Information Theory Notes

Vasco Villas-Boas

January 14, 2025

Contents

1 Entropy and Mutual Information

1.1 Initial Definitions and Results

In this section we will introduce the concept of entropy of a distribution and the mutual information between distributions.

Definition 1 (Entropy)**.** For a discrete random variable $X \in \mathcal{X}$ with pmf p, its *entropy* $H(X)$ (or $H(p)$) is defined by

$$H(X) := \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{1}{p(x)} \right)$$

For continuous random variable random variable Y, with a measure μ such that $Y \in \mathcal{Y}$ has a density f with respect to μ, we define the *differential entropy*

$$h(Y) := \int_{\mathcal{Y}} f(y) \log \left(\frac{1}{f(y)} \right) d\mu(y)$$

Generally, for a random variable Z with probability distribution p, we write the entropy of Z as

$$H(Z) := \mathbb{E}_p \left[\log \left(\frac{1}{p(Z)} \right) \right]$$

To make some remarks, we note that

- $0 \leq H(X) \leq \log(|\mathcal{X}|)$.
- $H(X)$ can be finite or infinite when $|\mathcal{X}| = \infty$

Shannon (1948) shows that entropy characterizes the fundamental limit of source coding. The source coding problem (for the iid case) is given:

- An input alphabet $\mathcal{X}$ (eg., all English letters)

- A known pmf p on $\mathcal{X}$.

Target: find a map (ie., code) $f : \mathcal{X} \to \{0,1\}^*$, such that[1]

- It's *uniquely decodable*. Ie., based on the concatenation $(f(x_1), ..., f(x_m))$, one can uniquely decode m and $(x_1, ..., x_m) \in \mathcal{X}^m$.

- The expected code length $\mathbb{E}[l(f(X))] = \sum_{x\in\mathcal{X}} p(x) l(f(x))$ is minimized.

Example 2 (Code Examples)**.** If $\mathcal{X} = \{a, b, c\}$ and $p = (1/4, 1/2, 1/4)$:

(a) The code ($a \mapsto$ 0, $b \mapsto$ 10, $c \mapsto$ 11) is uniquely decodable.

(b) The code ($a \mapsto$ 0, $b \mapsto$ 1, $c \mapsto$ 10) is *not* uniquely decodable.

(c) The code ($a \mapsto$ 10, $b \mapsto$ 1, $c \mapsto$ 11) is uniquely decodable and has a smaller expected code length than the proposal in (a).

Theorem 3 (Kraft-McMillan)**.** Given a length profile $\{l_x\}_{x\in\mathcal{X}}$, there is a uniquely decodable code f with $l(f(x)) = l_x \; \forall x \in \mathcal{X}$ if and only if $\sum_{x\in\mathcal{X}} 2^{-l_x} \leq 1$.

Proof.

[$\Longleftarrow$]:

First note that for a full binary tree T (ie., a binary tree where each node had 0 or 2 children), then $\sum_{v\in leaf(T)} 2^{-depth(v)}$ $= 1$. Given the encoding f, one can construct a full binary tree T such that $\mathcal{X} \subseteq leaf(T)$ and $depth(x) = l_x \; \forall x \in \mathcal{X}$. Now use a prefix coding scheme that reports the path to the node from the root where 0 corresponds to going left and 1 corresponds to going right at a node to reach the leaf node of interest.

