le \ \exp\Big(- \frac{n}{8} H^2(P_0, Q_0) \Big).$$

Proof. Since $(P, Q) \mapsto H^2(P, Q)$ is jointly convex (Lecture 3), both balls $B_H(P_0, \varepsilon)$ and $B_H(Q_0, \varepsilon)$ are convex. Therefore the previous lemma applies once we lower bound

$$\begin{aligned} H\big(B_H(P_0,\varepsilon), B_H(Q_0,\varepsilon)\big) &\coloneqq \inf_{P \in B_H(P_0,\varepsilon)} \inf_{Q \in B_H(Q_0,\varepsilon)} H(P, Q) \\ &\ge \inf_{P \in B_H(P_0,\varepsilon)} \inf_{Q \in B_H(Q_0,\varepsilon)} \Big(H(P_0, Q_0) - H(P, P_0) - H(Q, Q_0) \Big) \\ &\ge H(P_0, Q_0) - 2\varepsilon \\ &\ge \frac{1}{2} H(P_0, Q_0), \end{aligned}$$

where the last step uses $H(P_0, Q_0) \ge 4\varepsilon$. Plugging into the testing lemma yields the claimed exponent $\frac{n}{2} \cdot (\frac{1}{2} H(P_0, Q_0))^2 = \frac{n}{8} H^2(P_0, Q_0)$. □

10.4.4 Le Cam–Birgé pairwise comparison estimator

Let $P_1, \dots, P_N$ be a maximal ε-packing of $\mathcal{P}$ under H, i.e.

$$H(P_i, P_j) \ge \varepsilon, \qquad \forall i \neq j.$$

Since a maximal ε-packing is also an ε-covering,

$$\sup_{P \in \mathcal{P}} \min_{i \in [N]} H(P, P_i) \le \varepsilon.$$

For $\delta = 4\varepsilon$ and $H(P_i, P_j) > \delta$, construct a test T_{ij} for

$$H_0 : \ P \in B_H(P_i, \varepsilon) \qquad \text{vs.} \qquad H_1 : \ P \in B_H(P_j, \varepsilon).$$

By the above corollary, there exists T_{ij} (and $T_{ji} \coloneqq 1 - T_{ij}$) such that

$$\sup_{P \in B_H(P_i,\varepsilon)} \mathbb{P}(T_{ij} = 1) \le \exp\Big(- \frac{n}{8} H(P_i, P_j)^2 \Big).$$

Now define the following estimator.

- For $i \in [N]$, let
$$\psi_i \coloneqq \max \big\{ H(P_i, P_j) : \ T_{ij} = 1, \ H(P_i, P_j) > \delta \big\},$$
with the convention $\psi_i = 0$ if no such j exists.
- Set $\widehat{P} = P_{\widehat{i}}$, where $\widehat{i} = \arg\min_{i \in [N]} \psi_i$.

Theorem 10.10. *If $n\varepsilon_n^2 \geq \max\{\log N_H(\mathcal{P}, \varepsilon_n),\, 1\}$, then the above estimator $\widehat{P}$ with $\varepsilon = \varepsilon_n$ satisfies*

$$\sup_{P\in\mathcal{P}} \mathbb{P}\big(H(P, \widehat{P}) > 4t\varepsilon_n\big) \leq Ce^{-t^2}, \qquad \forall t \geq 1.$$

Consequently,

$$\sup_{P\in\mathcal{P}} \mathbb{E}\big[H^2(P, \widehat{P})\big] = O(\varepsilon_n^2).$$

Proof. Since $\{P_1, \ldots, P_N\}$ is an ε-covering, WLOG assume $H(P, P_1) \leq \varepsilon$. For $\delta = 4\varepsilon$ and $t \geq 1$,

$$\begin{aligned} \{H(\widehat{P}, P_1) \geq t\delta\} &= \{H(P_{\widehat{i}}, P_1) \geq t\delta\} \\ &\subseteq \Big\{ \max\{\psi_{\widehat{i}}, \psi_1\} \geq t\delta \Big\} = \{\psi_1 \geq t\delta\} \qquad (\psi_{\widehat{i}} \leq \psi_1) \\ &\subseteq \bigcup_{j:\, H(P_1, P_j) \geq t\delta} \{T_{1j} = 1\}. \end{aligned}$$

(One of $T_{\widehat{i},1}$ and $T_{1,\widehat{i}}$ must be 1.) By a union bound,

$$\mathbb{P}\big(H(\widehat{P}, P_1) \geq t\delta\big) \leq N \exp\Big(-\frac{n}{8}(t\delta)^2\Big) = N_H(\mathcal{P}, \varepsilon)\, e^{-2nt^2\varepsilon^2}.$$

Since $n\varepsilon^2 \geq \max\{1, \log N_H(\mathcal{P}, \varepsilon)\}$, this probability is at most $O(e^{-t^2})$. Finally,

$$\mathbb{P}\big(H(\widehat{P}, P) \geq t\delta\big) \leq \mathbb{P}\big(H(\widehat{P}, P_1) \geq t\delta - \varepsilon\big),$$

by the triangle inequality $H(\widehat{P}, P) \leq H(\widehat{P}, P_1) + H(P_1, P)$. □

Remark 10.11.

1. $\widehat{P}$ is proper, i.e. $\widehat{P} \in \mathcal{P}$.
2. A high-probability upper bound on $H(\widehat{P}, P)$ is established above.

10.5 Refinement via local entropy

It turns out that the global entropy $\log N_H(\mathcal{P}, \varepsilon)$ can be improved to a *local* entropy $\log N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)$, with

$$N_{\mathrm{loc}}(\mathcal{P}, \varepsilon) \;:=\; \sup_{P\in\mathcal{P}} \sup_{\eta\geq\varepsilon} N_H\Big(B_H(P, \eta) \cap \mathcal{P},\; \frac{\eta}{2}\Big).$$

(In other words, we are using balls of radius $\eta/2$ to cover balls of radius η.)

Example. For many d-dimensional families $\mathcal{P}$, we usually have

$$\log N_H(\mathcal{P}, \varepsilon) \simeq d \log\frac{1}{\varepsilon}, \qquad \log N_{\mathrm{loc}}(\mathcal{P}, \varepsilon) \simeq d.$$

Therefore, using local entropy improves the Hellinger result from $O\big(\frac{d\log n}{n}\big)$ to $O\big(\frac{d}{n}\big)$.

Theorem 10.12. *The same guarantee holds for the Le Cam–Birgé pairwise comparison estimator, with N_H replaced by N_{loc}.*

Proof. Let $2^\ell \le t < 2^{\ell+1}$. Decompose

$$\{j \in [N] :\ H(P_1, P_j) \ge t\delta\} \subseteq \bigcup_{k \ge \ell} A_k, \qquad A_k \coloneqq \{j \in [N] :\ 2^k\delta \le H(P_1, P_j) < 2^{k+1}\delta\}.$$

By a union bound,

$$\begin{aligned}\mathbb{P}\big(H(\widehat{P}, P_1) \ge t\delta\big) &\le \mathbb{P}(\psi_1 \ge t\delta)\\ &\le \sum_{k \ge \ell} \mathbb{P}\big(2^k\delta \le \psi_1 < 2^{k+1}\delta\big)\\ &\le \sum_{k \ge \ell} |A_k| \exp\Big(-\frac{n}{8}(2^k\delta)^2\Big).\end{aligned}$$

To upper bound $|A_k|$, since $\{P_1, \dots, P_N\}$ is an ε-packing,

$$\begin{aligned}|A_k| &\le M\Big(\{P \in \mathcal{P} : 2^k\delta \le H(P_1, P) < 2^{k+1}\delta\},\ \varepsilon\Big)\\ &\le M\big(B_H(P_1, 2^{k+1}\delta) \cap \mathcal{P},\ \varepsilon\big)\\ &\le N\big(B_H(P_1, 2^{k+1}\delta) \cap \mathcal{P},\ \varepsilon/2\big)\\ &\le N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)^{k+4}, \qquad \text{(see lemma below)}.\end{aligned}$$

Therefore,

$$\mathbb{P}\big(H(\widehat{P}, P_1) \ge t\delta\big) \le \sum_{k \ge \ell} \exp\Big((k+4)\log N_{\mathrm{loc}}(\mathcal{P}, \varepsilon) - 2n\varepsilon^2\, 4^k\Big) \le e^{-\Omega(4^\ell)} = e^{-\Omega(t^2)},$$

provided $n\varepsilon^2 \ge \max\{1, \log N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)\}$. □

Lemma 10.13. *For $\eta \ge \varepsilon$,*

$$N_H\Big(B_H(P, 2^k\eta) \cap \mathcal{P},\ \frac{\eta}{2}\Big) \le N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)^{k+1}.$$

Proof. Induction on k. The base case $k = 0$ is the definition of $N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)$. For the inductive step, first cover $B_H(P, 2^k\eta) \cap \mathcal{P}$ using balls of radius $2^{k-1}\eta$, then cover each such ball using balls of radius $\eta/2$. Writing $N_k \coloneqq N_H(B_H(P, 2^k\eta) \cap \mathcal{P}, \eta/2)$, this gives

$$N_k \le N_{k-1}N_0 \le N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)^k\, N_{\mathrm{loc}}(\mathcal{P}, \varepsilon) = N_{\mathrm{loc}}(\mathcal{P}, \varepsilon)^{k+1}.$$

□

10.6 Special topic: High-probability density estimation under KL

Guest lecture by J. Qian on his recent work on high-probability density estimation under KL.

Lecture 11: Universal Compression and Redundancy

11.1 Motivation: compressing without knowing the source

In Lecture 1 we saw that for any distribution P_{X^n} on sequences $x^n = (x_1, \ldots, x_n)$, there exists a uniquely-decodable code f whose expected length satisfies

$$\mathbb{E}_{P_{X^n}}[\ell(f(X^n))] \le H(P_{X^n}) + 1 \quad \text{(bits)}.$$

The catch: the code depends on P_{X^n}. In *universal compression* we want a single code (equivalently a single distribution Q_{X^n}) that performs well for every P in some class $\mathcal{P}$.

11.1.1 Warm-up: a Bernoulli example

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$, with unknown $p \in [0,1]$, and alphabet $\mathcal{X} = \{0,1\}$. Let

$$N_1 \coloneqq n_1(X^n) = \sum_{i=1}^{n} \mathbb{1}\{X_i = 1\}, \qquad N_0 \coloneqq n_0(X^n) = n - N_1.$$

A simple code:

(a) Encode the count $N_1 \in \{0, 1, \ldots, n\}$ using $\log(n+1)$ bits.

(b) Given N_1, the sequence X^n is determined by the set of indices where $X_i = 1$, of which there are $\binom{n}{N_1}$ possibilities; encode this using $\log \binom{n}{N_1}$ bits.

Thus

$$\ell\big(f(X^n)\big) = \log(n+1) + \log \binom{n}{N_1}.$$

Taking expectation under $\mathrm{Ber}(p)^{\otimes n}$ gives

$$\begin{aligned}
\mathbb{E}[\ell(f(X^n))] &= \log(n+1) + \mathbb{E}\Big[\log \binom{n}{N_1}\Big] \\
&\le \log(n+1) + n\,\mathbb{E}\Big[H\big(\mathrm{Ber}(N_1/n)\big)\Big] \qquad \Big(\log \binom{n}{k} \le n\,H\big(\mathrm{Ber}(k/n)\big)\Big) \\
&\le \log(n+1) + n\,H\Big(\mathrm{Ber}\big(\mathbb{E}[N_1/n]\big)\Big) \qquad \big(H(\mathrm{Ber}(\cdot)) \text{ is concave}\big) \\
&= \log(n+1) + n\,H\big(\mathrm{Ber}(p)\big),
\end{aligned}$$

where $H(\mathrm{Ber}(q)) = -q \log q - (1-q)\log(1-q)$. So we get a universal overhead of about $\log n$ bits.