[$\Longrightarrow$]:

WLOG assume that $|\mathcal{X}| < \infty$ and $l_{max} := \max_{x\in\mathcal{X}} l_x < \infty$. Next, we use a tensor power trick for a uniquely

[1] We define $\{0,1\}^* := \cup_{n=1}^{\infty} \{0,1\}^n$

decodable code f

$$\begin{aligned}
\left(\sum_{x\in\mathcal{X}} 2^{-l(f(x))}\right)^m &= \sum_{\{x_1,...,x_m\}\subseteq 2^{\mathcal{X}}} 2^{-l(f(x_1))+...+l(f(x_m))}\\
&= \sum_{\{x_1,...,x_m\}\subseteq 2^{\mathcal{X}}} 2^{-\underbrace{l(f(x_1),...,f(x_m))}_{concatenation}}\\
&= \sum_{l=1}^{ml_{max}} 2^{-l}(\textit{number of concatenated codewords of total length } l)\\
&\leq \sum_{l=1}^{ml_{max}} 2^{-l}2^{l}, \textit{ by uniquely decodable assumption}\\
&= ml_{max}\\
\implies \sum_{x\in\mathcal{X}} 2^{-l(f(x))} &\leq (ml_{max})^{1/m} \overset{m\to\infty}{\to} 1
\end{aligned}$$

□

Using the Kraft-McMillan result, we obtain the following characterization of the smallest expected code length.

Theorem 4 (Source Coding Theorem for Uniquely Decodable Code). $H(X) \leq \min_{\textit{uniquely decodable } f} \mathbb{E}[l(f(X))] < H(X)+1$

Proof.

[Upper Bound]:

We have that $l_x := \left\lceil \log_2\left(\frac{1}{p(x)}\right)\right\rceil\ \forall x\in\mathcal{X}$ satisfies the Kraft-McMillan inequaltiy and

$$\begin{aligned}
\sum_{x\in\mathcal{X}} p(x)l_x &< \sum_{x\in\mathcal{X}} p(x)\left(\log_2\left(\frac{1}{p(x)}\right)+1\right)\\
&= H(X)+1
\end{aligned}$$

[Lower Bound]:

It's easy to show via Lagrangian multipliers that

$$\begin{aligned}
V &= \min_{l\in\mathbb{R}_+^{|\mathcal{X}|}} \sum_{x\in\mathcal{X}} p(x)l_x \textit{ s.t. } \sum_{x\in\mathcal{X}} 2^{-l_x}\leq 1\\
&= \sum_{x\in\mathcal{X}} p(x)\log_2\left(\frac{1}{p(x)}\right)\\
&= H(X)
\end{aligned}$$

□

We have noted earlier that for a random variable X drawn from probability distribution p, we can write $H(X) = \mathbb{E}_p\left[\log\left(\frac{1}{p(X)}\right)\right]$. Therefore, if $X_1,...,X_n \overset{iid}{\sim} p$, the LLN leads to if $(H(X)<\infty)$,

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{p(X_i)}\right) \overset{a.s.}{\to} \mathbb{E}_p\left[\log\left(\frac{1}{p(X)}\right)\right] \text{ as } n \to \infty$$
$$= H(X)$$

As a result, we have the following theorem.

Theorem 5 (AEP)**.** If $X_1, ..., X_n \overset{iid}{\sim} p$, and we define $T_n^\epsilon := \{(X_1, ..., X_n) : p(X_1, ..., X_n) \in \left[2^{-n(H(X)+\epsilon)}, 2^{-n(H(X)-\epsilon)}\right]\}$, then we have that

- $\forall \epsilon > 0, \Pr\left((X_1, ..., X_n) \in T_n^\epsilon\right) \to 1$ *as* $n \to \infty$
- $\forall \epsilon > 0, (1 - o(1))2^{n(H(X)-\epsilon)} \leq |T_n^\epsilon| \leq 2^{n(H(X)+\epsilon)}$

This theorem basically states that for $X_1, ..., X_n \overset{iid}{\sim} p$, the joint distribution of $X_1, ..., X_n$ is "roughly" a uniform distribution over $2^{nH(p)}$ typical sequences. We say roughly uniformly sampled because the ratio of the probability of one typical sequence to another approaches 1 as $\epsilon \to 0$.[2]

1.2 Introducing Decoder Errors

Here we write about an encoder/ decoder theorem where we will permit some possibility of decode errors. Specifically, suppose that we have $X_1, ..., X_N \overset{iid}{\sim} p$ that's encoded into $Y \in \{0, 1\}^*$ which is then decoded into $(\hat{X}_1, ..., \hat{X}_m)$ with a block error guarantee $ep := \Pr((X_1, ..., X_n) \neq (\hat{X}_1, ..., \hat{X}_m)) \leq \delta$ for some $\delta \in [0, 1]$. We write ep to mean "error probability".

Theorem 6 (Source Coding Theorem with Error Probability)**.** Suppose that we have $X_1, ..., X_n \overset{iid}{\sim} p$. That is encoded into $Y \in \{0, 1\}^*$ which is then decoded into $(\hat{X}_1, ..., \hat{X}_m)$. (Achievability) There exists an (encoder, decoder) such that $\frac{1}{n}\mathbb{E}[l(Y)] \leq H(p) + o(1)$ and $\delta = o(1)$. (Converse) For any (encoder, decoder) pair with $ep \leq \delta = o(1)$, $\frac{1}{n}\mathbb{E}[l(Y)] \geq H(p) - o(1)$.

Proof.

[Achievability]:

Take any $\epsilon > 0$. Consider an (encoder, decoder) pair that enumerates all typical sequences in T_n^ϵ and ignores all others. Then, by AEP, we see that $ep = \Pr((X_1, ..., X_n) \notin T_n^\epsilon) \to 0$. Also, $l(Y) \leq \log_2(|T_n^\epsilon|) \leq n(H(p) + \epsilon)$ deterministically. Since ϵ is arbitrary, we conclude this part of the proof.

[Converse]:

Fix any $\epsilon > 0$. Define two sets

$$A := \{(X_1, ..., X_n) : l(Y) > n(H(p) - 2\epsilon)\}$$
$$B := \{(X_1, ..., X_n) : n = m \text{ and } (X_1, ..., X_n) = (\hat{X}_1, ..., \hat{X}_m)\}$$

[2] When we write that $z = o(1)$, we mean that $z \to 0$ as $n \to \infty$ (or whatever the parameter of interest is goes to infinity).

Then, we have that $\Pr(T_n^\epsilon \cap B) \geq 1 - \delta - o(1)$ by the AEP and the union bound. Moreover

$$\begin{aligned}|T_n^\epsilon \cap B \cap A^c| &= |\{(X_1, ..., X_n) \in T_n^\epsilon \cap B : l(Y(X_1, ..., X_n)) \leq nH(p) - 2\epsilon\}| \\ &\leq |\{y \in \{0,1\}^*, l(y) \leq n(H(p) - 2\epsilon\}| \\ &= \sum_{k=1}^{n(H(p)-2\epsilon)} 2^k < 2 \cdot 2^{n(H(p)-2\epsilon)}\end{aligned}$$

As a result and by the AEP, we then have that $\Pr(T_n^\epsilon \cap B \cap A^c) \leq 2^{-n(H(p)-\epsilon)} |T_n^\epsilon \cap B \cap A^c| < 2 \cdot 2^{-n\epsilon}$. Therefore, we have that $\Pr(T_n^\epsilon \cap A \cap B) \geq 1 - \delta - o(1) - 2 \cdot 2^{-n\epsilon} = 1 - o(1)$. That implies, by Markov's inequality that

$$\begin{aligned}\frac{1}{n}\mathbb{E}[l(Y)] &\geq (H(p) - 2\epsilon) \Pr(A) \\ &\geq (1 - o(1))(H(p) - 2\epsilon)\end{aligned}$$

Since ϵ is arbitrary, the converse follows. □

1.3 Joint Entropy and Mutual Information

Similar to entropy, we can also define the quantities: *joint entropy*, *conditional entropy*, and *mutual information*. We do that here and then look at various properties of these concepts.