11.2 Minimax redundancy and sequential estimators

Any uniquely-decodable code induces a sub-probability distribution Q_{X^n} via Kraft:

$$Q_{X^n}(x^n) := 2^{-\ell(f(x^n))}, \qquad \sum_{x^n} Q_{X^n}(x^n) \le 1.$$

Then for any distribution P_{X^n},

$$\mathbb{E}_{P_{X^n}}[\ell(f(X^n))] = \mathbb{E}_{P_{X^n}}\Big[\log \frac{1}{Q_{X^n}(X^n)}\Big].$$

The "overhead" relative to the entropy is

$$\mathbb{E}_{P_{X^n}}\Big[\log \frac{1}{Q_{X^n}(X^n)}\Big] - H(P_{X^n}) = \mathbb{E}_{P_{X^n}}\Big[\log \frac{P_{X^n}(X^n)}{Q_{X^n}(X^n)}\Big] = D_{\mathrm{KL}}\big(P_{X^n} \,\|\, Q_{X^n}\big).$$

Definition 11.1 (Minimax redundancy)**.** For a model class $\mathcal{P}$ of distributions over $\mathcal{X}^n$, the (expected) minimax redundancy is

$$\mathrm{Red}(\mathcal{P}) := \inf_{Q_{X^n}} \sup_{P_{X^n} \in \mathcal{P}} D_{\mathrm{KL}}\big(P_{X^n} \,\|\, Q_{X^n}\big).$$

In many cases $\mathrm{Red}(\mathcal{P}) = o(n)$ and is often on the order of $\log n$.

11.2.1 Bernoulli model via Laplace and Krichevsky–Trofimov

Let

$$\mathcal{P} \;=\; \big\{\mathrm{Ber}(p)^{\otimes n} : p \in [0,1]\big\}.$$

We construct Q_{X^n} sequentially. For a prefix $x^t = (x_1, \dots, x_t)$ define

$$n_1(x^t) := \sum_{i=1}^{t} \mathbb{1}\{x_i = 1\}, \qquad n_0(x^t) := t - n_1(x^t).$$

Laplace (add-1) estimator. Define for $t \ge 0$,

$$Q_{X_{t+1}|X^t}(1 \mid x^t) = \frac{n_1(x^t)+1}{t+2}, \qquad Q_{X_{t+1}|X^t}(0 \mid x^t) = \frac{n_0(x^t)+1}{t+2}.$$

(For $t = 0$ this gives $Q_{X_1}(1) = Q_{X_1}(0) = 1/2$.) Then

$$\begin{aligned} Q_{X^n}(x^n) &= \prod_{t=0}^{n-1} Q_{X_{t+1}|X^t}(x_{t+1} \mid x^t) \\ &= \frac{(1 \cdot 2 \cdots n_1(x^n))\,(1 \cdot 2 \cdots n_0(x^n))}{2 \cdot 3 \cdots (n+1)} = \frac{n_1(x^n)!\, n_0(x^n)!}{(n+1)!}. \end{aligned}$$

On the other hand, for any $p \in [0,1]$,

$$P_{X^n}(x^n) = p^{n_1(x^n)}(1-p)^{n_0(x^n)} \le \Big(\frac{n_1(x^n)}{n}\Big)^{n_1(x^n)} \Big(\frac{n_0(x^n)}{n}\Big)^{n_0(x^n)}.$$

Therefore

$$\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \le (n+1)\Big(\frac{n_1(x^n)}{n}\Big)^{n_1(x^n)} \Big(\frac{n_0(x^n)}{n}\Big)^{n_0(x^n)} \frac{n!}{n_1(x^n)!\, n_0(x^n)!} = O(n), \qquad \text{(by Stirling)}.$$

This implies

$$\mathrm{Red}(Q_{X^n};\mathcal{P}) := \sup_{P_{X^n}\in\mathcal{P}} D_{\mathrm{KL}}(P_{X^n} \| Q_{X^n}) = \sup_{P_{X^n}\in\mathcal{P}} \mathbb{E}_{P_{X^n}}\Big[\log \frac{P_{X^n}(X^n)}{Q_{X^n}(X^n)}\Big] \le \log n + O(1).$$

Krichevsky–Trofimov (add-$\frac{1}{2}$) estimator. Now define

$$Q_{X_{t+1}|X^t}(1 \mid x^t) = \frac{n_1(x^t) + \frac{1}{2}}{t+1}, \qquad Q_{X_{t+1}|X^t}(0 \mid x^t) = \frac{n_0(x^t) + \frac{1}{2}}{t+1}.$$

This is the add-$\frac{1}{2}$ / Krichevsky–Trofimov estimator. In this case

$$\begin{aligned} Q_{X^n}(x^n) &= \prod_{t=0}^{n-1} Q_{X_{t+1}|X^t}(x_{t+1} \mid x^t) \\ &= \frac{1}{n!}\Big(\frac{1}{2}\cdot\frac{3}{2}\cdots\big(n_1(x^n) - \tfrac{1}{2}\big)\Big)\Big(\frac{1}{2}\cdot\frac{3}{2}\cdots\big(n_0(x^n) - \tfrac{1}{2}\big)\Big) \\ &= \frac{(2n_1(x^n) - 1)!!\,(2n_0(x^n) - 1)!!}{2^n\, n!}. \end{aligned}$$

Moreover,

$$\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \le \frac{2^n n!\,\big(n_1(x^n)\big)^{n_1(x^n)}\big(n_0(x^n)\big)^{n_0(x^n)}}{n^n\,(2n_1(x^n) - 1)!!\,(2n_0(x^n) - 1)!!} = O(\sqrt{n}), \qquad \text{(Stirling)}.$$

Therefore

$$\mathrm{Red}(Q_{X^n}; \mathcal{P}) \le \log(C\sqrt{n}) = \frac{1}{2}\log n + O(1).$$

This constant 1/2 turns out to be tight:

$$\mathrm{Red}(\mathcal{P}) = \frac{1}{2}\log n + O(1).$$

11.3 Worst-case / pointwise redundancy and Shtarkov's theorem

In the previous examples we implicitly used the fact that $\mathrm{Red}(\mathcal{P}) \le R^*(\mathcal{P})$, where $R^*(\mathcal{P})$ is a worst-case (pointwise) analogue.

Definition 11.2 (Worst-case / pointwise redundancy)**.**

$$R^*(\mathcal{P}) := \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n\in\mathcal{X}^n} \log\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)}.$$

Remark 11.3 (Connection to log-loss regret). It is clear that $\mathrm{Red}(\mathcal{P}) \le R^*(\mathcal{P})$. Also, $R^*(\mathcal{P})$ treats x^n as an individual sequence rather than a random draw.

Let $Q_{X^n} = \prod_{t=1}^n Q_{X_t|X^{t-1}}$ be a sequential predictor and define log-loss $\ell_{\log}(q, x) = \log\frac{1}{q(x)}$. Then for any individual sequence x^n,

$$\log\frac{1}{Q_{X^n}(x^n)} = \sum_{t=1}^n \ell_{\log}\big(Q_{X_t|X^{t-1}}(\cdot \mid x^{t-1}), x_t\big).$$

Similarly, for any $P \in \mathcal{P}$,

$$\log\frac{1}{P_{X^n}(x^n)} = \sum_{t=1}^n \ell_{\log}\big(P_{X_t|X^{t-1}}(\cdot \mid x^{t-1}), x_t\big).$$

Hence

$$R^*(\mathcal{P}) = \inf_{Q_{X^n}} \sup_{x^n}\left\{\sum_{t=1}^n \ell_{\log}\big(Q_{X_t|X^{t-1}}(\cdot \mid x^{t-1}), x_t\big) - \inf_{P\in\mathcal{P}} \sum_{t=1}^n \ell_{\log}\big(P_{X_t|X^{t-1}}(\cdot \mid x^{t-1}), x_t\big)\right\}.$$

So $R^*(\mathcal{P})$ is the minimax regret under log-loss.

Unlike $\mathrm{Red}(\mathcal{P})$, which can be hard to characterize, $R^*(\mathcal{P})$ has a clean combinatorial expression.

Theorem 11.4 (Shtarkov sum / normalized maximum likelihood)**.**

$$R^*(\mathcal{P}) = \log\Big(\sum_{x^n\in\mathcal{X}^n} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)\Big).$$

The quantity $\sum_{x^n} \sup_{P\in\mathcal{P}} P(x^n)$ *is called the* Shtarkov sum.

Proof. (Upper bound.) Let

$$Z := \sum_{x^n\in\mathcal{X}^n} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n), \qquad Q^*_{X^n}(x^n) := \frac{1}{Z} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n).$$

This Q^* is the *normalized maximum likelihood* (NML) distribution. Then

$$\sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log \frac{P_{X^n}(x^n)}{Q^*_{X^n}(x^n)} = \log Z,$$

so $R^*(\mathcal{P}) \le \log Z$.

(Lower bound.) For any Q_{X^n},

$$\begin{aligned}
\sup_{P\in\mathcal{P}} \sup_{x^n} \log \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} &= \sup_{P\in\mathcal{P}} \sup_{x^n} \Big(\log \frac{P_{X^n}(x^n)}{Q^*_{X^n}(x^n)} + \log \frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)}\Big) \\
&= \log Z + \sup_{x^n} \log \frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)} \\
&\ge \log Z + \sum_{x^n} Q^*_{X^n}(x^n) \log \frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)} \\
&= \log Z + D_{\mathrm{KL}}(Q^*_{X^n} \| Q_{X^n}) \ge \log Z.
\end{aligned}$$

Taking the infimum over Q_{X^n} gives $R^*(\mathcal{P}) \ge \log Z$. □

11.3.1 Example: time-homogeneous Markov chains

This combinatorial nature of $R^*(\mathcal{P})$ makes it easy to upper bound $\mathrm{Red}(\mathcal{P})$ for non-i.i.d. families.

Example 11.5 (First-order Markov chains)**.** Let

$$\mathcal{P} = \Big\{P_{X^n} = p(x_1) \prod_{t=1}^{n-1} M(x_{t+1} \mid x_t)\Big\}$$

be the class of all time-homogeneous (first-order) Markov chains on state space $[k] = \{1, \dots, k\}$.

Claim 11.6.

$$\mathrm{Red}(\mathcal{P}) \le \frac{k(k-1)}{2} \log n + O_k(1).$$

Proof. Apply the add-$\frac{1}{2}$ estimator to each row of the transition matrix. Define, for $x^t = (x_1, \dots, x_t)$,

$$n_{j\to i}(x^t) := \sum_{s=1}^{t-1} \mathbb{1}\{x_s = j,\, x_{s+1} = i\}, \qquad n_j(x^t) := \sum_{s=1}^{t-1} \mathbb{1}\{x_s = j\}.$$

If $x_t = j$, set

$$Q_{X_{t+1}|X^t}(i \mid x^t) = \frac{n_{j\to i}(x^t) + \frac{1}{2}}{n_j(x^t) + \frac{k}{2}}.$$

Then for any $x^n \in [k]^n$,

$$\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} = \frac{p(x_1)}{1/k} \prod_{j=1}^{k} \prod_{\substack{t\in[n-1]:\\ x_t=j}} \frac{M(x_{t+1} \mid j)}{Q_{X_{t+1}|X^t}(x_{t+1} \mid x^t)}.$$

For each fixed j, the inner product behaves like a k-ary i.i.d. model and contributes a factor $O(\sqrt{n})^{k-1}$ by the i.i.d. analysis. Hence

$$\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} \le k\,(C\sqrt{n})^{k(k-1)}.$$

Taking logs and using $\mathrm{Red}(\mathcal{P}) \le R^*(\mathcal{P})$ yields

$$\mathrm{Red}(\mathcal{P}) \le R^*(\mathcal{P}) \le \log\bigl(k(C\sqrt{n})^{k(k-1)}\bigr) = \frac{k(k-1)}{2}\log n + O(k^2).$$

□

The same approach can be extended to other processes such as hidden Markov models (see, e.g., Gassiat (2018)).

11.4 Redundancy bounds for i.i.d. families

11.4.1 Entropic upper bound

By the global Fano proof (Lecture 9), we have the following entropy/covering-number upper bound.

Theorem 11.7 (Entropic upper bound)**.**

$$\mathrm{Red}\bigl(\mathcal{P}^{\otimes n}\bigr) \le \inf_{\varepsilon>0}\Bigl(n\varepsilon^2 + \log N_{\mathrm{KL}}(\mathcal{P},\varepsilon)\Bigr).$$

Example 11.8 (Parametric families)**.** If $\mathcal{P} = (P_\theta)_{\theta\in\mathbb{R}^d}$ is a d-parameter family, typically $\log N_{\mathrm{KL}}(\mathcal{P},\varepsilon) \asymp d\log(1/\varepsilon)$. Choosing $\varepsilon \asymp \sqrt{d/n}$ yields

$$\mathrm{Red}\bigl(\mathcal{P}^{\otimes n}\bigr) \le \frac{d}{2}\log\frac{n}{d} + O(d).$$

11.4.2 A variational formula: redundancy–capacity theorem

We begin with a variational representation of $\mathrm{Red}(\mathcal{P})$.

Theorem 11.9 (Redundancy–capacity theorem)**.** *Let $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$. Then*

$$\mathrm{Red}(\mathcal{P}) = \sup_{\rho\in\Delta(\Theta)} I(\theta;X), \qquad \textit{where } \theta\sim\rho,\ X\mid\theta\sim P_\theta.$$

The quantity $\sup_\rho I(\theta;X)$ is the capacity of the "channel" $\theta \mapsto X$ with law P_θ.

Proof. The "golden formula" for mutual information (Lecture 7) states

$$I(\theta; X) = \inf_{Q_X} \mathbb{E}_{\theta\sim\rho}\big[D_{\mathrm{KL}}(P_\theta \| Q_X)\big].$$

Therefore

$$\begin{aligned}
\sup_\rho I(\theta; X) &= \sup_\rho \inf_{Q_X} \mathbb{E}_{\theta\sim\rho}[D_{\mathrm{KL}}(P_\theta \| Q_X)] \\
&= \inf_{Q_X} \sup_\rho \mathbb{E}_{\theta\sim\rho}[D_{\mathrm{KL}}(P_\theta \| Q_X)] \qquad \text{(minimax theorem)} \\
&= \inf_{Q_X} \sup_{\theta\in\Theta} D_{\mathrm{KL}}(P_\theta \| Q_X) = \mathrm{Red}(\mathcal{P}).
\end{aligned}$$

□

11.4.3 Rissanen's lower bound

Rissanen's program: find an estimator $\hat\theta(X^n)$ such that

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta\big[\|\theta - \hat\theta(X^n)\|^2\big] \le \varepsilon_n^2.$$

Theorem 11.10 (Rissanen)**.** *Let $\Theta \subseteq \mathbb{R}^d$ have non-empty interior. Then*

$$\mathrm{Red}\big(\mathcal{P}^{\otimes n}\big) \ge \log \mathrm{Vol}_d(\Theta) - \frac{d}{2}\log\Big(\frac{2\pi e\, \varepsilon_n^2}{d}\Big).$$

Proof. Let $\theta \sim \rho = \mathrm{Unif}(\Theta)$, and let $h(\cdot)$ denote differential entropy on $\mathbb{R}^d$. Then

$$I(\theta; X^n) = h(\theta) - h(\theta \mid X^n) = \log \mathrm{Vol}_d(\Theta) - h(\theta \mid X^n).$$

Moreover,

$$\begin{aligned}
h(\theta \mid X^n) &= h(\theta - \hat\theta(X^n) \mid X^n) \le h(\theta - \hat\theta(X^n)) \qquad \text{(conditioning reduces entropy)} \\
&\le \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det\Big(\mathbb{E}\big[(\theta-\hat\theta)(\theta-\hat\theta)^\top\big]\Big) \qquad \text{(Gaussian maximizes entropy for fixed covariance)} \\
&\le \frac{d}{2}\log(2\pi e) + \frac{d}{2}\log\Big(\frac{\mathbb{E}\|\theta - \hat\theta(X^n)\|^2}{d}\Big) \qquad \Big(\det(A) = \prod_i \lambda_i \le \big(\tfrac{\mathrm{Tr}(A)}{d}\big)^d\Big) \\
&\le \frac{d}{2}\log\Big(\frac{2\pi e\,\varepsilon_n^2}{d}\Big).
\end{aligned}$$

By the redundancy–capacity theorem, $\mathrm{Red}(\mathcal{P}^{\otimes n}) \ge \sup_\rho I(\theta; X^n)$, so the claim follows. □

Example 11.11. In parametric families, typically $\mathrm{Vol}_d(\Theta) = \Omega\big((1/n)^{d/2}\big)$ and $\varepsilon_n^2 = O(d/n)$. Therefore Rissanen's bound gives

$$\mathrm{Red}\big(\mathcal{P}^{\otimes n}\big) \ge \frac{d}{2}\log\frac{n}{d} - O(d).$$

11.4.4 Haussler–Opper lower bound

The argument of Haussler & Opper (1997) chooses ρ to be a uniform mixture $\rho = \frac{1}{M}\sum_{i=1}^{M}\delta_{\theta_i}$.

Lemma 11.12. *For $X \mid \theta \sim P_\theta$ and $0 < \lambda \le 1$,*

$$I(\theta; X) \ge -\mathbb{E}_{\theta,X}\Big[\log \mathbb{E}_{\theta'}\Big(\frac{P_{\theta'}(X)}{P_\theta(X)}\Big)^{\lambda}\Big],$$

where θ' is an independent copy of θ (so $\theta' \perp\!\!\!\perp (\theta, X)$).

Proof. Let

$$f(\lambda) := -\mathbb{E}_{\theta,X}\Big[\log \mathbb{E}_{\theta'}\Big(\frac{P_{\theta'}(X)}{P_\theta(X)}\Big)^{\lambda}\Big].$$

Then $f(1) = I(\theta; X)$. Since cumulant generating functions are convex, f is concave in λ. A calculation gives

$$f'(\lambda) = \mathbb{E}_{\theta,X}[\log P_\theta(X)] - \mathbb{E}_{\theta,X}\left[\frac{\mathbb{E}_{\theta'}\big[P_{\theta'}(X)^{\lambda}\log P_{\theta'}(X)\big]}{\mathbb{E}_{\theta'}\big[P_{\theta'}(X)^{\lambda}\big]}\right].$$

At $\lambda = 1$,

$$\begin{aligned}\mathbb{E}_{\theta,X}\left[\frac{\mathbb{E}_{\theta'}\big[P_{\theta'}(X)\log P_{\theta'}(X)\big]}{\mathbb{E}_{\theta'}\big[P_{\theta'}(X)\big]}\right] &= \mathbb{E}_X\left[\frac{\int p(\theta')P_{\theta'}(X)\log P_{\theta'}(X)\,d\theta'}{\int p(\theta')P_{\theta'}(X)\,d\theta'}\right]\\ &= \int\int p(\theta)p(\theta')P_\theta(x)\log P_\theta(x)\,d\theta\,d\theta'\,dx = \mathbb{E}_{\theta,X}[\log P_\theta(X)],\end{aligned}$$

so $f'(1) = 0$. By concavity of f, for $\lambda \le 1$ we have $f'(\lambda) \ge 0$, hence $f(\lambda) \le f(1) = I(\theta; X)$. □

Theorem 11.13 (Haussler–Opper)**.**

$$\mathrm{Red}\big(\mathcal{P}^{\otimes n}\big) \ge \sup_{\varepsilon>0}\min\Big\{\frac{n\varepsilon^2}{2},\ \log M_H(\mathcal{P},\varepsilon)\Big\} - \log 2,$$

where $M_H(\mathcal{P},\varepsilon)$ is the ε-packing number of $\mathcal{P}$ under Hellinger distance.

Proof. Let $P_{\theta_1},\dots,P_{\theta_M}$ be an ε-packing of $\mathcal{P}$ under Hellinger distance, and take $\rho = \frac{1}{M}\sum_{i=1}^{M}\delta_{\theta_i}$. Apply Lemma 11.12 with $\lambda = \frac{1}{2}$ to $(X^n \mid \theta) \sim P_\theta^{\otimes n}$:

$$\begin{aligned}I(\theta; X^n) &\ge -\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{P_{\theta_i}^{\otimes n}}\left[\log\Big(\frac{1}{M}\sum_{j=1}^{M}\Big(\frac{P_{\theta_j}^{\otimes n}(X^n)}{P_{\theta_i}^{\otimes n}(X^n)}\Big)^{1/2}\Big)\right]\\ &\ge -\frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_{P_{\theta_i}^{\otimes n}}\Big[\Big(\frac{P_{\theta_j}^{\otimes n}(X^n)}{P_{\theta_i}^{\otimes n}(X^n)}\Big)^{1/2}\Big]\right) \qquad (x \mapsto -\log x \text{ convex}).\end{aligned}$$

The inner expectation is the Hellinger affinity:

$$\mathbb{E}_{P_{\theta_i}^{\otimes n}}\Big[\Big(\frac{P_{\theta_j}^{\otimes n}(X^n)}{P_{\theta_i}^{\otimes n}(X^n)}\Big)^{1/2}\Big] = \int\sqrt{dP_{\theta_i}^{\otimes n}\,dP_{\theta_j}^{\otimes n}} = \Big(\int\sqrt{dP_{\theta_i}\,dP_{\theta_j}}\Big)^n = \Big(1 - \frac{H^2(P_{\theta_i},P_{\theta_j})}{2}\Big)^n.$$

Since the family is an ε-packing, $H^2(P_{\theta_i}, P_{\theta_j}) \geq \varepsilon^2$ for $i \neq j$, hence

$$\Big(1 - \frac{H^2(P_{\theta_i}, P_{\theta_j})}{2}\Big)^n \leq \Big(1 - \frac{\varepsilon^2}{2}\Big)^n \leq e^{-n\varepsilon^2/2}.$$

Therefore

$$I(\theta; X^n) \geq -\log\Big(\frac{1}{M} + e^{-n\varepsilon^2/2}\Big) \geq \min\Big\{\log M,\ \frac{n\varepsilon^2}{2}\Big\} - \log 2,$$

using the elementary inequality $\frac{1}{a} + \frac{1}{b} \leq \frac{2}{\min\{a,b\}}$. Finally, by redundancy–capacity, $\mathrm{Red}(\mathcal{P}^{\otimes n}) \geq \sup_\rho I(\theta; X^n)$. □

Example 11.14 (Parametric families)**.** If typically $\log M_H(\mathcal{P}, \varepsilon) \asymp d \log(1/\varepsilon)$, then the Haussler–Opper bound gives

$$\mathrm{Red}\big(\mathcal{P}^{\otimes n}\big) \gtrsim \frac{d}{2} \log \frac{n}{C d \log n}.$$

11.5 Redundancy and prediction risk

Definition 11.15 (Prediction risk)**.** The (next-symbol) prediction risk under KL is

$$\mathrm{Risk}_n(\mathcal{P}) \coloneqq \inf_{Q_{X_{n+1}|X^n}} \sup_{P_{X^{n+1}} \in \mathcal{P}} \mathbb{E}_{P_{X^n}} \Big[D_{\mathrm{KL}}\big(P_{X_{n+1}|X^n} \,\|\, Q_{X_{n+1}|X^n}\big)\Big].$$

11.5.1 Mutual-information representation

If $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$, then

$$\mathrm{Risk}_n(\mathcal{P}) = \sup_{\rho \in \Delta(\Theta)} I(\theta; X_{n+1} \mid X^n), \qquad \theta \sim \rho,\ X \mid \theta \sim P_\theta.$$

Proof.

$$\begin{aligned}
\mathrm{Risk}_n(\mathcal{P}) &= \inf_{Q_{X_{n+1}|X^n}} \sup_\rho \mathbb{E}_{\theta\sim\rho}\big[D_{\mathrm{KL}}(P_{X_{n+1}|X^n,\theta} \| Q_{X_{n+1}|X^n})\big] \\
&= \sup_\rho \inf_{Q_{X_{n+1}|X^n}} \mathbb{E}_{\theta\sim\rho}\big[D_{\mathrm{KL}}(P_{X_{n+1}|X^n,\theta} \| Q_{X_{n+1}|X^n})\big] \qquad \text{(minimax theorem)} \\
&= \sup_\rho I(\theta; X_{n+1} \mid X^n).
\end{aligned}$$

□

11.5.2 Redundancy–risk inequality

Let $\mathrm{Red}_n(\mathcal{P})$ denote the minimax redundancy for sequences of length n. Then

$$\mathrm{Red}_n(\mathcal{P}) \leq \sum_{t=0}^{n-1} \mathrm{Risk}_t(\mathcal{P}).$$

Proof. By the chain rule for mutual information,

$$I(\theta; X^n) = \sum_{t=0}^{n-1} I(\theta; X_{t+1} \mid X^t).$$

Taking $\sup_\rho$ of both sides gives

$$\sup_\rho I(\theta; X^n) \le \sum_{t=0}^{n-1} \sup_\rho I(\theta; X_{t+1} \mid X^t) = \sum_{t=0}^{n-1} \mathrm{Risk}_t(\mathcal{P}).$$

Using redundancy–capacity, $\mathrm{Red}_n(\mathcal{P}) = \sup_\rho I(\theta; X^n)$. □

Remark 11.16 (Tightness for i.i.d. parametric models)*.* For i.i.d. $\mathcal{P}^{\otimes n}$ with $\Theta \subset \mathbb{R}^d$, the MLE $\hat{\theta}_t$ based on X^t typically satisfies

$$\mathbb{E}_\theta\big[D_{\mathrm{KL}}(P_\theta \| P_{\hat{\theta}_t})\big] \sim \frac{d}{2t} \qquad \text{(Wilks' theorem)},$$

so $\mathrm{Risk}_t \sim d/(2t)$ and $\mathrm{Red}_n \sim (d/2)\log n \sim \sum_{t=1}^n \mathrm{Risk}_t$.