Definition 7 (Joint Entropy, Conditional Entropy, and Mutual Information)**.** We define the *joint entropy* between randon variables X and Y as

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

We define the *conditional entropy* of Y given X as

$$\begin{aligned}H(Y|X) &:= \mathbb{E}_{X,Y}\left[\log\left(\frac{1}{p(Y|X)}\right)\right] \\ &= H(X, Y) - H(X)\end{aligned}$$

We define the *mutual information* between X and Y by

$$\begin{aligned}I(X; Y) &:= H(X) + H(Y) - H(X, Y) \\ &= H(Y) - H(Y|X) \\ &= \mathbb{E}_{X,Y}\left[\log\left(\frac{p(X, Y)}{p(X)p(Y)}\right)\right]\end{aligned}$$

Lemma 8 (Non-Negativity of Mutual Information)**.** We have that $I(X; Y) \geq 0$.

Proof.

[Method 1 – Typical Sets]:

Fix some $\epsilon > 0$ and define the following sets:

$$T_n^\epsilon(X) := \{(x^n, y^n) : \left| \left[\frac{1}{n} \sum_{i=1}^{n} \log_2 \left(\frac{1}{p_X(x_i)} \right) \right] - H(X) \right| \leq \epsilon \}$$
$$T_n^\epsilon(Y) := \{(x^n, y^n) : \left| \left[\frac{1}{n} \sum_{i=1}^{n} \log_2 \left(\frac{1}{p_Y(y_i)} \right) \right] - H(Y) \right| \leq \epsilon \}$$
$$T_n^\epsilon(X,Y) := \{(x^n, y^n) : \left| \left[\frac{1}{n} \sum_{i=1}^{n} \log_2 \left(\frac{1}{p_{XY}(x_i, y_i)} \right) \right] - H(X,Y) \right| \leq \epsilon \}$$
$$T_n^\epsilon := T_n^\epsilon(X) \cap T_n^\epsilon(Y) \cap T_n^\epsilon(X,Y), \text{ We call this one the } \textit{joint typical set}$$

For $(X_1, Y_1), ..., (X_n, Y_n) \overset{iid}{\sim} P_{XY}$ the LLN on each of $T_n^\epsilon(X), T_n^\epsilon(Y), T_n^\epsilon(X,Y)$ and a subsequent usage of the union bound implies that $\Pr((X^n, Y^n) \in T_n^\epsilon) \to 1$ as $n \to \infty$. From there, one deduces that $|T_n^\epsilon| \geq (1 - o(1))2^{n(H(X,Y)-\epsilon)}$ using the AEP. Next, draw $(\tilde{X}_1, \tilde{Y}_1), ..., (\tilde{X}_n, \tilde{Y}_n) \overset{iid}{\sim} P_X P_Y$. Then, we have that

$$\begin{aligned}
& 1 \geq \Pr((\tilde{X}^n, \tilde{Y}^n) \in T_n^\epsilon) \\
& \quad = \sum_{(X^n, Y^n) \in T_n^\epsilon} \Pr(\tilde{X}^n = X^n, \tilde{Y}^n = Y^n) \\
& \quad \geq (1 - o(1))2^{n(H(X,Y)-\epsilon)} \cdot 2^{-n(H(X)+\epsilon)} \cdot 2^{-n(H(Y)+\epsilon)} \\
& \quad = (1 - o(1))2^{-n(I(X;Y)+3\epsilon)} \\
\implies & 0 \leq I(X;Y) + 3\epsilon \\
\implies & 0 \leq I(X;Y) \text{ by taking } \epsilon \to 0
\end{aligned}$$

□

[Method 2 – KL Divergence]:

See Application 17.