11.5.3 Online-to-batch conversion for stationary processes

Assume each $P_{X^{n+1}} \in \mathcal{P}$ is stationary, i.e.

$$P_{X_{t_1},\ldots,X_{t_k}} = P_{X_{t_1+t_0},\ldots,X_{t_k+t_0}} \qquad \text{for all } t_0 \ge 0.$$

Then

$$\mathrm{Risk}_n(\mathcal{P}) \le \frac{1}{n}\mathrm{Red}(\mathcal{P}) + \mathrm{Mem}(\mathcal{P}),$$

where the *memory term* is

$$\mathrm{Mem}(\mathcal{P}) \coloneqq \sup_{P_{X^{n+1}} \in \mathcal{P}} \frac{1}{n}\sum_{t=1}^{n} I\big(X_{n+1}; X^{n-t} \mid X_{n-t+1}^n\big).$$

Proof. Let $Q_{X^{n+1}} = \prod_{t=1}^{n+1} Q_{X_t|X^{t-1}}$ attain the minimax redundancy $\mathrm{Red}(\mathcal{P})$. Choose a Yang–Barron type predictor

$$\widetilde{Q}_{X_{n+1}|X^n}(\cdot \mid X^n) \coloneqq \frac{1}{n}\sum_{t=1}^{n} Q_{X_{t+1}|X^t}(\cdot \mid X_{n-t+1}^n).$$

Then, by convexity of KL,

$$\mathbb{E}_{P_{X^n}}\Big[D_{\mathrm{KL}}\big(P_{X_{n+1}|X^n} \,\|\, \widetilde{Q}_{X_{n+1}|X^n}\big)\Big] \le \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{P_{X^{n+1}}}\left[\log \frac{P_{X_{n+1}|X^n}(X_{n+1} \mid X^n)}{Q_{X_{t+1}|X^t}(X_{n+1} \mid X_{n-t+1}^n)}\right].$$

Split the logarithm:

$$\log \frac{P_{X_{n+1}|X^n}(X_{n+1} \mid X^n)}{Q_{X_{t+1}|X^t}(X_{n+1} \mid X_{n-t+1}^n)} = \log \frac{P_{X_{n+1}|X_{n-t+1}^n}(X_{n+1} \mid X_{n-t+1}^n)}{Q_{X_{t+1}|X^t}(X_{n+1} \mid X_{n-t+1}^n)} + \log \frac{P_{X_{n+1}|X^n}(X_{n+1} \mid X^n)}{P_{X_{n+1}|X_{n-t+1}^n}(X_{n+1} \mid X_{n-t+1}^n)}.$$

Taking expectation, the second term becomes $I(X_{n+1}; X^{n-t} \mid X_{n-t+1}^n)$ (by stationarity). Hence

$$\begin{aligned}
\mathbb{E}_{P_{X^n}}\Big[D_{\mathrm{KL}}(P_{X_{n+1}|X^n} \| \widetilde{Q}_{X_{n+1}|X^n})\Big] &\le \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{P_{X^n}}\big[D_{\mathrm{KL}}(P_{X_{t+1}|X^t} \| Q_{X_{t+1}|X^t})\big] + \mathrm{Mem}(\mathcal{P}) \\
&\le \frac{1}{n} D_{\mathrm{KL}}(P_{X^{n+1}} \| Q_{X^{n+1}}) + \mathrm{Mem}(\mathcal{P}) \qquad \text{(chain rule)} \\
&\le \frac{1}{n}\mathrm{Red}(\mathcal{P}) + \mathrm{Mem}(\mathcal{P}).
\end{aligned}$$

Finally take $\sup_{P \in \mathcal{P}}$ and $\inf_Q$ to match the definition of $\mathrm{Risk}_n(\mathcal{P})$. □

Example 11.17 (Markov chain prediction)**.** Let $\mathcal{P}$ be the class of stationary Markov chains on $[k]$ of length $n+1$. Then

$$\mathrm{Red}(\mathcal{P}) = O(k^2 \log n), \qquad \mathrm{Mem}(\mathcal{P}) = \sup_{P\in\mathcal{P}} \frac{1}{n} I(X_{n+1}; X^n) \le \frac{\log k}{n},$$

so

$$\mathrm{Risk}_n(\mathcal{P}) = O\Big(\frac{k^2 \log n}{n}\Big).$$

A surprising feature is that this upper bound does not depend on the mixing property of the Markov chain. A purely statistical proof of this upper bound is unknown without mixing conditions. This bound is tight for $3 \le k \ll \sqrt{n}$.

11.6 Special topic: characterizing R^* in Gaussian models (Mourtada, 2023)

Consider the Gaussian shift family

$$\mathcal{P}_A \coloneqq \{\mathcal{N}(\theta, I_d) : \theta \in A\}, \qquad A \subset \mathbb{R}^d.$$

We use the facts

$$D_{\mathrm{KL}}\big(\mathcal{N}(\theta, I_d) \,\|\, \mathcal{N}(\theta', I_d)\big) = \frac{1}{2}\|\theta - \theta'\|_2^2, \qquad \int \sqrt{d\mathcal{N}(\theta, I_d)\, d\mathcal{N}(\theta', I_d)} = \exp\Big(-\frac{1}{8}\|\theta - \theta'\|_2^2\Big).$$

By the entropic upper bound and Haussler–Opper lower bound, one obtains the characterization

$$\mathrm{Red}(\mathcal{P}_A) \asymp \inf_{r>0}\big(\log N(A, \|\cdot\|_2, r) + r^2\big).$$

The main result of this section is an analogous characterization of $R^*(\mathcal{P}_A)$:

$$R^*(\mathcal{P}_A) \asymp \inf_{r>0}\big(\log N(A, \|\cdot\|_2, r) + w_A(r)\big),$$

where $w_A(r)$ is the *local Gaussian width*

$$w_A(r) \coloneqq \sup_{\theta\in A} w\big(A \cap B(\theta, r)\big) = \sup_{\theta\in A} \mathbb{E}\Big[\sup_{w\in A\cap B(\theta,r)} \langle w, Z\rangle\Big], \qquad Z \sim \mathcal{N}(0, I_d).$$

Remark 11.18 (Alternative representation)*.* Let

$$r_N \coloneqq \sup\big\{r > 0 : \log N(A, \|\cdot\|_2, r) \ge r^2\big\}, \qquad r_w \coloneqq \sup\big\{r > 0 : w_A(r) \ge r^2\big\}.$$

Then

$$\mathrm{Red}(\mathcal{P}_A) \asymp r_N^2, \qquad R^*(\mathcal{P}_A) \asymp r_N^2 + r_w^2.$$

11.6.1 Example: ellipsoids

If

$$A = \Big\{\theta \in \mathbb{R}^d : \sum_{i=1}^d \frac{\theta_i^2}{a_i^2} \le 1\Big\},$$

then

$$\mathrm{Red}(\mathcal{P}_A) \asymp \inf_{r>0}\Big(\sum_{i:a_i>2r} \log\frac{a_i}{r} + r^2\Big), \qquad R^*(\mathcal{P}_A) \asymp \inf_{r>0}\Big(\sum_{i=1}^d \log\Big(1 + \frac{a_i^2}{r^2}\Big) + r^2\Big).$$

11.6.2 A key lemma relating R^* and Gaussian width

The proof hinges on the following lemma.

Lemma 11.19. *Let $w(A) = \mathbb{E}[\sup_{u\in A}\langle u, Z\rangle]$ denote the Gaussian width. Then*

$$w(A) - \sup_{\theta\in A}\frac{\|\theta\|_2^2}{2} \le R^*(\mathcal{P}_A) \le w(A).$$

11.6.3 How the lemma implies the covering/width characterization

Upper bound on R^*. First observe the simple inequality: for families $\mathcal{P}_1, \dots, \mathcal{P}_N$,

$$R^*\Big(\bigcup_{i=1}^{N}\mathcal{P}_i\Big) \le \max_{i\in[N]} R^*(\mathcal{P}_i) + \log N.$$

Indeed, if Q_i attains $R^*(\mathcal{P}_i)$, then $\bar{Q} = \frac{1}{N}\sum_{i=1}^{N} Q_i$ attains the stated upper bound.

Now take an r-covering $\theta_1, \dots, \theta_N$ of A under $\|\cdot\|_2$. Then

$$\begin{aligned} R^*(\mathcal{P}_A) &\le \max_{i\in[N]} R^*\big(\mathcal{P}_{A\cap B(\theta_i,r)}\big) + \log N \\ &\le \max_{i\in[N]} w\big(A\cap B(\theta_i,r)\big) + \log N \qquad \text{(by Lemma 11.19)} \\ &\le w_A(r) + \log N. \end{aligned}$$

Lower bound on R^*. First, $R^*(\mathcal{P}_A) \ge \mathrm{Red}(\mathcal{P}_A) \gtrsim r_N^2$ (by the Haussler–Opper lower bound). Second, for $r = r_w$ and any $\theta \in A$,

$$R^*(\mathcal{P}_A) \ge R^*\big(\mathcal{P}_{A\cap B(\theta,r)}\big) \ge w\big(A\cap B(\theta,r)\big) - \frac{r^2}{2} \qquad \text{(Lemma 11.19 and translation invariance).}$$

Hence $R^*(\mathcal{P}_A) \ge w_A(r) - r^2/2 \ge r^2/2$ at $r = r_w$. Combining these bounds leads to $R^*(\mathcal{P}_A) \asymp r_N^2 + r_w^2$.

11.6.4 Proof of Lemma 11.19

We first write out the Shtarkov sum.

Lemma 11.20 (Shtarkov sum for the Gaussian shift family)**.**

$$R^*(\mathcal{P}_A) = \log\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}}\exp\Big(-\frac{1}{2}d(x,A)^2\Big)\,dx, \qquad d(x,A) := \inf_{y\in A}\|x-y\|_2.$$

Proof. This follows directly from Theorem 11.4 by computing $\sup_{\theta\in A}\varphi_d(x-\theta)$, where $\varphi_d(u) = (2\pi)^{-d/2}e^{-\|u\|_2^2/2}$ is the $\mathcal{N}(0, I_d)$ density. The supremum over θ occurs at the closest point in A, producing the distance term. □

Using an auxiliary $Z \sim \mathcal{N}(0, I_d)$,

$$\begin{aligned} \int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}}\exp\Big(-\frac{1}{2}d(x,A)^2\Big)\,dx &= \mathbb{E}\Big[\exp\Big(\frac{1}{2}(\|Z\|_2^2 - \mathrm{dist}(Z,A)^2)\Big)\Big] \\ &= \mathbb{E}\Big[\exp\Big(\sup_{w\in A}\Big(\frac{\|Z\|_2^2}{2} - \frac{\|Z-w\|_2^2}{2}\Big)\Big)\Big] \\ &= \mathbb{E}\Big[\exp\Big(\sup_{w\in A}(\langle w, Z\rangle - \tfrac{1}{2}\|w\|_2^2)\Big)\Big]. \end{aligned}$$

Denote

$$f(z) := \sup_{w\in A}\big(\langle w, z\rangle - \tfrac{1}{2}\|w\|_2^2\big), \qquad \nu := \mathcal{N}(0, I_d).$$

Then Lemma 11.20 becomes $R^*(\mathcal{P}_A) = \log \mathbb{E}_{Z\sim\nu}[e^{f(Z)}]$.

Lower bound.

$$\begin{aligned}\log \mathbb{E}\big[e^{f(Z)}\big] &= \log \mathbb{E}\Big[\exp\big(\sup_{w\in A}(\langle w, Z\rangle - \tfrac{1}{2}\|w\|_2^2)\big)\Big] \\ &\geq \log \mathbb{E}\Big[\exp\big(\sup_{w\in A}\langle w, Z\rangle\big)\Big] - \frac{1}{2}\sup_{w\in A}\|w\|_2^2 \\ &\geq \mathbb{E}\Big[\sup_{w\in A}\langle w, Z\rangle\Big] - \frac{1}{2}\sup_{w\in A}\|w\|_2^2 \qquad \text{(Jensen)} \\ &= w(A) - \frac{1}{2}\sup_{w\in A}\|w\|_2^2.\end{aligned}$$

Upper bound. By Gibbs' variational principle,

$$\log \mathbb{E}_{Z\sim\nu}[e^{f(Z)}] = \sup_{\mu}\Big\{\mathbb{E}_{Z\sim\mu}[f(Z)] - D_{\mathrm{KL}}(\mu\|\nu)\Big\}.$$

Using Talagrand's T_2 inequality for $\nu = \mathcal{N}(0, I_d)$, $W_2^2(\mu,\nu) \leq 2D_{\mathrm{KL}}(\mu\|\nu)$, we get

$$\log \mathbb{E}_{Z\sim\nu}[e^{f(Z)}] \leq \sup_{\mu}\Big\{\mathbb{E}_{Z\sim\mu}[f(Z)] - \frac{1}{2}W_2^2(\mu,\nu)\Big\}.$$

By Kantorovich duality for quadratic cost, $\frac{1}{2}W_2^2(\mu,\nu) = \sup_g\{\mathbb{E}_\mu g + \mathbb{E}_\nu g^c\}$, where $g^c(z) = \inf_x\{\frac{1}{2}\|x - z\|_2^2 - g(x)\}$. Thus

$$\begin{aligned}\log \mathbb{E}_{Z\sim\nu}[e^{f(Z)}] &\leq \sup_{\mu}\Big\{\mathbb{E}_\mu[f(Z)] - \sup_g(\mathbb{E}_\mu g + \mathbb{E}_\nu g^c)\Big\} \\ &\leq \mathbb{E}_{Z\sim\nu}\Big[\sup_x\big(f(x) - \tfrac{1}{2}\|x - Z\|_2^2\big)\Big].\end{aligned}$$

On the other hand,

$$\begin{aligned}\sup_x\Big(f(x) - \frac{1}{2}\|x - z\|_2^2\Big) &= \sup_x\sup_{w\in A}\Big(\langle w, x\rangle - \frac{1}{2}\|w\|_2^2 - \frac{1}{2}\|x - z\|_2^2\Big) \\ &= \sup_{w\in A}\langle w, z\rangle.\end{aligned}$$

Therefore

$$\log \mathbb{E}_{Z\sim\nu}[e^{f(Z)}] \leq \mathbb{E}\Big[\sup_{w\in A}\langle w, Z\rangle\Big] = w(A).$$

Combining the lower and upper bounds proves Lemma 11.19.

11.6.5 Alternative proof of the upper bound via convex geometry (Mourtada, 2023)

(An alternative proof in (Mourtada, 2023), using convex geometry.)

Definition 11.21 (Mixed volume)**.** Let $K_1, \dots, K_r$ be convex bodies in $\mathbb{R}^d$. Write

$$\mathrm{Vol}_d(\lambda_1 K_1 + \dots + \lambda_r K_r) = \sum_{j_1,\dots,j_d=1}^{r} V(K_{j_1}, \dots, K_{j_d})\, \lambda_{j_1} \cdots \lambda_{j_d}.$$

The quantity $V(K_{j_1}, \dots, K_{j_d})$ is called the *mixed volume.*

Definition 11.22 (Intrinsic volume)**.** Let $B \subset \mathbb{R}^d$ be the unit Euclidean ball. For $j \in \{0, 1, \dots, d\}$, define

$$V_j(K) := \binom{d}{j} \frac{V(\underbrace{K, \dots, K}_{j}, \underbrace{B, \dots, B}_{d-j})}{\kappa_{d-j}},$$

where $\kappa_m := \frac{\pi^{m/2}}{\Gamma(m/2+1)}$ is the volume of the unit ball in $\mathbb{R}^m$.

Theorem 11.23 (Steiner formula)**.**

$$\mathrm{Vol}_d(K + tB) = \sum_{j=0}^{d} V_{d-j}(K)\, \kappa_j\, t^j.$$

Theorem 11.24 (Alexandrov–Fenchel)**.** *For convex bodies $K_1, \dots, K_d$,*

$$V(K_1, K_2, K_3, \dots, K_d)^2 \geq V(K_1, K_1, K_3, \dots, K_d)\, V(K_2, K_2, K_3, \dots, K_d).$$

Remark 11.25 (A corollary). By choosing $(K_1, \dots, K_d) = (K, B, \underbrace{K, \dots, K}_{j-1}, \underbrace{B, \dots, B}_{d-j-1})$, we get

$$j\, V_j(K)^2 \geq (j+1)\, V_{j+1}(K)\, V_{j-1}(K).$$

In particular,

$$V_j(K) \leq \frac{V_1(K)^j}{j!}.$$

Back to the proof of the upper bound. Since $R^*(\mathcal{P}_A) \leq R^*(\mathcal{P}_{\mathrm{conv}(A)})$ and $w(A) = w(\mathrm{conv}(A))$, we may assume without loss of generality that $A = K$ is convex. Then

$$\begin{aligned}
\int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2} d(x,K)^2\Big)\, dx &= \int_0^\infty \mathrm{Vol}_d\Big(\big\{x \in \mathbb{R}^d : e^{-\frac{1}{2} d(x,K)^2} \geq t\big\}\Big)\, dt \\
&= \int_0^\infty \mathrm{Vol}_d\big(\{x \in \mathbb{R}^d : d(x,K) \leq r\}\big)\, r e^{-r^2/2}\, dr \\
&= \int_0^\infty \mathrm{Vol}_d(K + rB)\, r e^{-r^2/2}\, dr \\
&= \int_0^\infty \sum_{j=0}^{d} V_{d-j}(K)\, \kappa_j\, r^j\, r e^{-r^2/2}\, dr \\
&= \sum_{j=0}^{d} V_{d-j}(K)\, (2\pi)^{j/2},
\end{aligned}$$

where we used $\int_0^\infty r^{j+1} e^{-r^2/2}\, dr = 2^{j/2}\Gamma(\frac{j}{2}+1)$ so that $\kappa_j \int_0^\infty r^{j+1} e^{-r^2/2}\, dr = (2\pi)^{j/2}$.

Therefore,

$$\begin{aligned} R^*(\mathcal{P}_K) &= \log \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\Big(-\frac{1}{2} d(x,K)^2\Big)\, dx \\ &= \log \sum_{j=0}^{d} V_j(K)\,(2\pi)^{-j/2} = \log \sum_{j=0}^{d} V_j\Big(\frac{K}{\sqrt{2\pi}}\Big). \end{aligned}$$

The last quantity is called the *Wills functional*. Using the corollary above,

$$R^*(\mathcal{P}_K) \le \log \sum_{j=0}^{d} \frac{V_1(K/\sqrt{2\pi})^j}{j!} < \log \exp\big(V_1(K/\sqrt{2\pi})\big) = V_1(K/\sqrt{2\pi}) = w(K).$$

This recovers the upper bound in Lemma 11.19.

Lecture 12: Strong Data Processing Inequalities

12.1 Recall: DPI and SDPI

We recall the standard data processing inequality (DPI) and its “strong” variant.

Consider a channel $P_{Y|X}$, and two possible input distributions P_X and Q_X. Let the induced output distributions be

$$P_Y = P_X P_{Y|X}, \qquad Q_Y = Q_X P_{Y|X}.$$

$$\begin{tikzcd}
P_X \arrow[r] & P_{Y|X} \arrow[r] \arrow[dr] & P_Y \\
Q_X \arrow[ur] & & Q_Y
\end{tikzcd}$$

The (relative-entropy) DPI states

$$D_{\mathrm{KL}}\big(Q_Y \,\|\, P_Y\big) \le D_{\mathrm{KL}}\big(Q_X \,\|\, P_X\big).$$

A strong data processing inequality (SDPI) is a contraction version:

$$D_{\mathrm{KL}}\big(Q_Y \,\|\, P_Y\big) \le \eta(P_{Y|X})\, D_{\mathrm{KL}}\big(Q_X \,\|\, P_X\big) \qquad \text{for some } \eta(P_{Y|X}) < 1.$$

12.2 Input-independent SDPI

12.2.1 Definition

Definition 12.1 (Input-independent SDPI constant)**.** Given a channel $P_{Y|X}$, define

$$\eta(P_{Y|X}) \coloneqq \sup_{P_X \neq Q_X} \frac{D_{\mathrm{KL}}\big(Q_Y \,\|\, P_Y\big)}{D_{\mathrm{KL}}\big(Q_X \,\|\, P_X\big)}.$$

12.2.2 Mutual-information characterization

Proposition 12.2. *For any channel $P_{Y|X}$,*

$$\eta(P_{Y|X}) = \sup_{U - X - Y} \frac{I\big(U;Y\big)}{I\big(U;X\big)}.$$

Proof. We prove both directions.

($\geq$) Fix any Markov chain $U - X - Y$. Recall the identity

$$I\big(U;Y\big) = \mathbb{E}_U\Big[D_{\mathrm{KL}}\big(P_{Y|U} \,\|\, P_Y\big)\Big].$$

By the definition of $\eta(P_{Y|X})$, for each u we have

$$D_{\mathrm{KL}}\big(P_{Y|U=u} \,\|\, P_Y\big) \leq \eta(P_{Y|X})\, D_{\mathrm{KL}}\big(P_{X|U=u} \,\|\, P_X\big),$$

and therefore

$$I\big(U;Y\big) = \mathbb{E}_U\Big[D_{\mathrm{KL}}\big(P_{Y|U} \,\|\, P_Y\big)\Big] \leq \mathbb{E}_U\Big[\eta(P_{Y|X})\, D_{\mathrm{KL}}\big(P_{X|U} \,\|\, P_X\big)\Big] = \eta(P_{Y|X})\, I\big(U;X\big).$$

Hence $I\big(U;Y\big)/I\big(U;X\big) \leq \eta(P_{Y|X})$, and taking the supremum over $U-X-Y$ yields $\sup_{U-X-Y} I\big(U;Y\big)/I\big(U;X\big) \leq \eta(P_{Y|X})$.

($\leq$) Choose $U \sim \mathrm{Bern}(p)$ and two (fixed) distributions $\widetilde{P}_X, \widetilde{Q}_X$. Set

$$P_{X|U=1} = \widetilde{P}_X, \qquad P_{X|U=0} = \widetilde{Q}_X, \qquad P_X = p\widetilde{P}_X + (1-p)\widetilde{Q}_X.$$

Then

$$\begin{aligned} I\big(U;X\big) &= \mathbb{E}_U\Big[D_{\mathrm{KL}}\big(P_{X|U} \,\|\, P_X\big)\Big] \\ &= p\, D_{\mathrm{KL}}\big(\widetilde{P}_X \,\|\, p\widetilde{P}_X + (1-p)\widetilde{Q}_X\big) + (1-p)\, D_{\mathrm{KL}}\big(\widetilde{Q}_X \,\|\, p\widetilde{P}_X + (1-p)\widetilde{Q}_X\big). \end{aligned}$$

Differentiate at $p=0$ (this is the step shown in the notes):

$$\frac{\mathrm{d}}{\mathrm{d}p} I\big(U;X\big)\Big|_{p=0} = D_{\mathrm{KL}}\big(\widetilde{P}_X \,\|\, \widetilde{Q}_X\big) + \mathbb{E}_{\widetilde{Q}_X}\Big[\frac{\widetilde{P}_X - \widetilde{Q}_X}{\widetilde{Q}_X}\Big] = D_{\mathrm{KL}}\big(\widetilde{P}_X \,\|\, \widetilde{Q}_X\big).$$

Hence

$$I\big(U;X\big) = p\, D_{\mathrm{KL}}\big(\widetilde{P}_X \,\|\, \widetilde{Q}_X\big) + o(p).$$

Let $\widetilde{P}_Y, \widetilde{Q}_Y$ be the corresponding output distributions induced by $\widetilde{P}_X, \widetilde{Q}_X$ through $P_{Y|X}$. By the same reasoning,

$$I\big(U;Y\big) = p\, D_{\mathrm{KL}}\big(\widetilde{P}_Y \,\|\, \widetilde{Q}_Y\big) + o(p).$$

Therefore,

$$\frac{I\big(U;Y\big)}{I\big(U;X\big)} \to \frac{D_{\mathrm{KL}}\big(\widetilde{P}_Y \,\|\, \widetilde{Q}_Y\big)}{D_{\mathrm{KL}}\big(\widetilde{P}_X \,\|\, \widetilde{Q}_X\big)} \qquad \text{as } p \to 0^+.$$

Taking a supremum over $\widetilde{P}_X, \widetilde{Q}_X$ gives $\eta(P_{Y|X}) \leq \sup_{U-X-Y} I\big(U;Y\big)/I\big(U;X\big)$. □

12.2.3 Binary reduction

Proposition 12.3.

$$\eta(P_{Y|X}) = \sup_{P_X, Q_X \ \textit{binary}} \frac{D_{\mathrm{KL}}\big(Q_Y \,\|\, P_Y\big)}{D_{\mathrm{KL}}\big(Q_X \,\|\, P_X\big)},$$

where "binary" means P_X and Q_X are supported on at most two points of $\mathcal{X}$.