This is a fundamental inequality to prove other inequalities. For instance, it can be used to show that

- $H(X|Y) \leq H(X)$
- $H(X_1, ..., X_n) \leq \sum_{k=1}^{n} H(X_k)$.
- If $P_{Y^n|X^n} = \Pi_{i=1}^{n} P_{Y_i|X_i}$, then $I(X^n, Y^n) \leq \sum_{i=1}^{n} I(X_i, Y_i)$.
- If $P_{X^n} = \Pi_{i=1}^{n} P_{X_i}$, then $I(X^n, Y^n) \geq \sum_{i=1}^{n} I(X_i, Y_i)$.

We note that all inequalities that can be shown via monotonocity (ie., $H(X|Y) \leq H(X)$) and submodularity (ie., $H(X_A) + H(X_B) \geq H(X_{A \cup B}) + H(X_{A \cap B})$) are called *Shannon-type inequalities*.[3]

1.4 Channel Coding Problem

We consider a problem where an individual wishes to send a message $m \sim Unif(\{1, ..., M\})$ uniformly through a channel. They encode the message into the channel as $x^n \in \mathcal{X}^n$. The channel transforms the message into $y^n \sim P_{Y^n|X^n}(Y^n|X^n = x^n)$. This channel is given by nature and multiple usages of the sample are independent so that $P_{Y^n|X^n}(Y^n|X^n = x^n) = \Pi_{i=1}^{n} P_{Y|X}(Y_i|X_i = x_i)$. Then, a recipient on the other side decodes the message to create the decoded message $\hat{m}$. Given a (block) error probability guarantee (ie., $\Pr(m \neq \hat{m}) \leq \delta$ for some $\delta \in [0, 1]$), we aim to send as many messages as possible or equivalently maximize the rate of communication $R_n := \frac{\log(M)}{n}$, which is the

[3] We define $X_A := \{X_i : i \in A\}$ and all other set analogously.

number of bits used per channel use. The intuition is that if we require more channel usages to communicate a message, then our rate of channel usage is bad.

Definition 9 (Channel Capacity)**.** The channel capacity C is given by $C := C(P_{Y|X}) = \max_{P_X} I(X;Y)$ with $P_{XY} = P_X P_{Y|X}$.

In other words, the channel capacity is reached when given the transition probabilities from X to Y, we design an input distribution P_X so that $I(X;Y)$ is maximized.

Example 10 (Binary Symmetric Channel (BSC))**.** We consider $X, Y \in \{0,1\}$. We define $P_{Y|X}$ using a matrix as

$$\begin{aligned} P_{Y|X} &= \begin{bmatrix} P(Y=0|X=0) & P(Y=1|X=0) \\ P(Y=0|X=1) & P(Y=1|X=1) \end{bmatrix} \\ &= \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix} \end{aligned}$$

In this case, we note that $I(X;Y) = H(Y) - H(Y|X) \leq H(Y) - h_2(\epsilon)$. The maximum mutual information is only reached when $P_X = (1/2, 1/2)$ as then Y will forcibly be unconditionally uniform.[a]

[a] We let $h_2(\epsilon) := \epsilon \log\left(\frac{1}{\epsilon}\right) + (1-\epsilon)\log\left(\frac{1}{1-\epsilon}\right)$ to be the binary entropy function.

Example 11 (Binary Erasure Channel (BEC))**.** We consider $X \in \{0,1\}, Y \in \{0,1,2\}$. We define $P_{Y|X}$ using a matrix as

$$\begin{aligned} P_{Y|X} &= \begin{bmatrix} P(Y=0|X=0) & P(Y=1|X=0) & P(Y=2|X=0) \\ P(Y=0|X=1) & P(Y=1|X=1) & P(Y=2|X=1) \end{bmatrix} \\ &= \begin{bmatrix} 1-\epsilon & 0 & \epsilon \\ 0 & 1-\epsilon & \epsilon \end{bmatrix} \end{aligned}$$

In this case, we have that

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - P(Y \neq 2)\cancelto{0}{H(X|Y \neq 2)} - P(Y=2)\underbrace{H(X|Y=2)}_{=H(X)} \\ &= (1-\epsilon)H(X) \\ &\leq H(X) \end{aligned}$$

The maximum mutual information is only reached when $P_X = (1/2, 1/2)$.