Proof. It suffices to show that for any $\gamma > 0$ and any pair (P_X, Q_X), defining

$$f(P_X, Q_X) := D_{\mathrm{KL}}\big(Q_Y \,\|\, P_Y\big) - \gamma\, D_{\mathrm{KL}}\big(Q_X \,\|\, P_X\big),$$

we can always find binary distributions $(\widehat{P}_X, \widehat{Q}_X)$ such that $f(P_X, Q_X) \le f(\widehat{P}_X, \widehat{Q}_X)$.

To prove it, consider the map

$$\widehat{P} \mapsto f\Big(\widehat{P}, \frac{Q_X}{P_X}\widehat{P}\Big) = D_{\mathrm{KL}}\big(P_{Y|X} \cdot \frac{Q_X}{P_X}\widehat{P} \,\|\, P_{Y|X} \cdot \widehat{P}\big) - \gamma\, D_{\mathrm{KL}}\big(\frac{Q_X}{P_X}\widehat{P} \,\|\, \widehat{P}\big).$$

This map is convex over the set

$$\Big\{\widehat{P} : \sum_x \frac{Q_X(x)}{P_X(x)}\widehat{P}(x) = 1,\ \sum_x \widehat{P}(x) = 1\Big\}.$$

When $\widehat{P} = P_X$, its value is $f(P_X, Q_X)$. A maximizer $\widehat{P}^*$ of a convex function over this polytope must lie at an extreme point. Extreme points here correspond to distributions supported on at most two atoms (hence binary). Letting $\widehat{Q}_X = \frac{Q_X}{P_X}\widehat{P}_X$ gives the desired binary pair. □

12.2.4 Characterization via Le Cam divergence and Hellinger diameter

Proposition 12.4.

$$\eta(P_{Y|X}) = \sup_{x,x' \in \mathcal{X}} \mathrm{LC}_{\max}\big(P_{Y|X=x}, P_{Y|X=x'}\big),$$

where

$$\mathrm{LC}_{\max}(P, Q) := \sup_{0<\beta<1} \beta(1-\beta) \int \frac{(p-q)^2}{(1-\beta)p + \beta q}\, \mathrm{d}\mu,$$

and p, q are densities of P, Q with respect to a common dominating measure μ. In particular, if

$$\mathrm{diam}_H^2(P_{Y|X}) := \sup_{x,x' \in \mathcal{X}} H^2\big(P_{Y|X=x}, P_{Y|X=x'}\big),$$

then

$$\frac{1}{2}\mathrm{diam}_H^2(P_{Y|X}) \le \eta(P_{Y|X}) \le \mathrm{diam}_H^2(P_{Y|X}) - \frac{\mathrm{diam}_H^4(P_{Y|X})}{4}.$$

Proof. The first claim (the Le Cam characterization) follows from the binary reduction above and explicit computations for binary input distributions; see the textbook for details. We prove the stated Hellinger bounds.

Lower bound. Fix two distributions P, Q (with densities p, q). For $\beta = 1/2$,

$$\mathrm{LC}_{\max}(P, Q) \ge \frac{1}{4} \int \frac{(p-q)^2}{\frac{1}{2}(p+q)}\, \mathrm{d}\mu = \frac{1}{2} \int \frac{(p-q)^2}{p+q}\, \mathrm{d}\mu.$$

Now note that $(p-q)^2 = (\sqrt{p} - \sqrt{q})^2(\sqrt{p} + \sqrt{q})^2 \ge (\sqrt{p} - \sqrt{q})^2(p+q)$. Hence

$$\frac{1}{2} \int \frac{(p-q)^2}{p+q}\, \mathrm{d}\mu \ge \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2\, \mathrm{d}\mu = \frac{1}{2} H^2(P, Q).$$

Taking the supremum over x, x' yields $\eta(P_{Y|X}) \ge \frac{1}{2}\mathrm{diam}_H^2(P_{Y|X})$.

Upper bound. For any $0 < \beta < 1$, one can check the identity

$$1 - \beta(1-\beta)\int \frac{(p-q)^2}{(1-\beta)p+\beta q}\,\mathrm{d}\mu = \int \frac{pq}{(1-\beta)p+\beta q}\,\mathrm{d}\mu.$$

Using Cauchy–Schwarz,

$$\int \frac{pq}{(1-\beta)p+\beta q}\,\mathrm{d}\mu = \int \frac{(\sqrt{pq})^2}{(1-\beta)p+\beta q}\,\mathrm{d}\mu \geq \frac{\left(\int \sqrt{pq}\,\mathrm{d}\mu\right)^2}{\int((1-\beta)p+\beta q)\,\mathrm{d}\mu} = \left(\int \sqrt{pq}\,\mathrm{d}\mu\right)^2.$$

Therefore

$$\beta(1-\beta)\int \frac{(p-q)^2}{(1-\beta)p+\beta q}\,\mathrm{d}\mu \leq 1 - \left(\int \sqrt{pq}\,\mathrm{d}\mu\right)^2.$$

Since $H^2(P,Q) = \int(\sqrt{p}-\sqrt{q})^2\,\mathrm{d}\mu = 2 - 2\int\sqrt{pq}\,\mathrm{d}\mu$, we have $\int\sqrt{pq}\,\mathrm{d}\mu = 1 - H^2(P,Q)/2$. Thus

$$\mathrm{LC}_{\max}(P,Q) \leq 1 - \left(1 - \frac{H^2(P,Q)}{2}\right)^2 = H^2(P,Q) - \frac{H^4(P,Q)}{4}.$$

Taking the supremum over x, x' yields the desired upper bound for $\eta(P_{Y|X})$. □

12.2.5 Examples and tensorization

Example 12.5 (Erasure channel)**.** Let EC_δ be the erasure channel with erasure probability δ, i.e.

$$P_{Y|X} = \begin{cases} Y = X, & \text{w.p. } 1-\delta, \\ Y = ?, & \text{w.p. } \delta. \end{cases}$$

Then (as shown in HW1) for all $U - X - Y$,

$$I\big(U;Y\big) = (1-\delta)\,I\big(U;X\big).$$

Therefore, $\eta(\mathrm{EC}_\delta) = 1 - \delta$.

Example 12.6 (Binary symmetric channel)**.** Let BSC_δ be the binary symmetric channel with crossover probability δ: $X \in \{0,1\}$ and $Y = X \oplus \mathrm{Bern}(\delta)$. In this case,

$$\begin{aligned}\mathrm{LC}_{\max}(P_{Y|X=0}, P_{Y|X=1}) &= \sup_{\beta\in(0,1)} \beta(1-\beta)\left(\frac{(1-2\delta)^2}{(1-\beta)(1-\delta)+\beta\delta} + \frac{(1-2\delta)^2}{(1-\beta)\delta+\beta(1-\delta)}\right) \\ &= (1-2\delta)^2 \sup_{\beta\in(0,1)} \frac{\beta(1-\beta)}{\big((1-\beta)(1-\delta)+\beta\delta\big)\big((1-\beta)\delta+\beta(1-\delta)\big)}.\end{aligned}$$

Let $A = (1-\beta)(1-\delta)+\beta\delta,\ B = (1-\beta)\delta + \beta(1-\delta)$. Then $A + B = 1$ and one can compute $AB = \delta(1-\delta) + \beta(1-\beta)(1-2\delta)^2$. Hence

$$AB - \beta(1-\beta) = \delta(1-\delta)\big(1 - 4\beta(1-\beta)\big) \geq 0,$$

so $\beta(1-\beta)/(AB) \leq 1$, with equality at $\beta = 1/2$. Therefore

$$\eta(\mathrm{BSC}_\delta) = (1-2\delta)^2.$$

Example 12.7 (Tensorization bound)**.** For the n-fold product channel, one has

$$\eta\big(P_{Y|X}^{\otimes n}\big) \leq 1 - (1 - \eta(P_{Y|X}))^n.$$

Proof. Let $U - X^n - Y^n$. Write $Y^n = (Y_1, Y_2^n)$. Then

$$\begin{aligned} I\big(U; Y^n\big) &= I\big(U; Y_2^n\big) + I\big(U; Y_1 \mid Y_2^n\big) \\ &\le I\big(U; Y_2^n\big) + \eta(P_{Y|X})\, I\big(U; X_1 \mid Y_2^n\big) \\ &= (1 - \eta(P_{Y|X}))\, I\big(U; Y_2^n\big) + \eta(P_{Y|X})\, I\big(U; X_1, Y_2^n\big) \\ &\le (1 - \eta(P_{Y|X}))\, I\big(U; Y_2^n\big) + \eta(P_{Y|X})\, I\big(U; X^n\big). \end{aligned}$$

Iterating this decomposition gives

$$\frac{I\big(U; Y^n\big)}{I\big(U; X^n\big)} \le \eta(P_{Y|X}) \sum_{t=0}^{n-1} (1 - \eta(P_{Y|X}))^t = 1 - (1 - \eta(P_{Y|X}))^n.$$

Taking the supremum over $U - X^n - Y^n$ proves the claim. □

Remark 12.8. (A general result recorded in the notes.) In a Bayesian network, suppose each vertex v is declared "open" with probability $\eta\big(P_{X_v|\mathrm{Pa}(v)}\big)$. Then for S a set of vertices,

$$\eta(P_{X_S|X_0}) \le \mathbb{P}(\text{there exists an open path from 0 to some vertex in } S),$$

which is a "percolation" probability from 0 to S.

12.3 Input-dependent SDPI

12.3.1 Definition

Definition 12.9 (Input-dependent SDPI constant)**.** Given a channel $P_{Y|X}$ and an input distribution P_X, define

$$\eta(P_X, P_{Y|X}) \coloneqq \sup_{Q_X} \frac{D_{\mathrm{KL}}\big(Q_Y \,\|\, P_Y\big)}{D_{\mathrm{KL}}\big(Q_X \,\|\, P_X\big)}.$$

12.3.2 Properties

Proposition 12.10. *(1)*

$$\eta(P_X, P_{Y|X}) = \sup_{U - X - Y} \frac{I\big(U; Y\big)}{I\big(U; X\big)}.$$

(2) (Tensorization)

$$\eta(P_X^{\otimes n}, P_{Y|X}^{\otimes n}) = \eta(P_X, P_{Y|X}).$$

Proof. The mutual-information characterization is analogous to the input-independent case. We prove the tensorization statement.

By induction, it suffices to prove the case $n = 2$. Let $U - (X_1, X_2) - (Y_1, Y_2)$ under the product channel. Then

$$\begin{aligned} I\big(U; Y_1, Y_2\big) &= I\big(U; Y_1\big) + I\big(U; Y_2 \mid Y_1\big) \\ &\le \eta\big(I\big(U; X_1\big) + I\big(U; X_2 \mid Y_1\big)\big), \end{aligned}$$

where $\eta = \eta(P_X, P_{Y|X})$ and the second inequality uses that $(U, Y_1) - X_2 - Y_2$ and $P_{X_2|Y_1} = P_{X_2} = P_X$. Now expand

$$\begin{aligned} I\big(U; X_2 \mid Y_1\big) &= I\big(U; X_2 \mid X_1, Y_1\big) + I\big(X_1; X_2 \mid Y_1\big) - I\big(X_1; X_2 \mid Y_1, U\big) \\ &= I\big(U; X_2 \mid X_1\big) + 0 - I\big(X_1; X_2 \mid Y_1, U\big) \\ &\le I\big(U; X_2 \mid X_1\big), \end{aligned}$$

where $I\big(X_1; X_2 \mid Y_1\big) = 0$ because X_2 is independent of (X_1, Y_1) under the product input. Therefore,

$$I\big(U; Y_1, Y_2\big) \le \eta\big(I\big(U; X_1\big) + I\big(U; X_2 \mid X_1\big)\big) = \eta\, I\big(U; X_1, X_2\big).$$

Taking the supremum over $U - (X_1, X_2) - (Y_1, Y_2)$ yields $\eta(P_X^{\otimes 2}, P_{Y|X}^{\otimes 2}) \le \eta(P_X, P_{Y|X})$. The reverse inequality is immediate by restricting to product auxiliaries, hence equality holds. □

Remark 12.11. Unlike $\eta(P_{Y|X})$, the input-dependent SDPI constant $\eta(P_X, P_{Y|X})$ can be much more challenging to characterize. An example is when $P_{Y|X}$ is the transition matrix of a Markov chain and $P_X = \pi$ is its stationary distribution. Then SDPI implies (for all initial distributions π_0)

$$D_{\mathrm{KL}}\big(\pi_0 P^n \,\|\, \pi\big) = D_{\mathrm{KL}}\big(\pi_0 P^n \,\|\, \pi P^n\big) \le \eta(\pi, P)^n\, D_{\mathrm{KL}}\big(\pi_0 \,\|\, \pi\big).$$

This is called the modified log-Sobolev inequality and leads to upper bounds on mixing times; both tasks can be challenging for general Markov chains.

12.3.3 Example: jointly Gaussian pair

Let (X, Y) be jointly Gaussian with zero mean and covariance

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Claim 12.12.

$$\eta(P_X, P_{Y|X}) = \eta(P_Y, P_{X|Y}) = \rho^2.$$

Proof. We only prove the upper bound $\eta(P_X, P_{Y|X}) \le \rho^2$; the notes indicate a matching lower bound is obtained later.

Step 1: scaling. By scaling, we may assume

$$Y = X + Z, \qquad Z \sim N(0, \rho^{-2} - 1), \qquad X \sim N(0, 1), \qquad Y \sim N(0, \rho^{-2}).$$

Step 2: relate KL to entropy and second moment. For any random variable $\widetilde{X}$ and $\widetilde{Y} = \widetilde{X} + Z$,

$$\begin{aligned} D_{\mathrm{KL}}\big(P_{\widetilde{Y}} \,\|\, P_Y\big) &= -h(\widetilde{Y}) + \log \frac{\sqrt{2\pi}}{\rho} + \frac{\rho^2}{2}\, \mathbb{E}[\widetilde{Y}^2] \\ &\le -\frac{1}{2} \log\Big(2\pi e(\rho^{-2} - 1) + e^{2h(\widetilde{X})}\Big) + \log \frac{\sqrt{2\pi}}{\rho} + \frac{\rho^2}{2}\, \mathbb{E}[\widetilde{Y}^2], \end{aligned}$$

where the inequality is the entropy power inequality (EPI) applied to $\widetilde{Y} = \widetilde{X} + Z$. Also

$$D_{\mathrm{KL}}\big(P_{\widetilde{X}} \,\|\, P_X\big) = -h(\widetilde{X}) + \log \sqrt{2\pi} + \frac{1}{2}\mathbb{E}[\widetilde{X}^2].$$

Step 3: rearrange. Using $h(\widetilde{X}) = \log\sqrt{2\pi} + \frac{1}{2}\mathbb{E}[\widetilde{X}^2] - D_{\mathrm{KL}}\big(P_{\widetilde{X}} \,\|\, P_X\big)$, we have

$$e^{2h(\widetilde{X})} = 2\pi\, \exp\Big(\mathbb{E}[\widetilde{X}^2] - 2D_{\mathrm{KL}}\big(P_{\widetilde{X}} \,\|\, P_X\big)\Big).$$

Substitute this into the previous bound, and use $\mathbb{E}[\widetilde{Y}^2] = \mathbb{E}[\widetilde{X}^2] + (\rho^{-2} - 1)$. After simplifying (this is the algebra shown in the notes), we obtain

$$D_{\mathrm{KL}}\big(P_{\widetilde{Y}} \,\|\, P_Y\big) \le -\frac{1}{2}\log\Big(1 - \rho^2 + \rho^2\exp\big(\mathbb{E}[\widetilde{X}^2] - 2D_{\mathrm{KL}}\big(P_{\widetilde{X}} \,\|\, P_X\big) - 1\big)\Big) + \frac{\rho^2}{2}\big(\mathbb{E}[\widetilde{X}^2] - 1\big).$$

Step 4: concavity of log**.** Using concavity of log, for $x > 0$,

$$\log(1 - \rho^2 + \rho^2 x) \ge \rho^2 \log x \qquad (\text{equivalently, } \log(1 - p + px) \ge p\log x).$$

Apply this with $x = \exp(\mathbb{E}[\widetilde{X}^2] - 2D_{\mathrm{KL}}\big(P_{\widetilde{X}} \,\|\, P_X\big) - 1)$:

$$\begin{aligned} D_{\mathrm{KL}}\big(P_{\widetilde{Y}} \,\|\, P_Y\big) &\le -\frac{\rho^2}{2}\Big(\mathbb{E}[\widetilde{X}^2] - 2D_{\mathrm{KL}}\big(P_{\widetilde{X}} \,\|\, P_X\big) - 1\Big) + \frac{\rho^2}{2}\big(\mathbb{E}[\widetilde{X}^2] - 1\big) \\ &= \rho^2\, D_{\mathrm{KL}}\big(P_{\widetilde{X}} \,\|\, P_X\big). \end{aligned}$$

This shows the KL contraction factor is at most ρ^2, i.e. $\eta(P_X, P_{Y|X}) \le \rho^2$.

Lower bound via another SDPI constant. The notes introduce the χ^2-based SDPI constant

$$\eta_{\chi^2}(P_X, P_{Y|X}) \coloneqq \sup_{Q_X} \frac{\chi^2(Q_Y \,\|P_Y)}{\chi^2(Q_X \,\|P_X)},$$

with the following properties:

(1) $\eta_{\chi^2} \le \eta$ (KL dominates χ^2 in this SDPI sense).

(2) $\eta_{\chi^2} = \sigma_2(M)^2$, where $\sigma_1(M) \ge \sigma_2(M) \ge \cdots \ge 0$ are the singular values of

$$M_{x,y} = \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)P_Y(y)}}.$$

(3) $\sqrt{\eta_{\chi^2}}$ equals the maximal correlation between X and Y:

$$\sup_{g_1,g_2} \mathrm{corr}(g_1(X), g_2(Y)) = \sup_{g_1,g_2} \frac{\mathrm{Cov}(g_1(X), g_2(Y))}{\sqrt{\mathrm{Var}(g_1(X))\,\mathrm{Var}(g_2(Y))}}.$$

(4) In Markov chains,

$$\chi^2(\pi_0 P^n \,\|\pi) \le \eta_{\chi^2}(\pi, P)^n\, \chi^2(\pi_0 \,\|\pi),$$

which is Poincaré's inequality.

By (1) and (3), for a jointly Gaussian pair (X, Y), $\eta \ge \eta_{\chi^2} = (\text{maximal correlation})^2 = \rho^2$. Combined with the upper bound, this yields $\eta = \rho^2$. □

12.4 Applications of SDPI

12.4.1 Example 1: noisy gates

Suppose a noisy gate is an $\{\mathrm{AND}, \mathrm{OR}, \mathrm{NOT}\}$ gate with output corrupted by a $\mathrm{Bern}(\delta)$ noise. A natural question is:

Question. For every $\delta < 1/2$, can we still reliably compute all Boolean functions $\{0,1\}^n \to \{0,1\}$?

Claim 12.13. *For each input bit X_i,*

$$I\big(X_i; Y\big) \le \big(2(1-2\delta)^2\big)^{d_i},$$

where d_i is the minimum (graph) distance from X_i to the output Y.

Answer to the question (as in the notes): No. Suppose we'd like to compute

$$\mathrm{XOR}(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i \bmod 2.$$

Then there exists $i \in [n]$ with $d_i \ge \log_2 n$. For this i, if

$$\delta > \frac{1}{2} - \frac{1}{2\sqrt{2}} \approx 0.15,$$

then $2(1-2\delta)^2 < 1$ and

$$I\big(X_i; Y\big) \le \big(2(1-2\delta)^2\big)^{\log_2 n} \to 0 \qquad (n \to \infty).$$

Since $\mathrm{XOR}(X_1, \dots, X_n)$ is sensitive to every X_i, its computation is impossible in this noise regime.

Proof of the claim. As written in the notes,

$$I\big(X_i; Y\big) \le \eta(P_{Y|X_i})\, H(X_i) \le \eta(P_{Y|X_i}).$$

Using the percolation interpretation of η for Bayesian networks,

$$\eta(P_{Y|X_i}) \le (\text{percolation probability from } X_i \text{ to } Y) = \sum_{\text{paths } X_i \to Y} (1-2\delta)^{2\,\mathrm{length(path)}}.$$

When $\mathrm{length(path)} \ge d_i$ and $2(1-2\delta)^2 \le 1$, this sum is bounded by $\big(2(1-2\delta)^2\big)^{d_i}$. □

12.4.2 Example 2: broadcast on trees

Let $(\pi, P_{X'|X})$ be a reversible Markov chain. Consider the broadcasting problem on an infinite b-ary tree. The root is $X_0 \sim \pi$, and each edge transmits the parent state through the same channel $P_{X'|X}$.

Question. Given all variables on level D (denote the set of vertices at level D by L_D), as $D \to \infty$, can you recover X_0 reliably?

Claim 12.14. *No if*

$$b\,\eta(\pi, P_{X'|X}) < 1.$$

Proof. Let $X_{L_D} = (X_v)_{v\in L_D}$. The notes argue

$$\begin{aligned} I\big(X_0; X_{L_D}\big) &\le \sum_{v\in L_1} I\big(X_0; X_{L_{D,v}}\big), \qquad L_{D,v} \coloneqq \{u \in L_D : v \text{ is an ancestor of } u\} \\ &\le \eta(\pi, P_{X'|X}) \sum_{v\in L_1} I\big(X_v; X_{L_{D,v}}\big) \qquad (X_{L_{D,v}} \to X_v \to X_0, \text{ and reversibility}) \\ &= b\,\eta(\pi, P_{X'|X})\, I\big(X_0; X_{L_{D-1}}\big). \end{aligned}$$

Iterating gives

$$I\big(X_0; X_{L_D}\big) \le \big(b\,\eta(\pi, P_{X'|X})\big)^D H(X_0) \to 0 \qquad (D\to\infty)$$

whenever $b\,\eta(\pi, P_{X'|X}) < 1$. □

12.4.3 Application: stochastic block model

In the 2-SBM$(a/n, b/n)$, a label vector $X \sim \mathrm{Unif}(\{\pm 1\}^n)$ is drawn and edges are generated conditionally independently as

$$\mathbb{P}\big((i,j) \text{ is connected} \mid X\big) = \begin{cases} \frac{a}{n}, & X_iX_j = 1 \quad \text{(same community)}, \\ \frac{b}{n}, & X_iX_j = -1 \quad \text{(different community)}. \end{cases}$$

Question. When can we recover $X_1, X_2 \in \{\pm 1\}$ with nontrivial probability as $n\to\infty$?

Claim 12.15. *We cannot if*

$$\frac{(a-b)^2}{2(a+b)} < 1 \qquad \textit{(Kesten–Stigum threshold).}$$

Proof. As written in the notes: since all edge probabilities are of order $\Theta(1/n)$, with high probability the neighborhood of a vertex out to distance d is a tree (no cycles) for some $d = d_n \to \infty$. Moreover, the number of children is approximately $\mathrm{Poi}((a+b)/2)$, and the label flipping probability along an edge is $\frac{b}{a+b}$. With high probability, vertex 2 does not belong to the local neighborhood of vertex 1, so

$$\begin{aligned} I\big(X_1; X_2 \mid G\big) &\le I\big(X_1; (X_i)_{i\in L_d} \mid G\big) \\ &\le \Big(\frac{a+b}{2}\big(1 - 2\tfrac{b}{a+b}\big)^2\Big)^d \qquad \text{(see HW3 for details)} \\ &= \Big(\frac{(a-b)^2}{2(a+b)}\Big)^d \to 0 \qquad \text{if } \frac{(a-b)^2}{2(a+b)} < 1. \end{aligned}$$

□

12.4.4 Example 3: spiked Wigner model

Let $X \sim \mathrm{Unif}(\{\pm 1\}^n)$ be unknown. Observe a noisy rank-one matrix

$$Y = \frac{\lambda}{\sqrt{n}} XX^\top + W, \qquad (W_{ij} = W_{ji} \sim N(0,1) \text{ i.i.d.}).$$

Claim 12.16. *If $\lambda < 1$, then*

$$I\big(X_1; X_2 \mid Y\big) = o(1),$$

i.e. weak recovery of X is impossible. (The threshold $\lambda \le 1$ is the BBP transition.)

Proof. The idea is that Y_{ij} is determined by X_iX_j through

$$Y_{ij} \mid (X_iX_j) \sim N\Big(\sqrt{\frac{\lambda}{n}}\, X_iX_j, 1\Big).$$

Let $\theta_{ij} \in \{\pm 1\}$ denote X_iX_j. For the Gaussian binary-input channel $P = \{N(\sqrt{\lambda/n}, 1),\ N(-\sqrt{\lambda/n}, 1)\}$, one can show

$$\eta := \eta(P) = \mathrm{LC}_{\chi^2}\big(N(\sqrt{\lambda/n}, 1), N(-\sqrt{\lambda/n}, 1)\big) = \frac{\lambda}{n}(1 + o(1)).$$

Next, replace Y_{ij} by an erasure variable Z_{ij} defined by

$$Z_{ij} \mid \theta_{ij} = \begin{cases} \theta_{ij}, & \text{w.p. } \eta, \\ ?, & \text{w.p. } 1-\eta, \end{cases} \qquad \text{i.e. EC}(1-\eta).$$

Then for any $U \to \theta_{ij} \to (Y_{ij}, Z_{ij})$,

$$I\big(U; Y_{ij}\big) \le \eta\, I\big(U; \theta_{ij}\big) = I\big(U; Z_{ij}\big).$$

We claim that

$$I\big(X_1; Y \mid X_2\big) \le I\big(X_1; Z \mid X_2\big). \tag{$*$}$$

Assuming $(*)$,

$$\begin{aligned} I\big(X_1; X_2 \mid Y\big) &= I\big(X_1; X_2, Y\big) && \big(I\big(X_1; Y\big) = 0\big) \\ &= I\big(X_1; Y \mid X_2\big) && \big(I\big(X_1; X_2\big) = 0\big) \\ &\le I\big(X_1; Z \mid X_2\big) && (\text{by } (*)) \\ &= I\big(X_1; X_2 \mid Z\big) \\ &\le \mathbb{P}(1 \text{ and } 2 \text{ are connected in the graph induced by } Z), \end{aligned}$$

where the induced graph has an edge (i, j) iff $Z_{ij} \neq ?$. Since this graph is Erdős–Rényi with edge probability $\eta = \frac{\lambda}{n}(1 + o(1))$, it is known that when $\lambda < 1$, the largest connected component has size $O(\log n)$. Therefore $\mathbb{P}(1 \text{ and } 2 \text{ connected}) \to 0$, giving the claim.

Proof of $(*)$. Write $Y = (Y_1, Y_2)$ where Y_1 corresponds to some subset of entries and Y_2 to the remaining ones. Then

$$\begin{aligned} I\big(X_1; Y \mid X_2\big) &= I\big(X_1; Y_1 \mid X_2\big) + I\big(X_1; Y_2 \mid X_2, Y_1\big) \\ &\le I\big(X_1; Y_1 \mid X_2\big) + \eta\, I\big(X_1; \theta_2 \mid X_2, Y_1\big) \\ &= I\big(X_1; Y_1 \mid X_2\big) + I\big(X_1; Z_2 \mid X_2, Y_1\big) \\ &= I\big(X_1; Y_1, Z_2 \mid X_2\big). \end{aligned}$$

Proceeding with the same argument entry-by-entry replaces all Y coordinates by their erasure counterparts, yielding $I\big(X_1; Y \mid X_2\big) \le I\big(X_1; Z \mid X_2\big)$. □

12.5 Example 4: proximal sampling

Suppose we would like to sample from

$$\pi(x) \propto e^{-f(x)}, \qquad f : \mathbb{R}^d \to \mathbb{R}.$$

A proximal sampler aims to sample from the joint density

$$\pi(x,y) \propto \exp\Big(- f(x) - \frac{1}{2\eta} \|x-y\|^2 \Big)$$

via an iterative procedure. Given initialization $X_0 \sim P_{X_0}$, for each $t = 0, 1, \dots$:

- Given X_t, sample $Y_t \mid X_t \sim N(X_t, \eta I)$.
- Given Y_t, sample $X_{t+1} \mid Y_t \sim \pi^{x|y}(\cdot \mid Y_t)$.

(For convex f, the conditional $\pi^{x|y}(\cdot \mid y)$ is η-strongly log-concave.)

Claim 12.17. *If π satisfies an α-log-Sobolev inequality (LSI), i.e.*

$$D_{\mathrm{KL}}\big(p \,\|\, \pi\big) \le \frac{1}{2\alpha}\,\mathrm{FI}(p \,\|\pi) := \frac{1}{2\alpha}\mathbb{E}_p\big[\, \|\nabla \log(p/\pi)\|^2 \,\big], \qquad \forall p,$$

then

$$D_{\mathrm{KL}}\big(P_{X_t} \,\|\, \pi\big) \le \frac{D_{\mathrm{KL}}\big(P_{X_0} \,\|\, \pi\big)}{(1+\alpha\eta)^{2t}}.$$

Proof. We show the two one-step contractions written in the notes:

$$D_{\mathrm{KL}}\big(P_{Y_t} \,\|\, \pi_\eta\big) \le \frac{D_{\mathrm{KL}}\big(P_{X_t} \,\|\, \pi\big)}{1+\alpha\eta}, \tag{1}$$

$$D_{\mathrm{KL}}\big(P_{X_{t+1}} \,\|\, \pi\big) \le \frac{D_{\mathrm{KL}}\big(P_{Y_t} \,\|\, \pi_\eta\big)}{1+\alpha\eta}, \tag{2}$$

where $\pi_\eta = \pi * N(0, \eta I)$. Iterating (1) and (2) yields the claimed rate. The notes explain these are equivalent to input-dependent SDPI bounds $\eta(\pi, N(\cdot, \eta I)) \le 1/(1+\alpha\eta)$ and $\eta(\pi_\eta, \pi^{x|y}(\cdot \mid y)) \le 1/(1+\alpha\eta)$.

Forward step. Let $p_t = P_{X_t}$ and $\pi_t = \pi$. Consider the heat flow

$$\partial_t p_t = \frac{1}{2}\Delta p_t, \qquad \partial_t \pi_t = \frac{1}{2}\Delta \pi_t.$$

Then $p_\eta = P_{Y_t}$ and $\pi_\eta = \pi * N(0, \eta I)$. Now compute (as in the notes):

$$\begin{aligned}
\partial_t D_{\mathrm{KL}}\big(p_t \,\|\, \pi_t\big) &= \partial_t \int p_t \log \frac{p_t}{\pi_t} \\
&= \frac{1}{2}\int \Delta p_t \Big(\log \frac{p_t}{\pi_t} + 1 \Big) - \frac{1}{2}\int \Delta \pi_t \frac{p_t}{\pi_t} \\
&= -\frac{1}{2}\int \nabla p_t \cdot \nabla \log \frac{p_t}{\pi_t} + \frac{1}{2}\int \nabla \pi_t \cdot \nabla \frac{p_t}{\pi_t} \\
&= -\frac{1}{2}\mathbb{E}_{p_t}\Big[\nabla \log p_t \cdot \nabla \log \frac{p_t}{\pi_t} \Big] + \frac{1}{2}\mathbb{E}_{p_t}\Big[\nabla \log \pi_t \cdot \nabla \log \frac{p_t}{\pi_t} \Big] \\
&= -\frac{1}{2}\,\mathrm{FI}(p_t \,\|\pi_t).
\end{aligned}$$

Since π is α-LSI, one can show that $\pi_t = \pi * N(0, tI)$ is $(\frac{1}{\alpha} + t)^{-1}$-LSI. Therefore

$$\partial_t D_{\mathrm{KL}}\big(p_t \,\|\, \pi_t\big) = -\frac{1}{2}\mathrm{FI}(p_t \,\|\pi_t) \le -\frac{1}{\frac{1}{\alpha} + t}\, D_{\mathrm{KL}}\big(p_t \,\|\, \pi_t\big).$$

Integrating from $t = 0$ to $t = \eta$ yields

$$\frac{D_{\mathrm{KL}}\big(p_\eta \,\|\, \pi_\eta\big)}{D_{\mathrm{KL}}\big(p_0 \,\|\, \pi_0\big)} \le \exp\Big(-\int_0^\eta \frac{1}{\frac{1}{\alpha}+t}\,\mathrm{d}t\Big) = \frac{1}{1+\alpha\eta},$$

which is (1).

Backward step. Let $p_0^- = P_{Y_t}$ and $\pi_0^- = \pi_\eta$. Consider the reverse-time evolution written in the notes:

$$\begin{aligned}
\partial_t p_t^- &= -\mathrm{div}(p_t^- \nabla \log \pi_t^-) + \frac{1}{2}\Delta p_t^- = \mathrm{div}\Big(p_t^- \nabla \log \frac{p_t^-}{\pi_t^-}\Big) - \frac{1}{2}\Delta p_t^-, \\
\partial_t \pi_t^- &= -\mathrm{div}(\pi_t^- \nabla \log \pi_t^-) + \frac{1}{2}\Delta \pi_t^- = -\frac{1}{2}\Delta \pi_t^-.
\end{aligned}$$

Then $p_\eta^- = P_{X_{t+1}}$ and $\pi_\eta^- = \pi$ ("by the reverse process of diffusion model"). A similar computation gives

$$\begin{aligned}
\partial_t D_{\mathrm{KL}}\big(p_t^- \,\|\, \pi_t^-\big) &= \partial_t \int p_t^- \log \frac{p_t^-}{\pi_t^-} \\
&= \int \Big(\mathrm{div}(p_t^- \nabla \log \tfrac{p_t^-}{\pi_t^-}) - \frac{1}{2}\Delta p_t^-\Big)\Big(\log \frac{p_t^-}{\pi_t^-} + 1\Big) - \int \Big(-\frac{1}{2}\Delta \pi_t^-\Big)\frac{p_t^-}{\pi_t^-} \\
&= -\int p_t^- \nabla \log \frac{p_t^-}{\pi_t^-} \cdot \nabla \log \frac{p_t^-}{\pi_t^-} + \frac{1}{2}\,\mathrm{FI}(p_t^- \,\|\pi_t^-) \\
&= -\frac{1}{2}\,\mathrm{FI}(p_t^- \,\|\pi_t^-).
\end{aligned}$$

Since $\pi_t^- = \pi_{\eta - t}$ is $(\frac{1}{\alpha} + \eta - t)^{-1}$-LSI,

$$\partial_t D_{\mathrm{KL}}\big(p_t^- \,\|\, \pi_t^-\big) \le -\frac{1}{\frac{1}{\alpha}+\eta-t}\,D_{\mathrm{KL}}\big(p_t^- \,\|\, \pi_t^-\big).$$

Integrating from $t = 0$ to $t = \eta$ yields

$$\frac{D_{\mathrm{KL}}\big(p_\eta^- \,\|\, \pi_\eta^-\big)}{D_{\mathrm{KL}}\big(p_0^- \,\|\, \pi_0^-\big)} \le \exp\Big(-\int_0^\eta \frac{1}{\frac{1}{\alpha}+\eta-t}\,\mathrm{d}t\Big) = \frac{1}{1+\alpha\eta},$$

which is (2). □

12.6 Special topic: SDPIs

Guest lecture by Y. Gu on SDPIs.