Lemma 12 (Mutual Information Markov Property)**.** If $X - Y - Z$ form a Markov Chain (ie., $P_{XYZ} = P_X P_{Y|X} P_{Z|Y}$), then $I(X;Y) \geq I(X;Z)$.

Proof.

[Method 1 – Non-Negativity of Mutual Information]:

$$\begin{aligned} I(X;Y) - I(X;Z) &= H(X|Z) - H(X|Y) \\ &= H(X|Z) - H(X|Y,Z), \text{ by Markov} \\ &= I(X;Y|Z) \\ &\geq 0, \text{ by non-negativity of mutual information} \end{aligned}$$

□

[Method 2 – DPI]:

See Application 21.

Theorem 13 (Fano's Inequality Channel Coding)**.** If $X \sim Unif([M])$ and Y is distributed according to some distribution with support at most $[M]$, then $\Pr(X \neq Y) \geq 1 - \frac{I(X;Y)+\log(2)}{\log(M)}$.

Proof.

[Method 1 – Leveraging Bounds on Mutual Information and Entropy]:

Let $E := \mathbb{1}_{\{X \neq Y\}}$. Then,

$$\begin{aligned} H(X|Y) &= H(X|Y,E) + \underbrace{I(X,E|Y)}_{\leq H(E) \leq \log(2)} \\ &\leq \Pr(E=1)\underbrace{H(X|Y,E=1)}_{\leq H(X) = \log(M)} + \Pr(E=0)\overset{0}{\cancel{H(X|Y,E=0)}} + \log(2) \\ &\leq \Pr(X \neq Y)\log(M) + \log(2) \end{aligned}$$

On the other hand, we also have

$$\begin{aligned} H(X|Y) &= H(X) - I(X;Y) \\ &= \log(M) - I(X;Y) \end{aligned}$$

Rearranging terms, we get the claim. □

[Method 2 – DPI]:

See Application 22.

Theorem 14 (Shannon's Channel Coding Theorem)**.** Fix any $\epsilon > 0$. (Achievability) If $R_n < C - \epsilon$, then there exists an (encoder, decoder) pair such that $\Pr(m \neq \hat{m}) \to 0$ as $n \to \infty$. (Weak Converse) If $R_n > C + \epsilon$, then for any (encoder, decoder) pair, $\liminf_{n\to\infty} \Pr(m \neq \hat{m}) > 0$. (Strong Converse) If $R_n > C + \epsilon$, then for any (encoder, decoder) pair, $\liminf_{n\to\infty} \Pr(m \neq \hat{m}) = 1$.

Proof.

[Achievability]:

Fix any $\epsilon > 0$. We will show a constructive proof of this result. Construct a random code book $X^n_{(1)}, ..., X^n_{(M)} \overset{iid}{\sim} P_X^{\otimes n}$. As an encoder, for message $m \in [M]$, send $X^n_{(m)}$. As a decoder, find the unique message $\hat{m} \in [M]$ such that

$(X^n_{(\hat{m})}, Y^n)$ is joint typical (see definition in Lemma 8). If None is joint typical or not unique then report a failure.

WLOG, assume that the true message is $m = 1$. Then, we have that $\hat{m} = m$ if and only if $(X^n_{(1)}, Y^n)$ is joint typical *and* none of $(X^n_{(2)}, Y^n)$, ..., $(X^n_{(M)}, Y^n)$ are joint typical (call this event that none of these are joint typical G). By the LLN, we have that $\Pr((X^n_{(1)}, Y^n) \in T^\epsilon_n) = 1 - o(1)$. Reversing the analysis in Lemma 8, since $(X^n_{(2)}, Y^n) \sim P_X^{\otimes n} P_Y^{\otimes n}$ (ie., the components are independent), we have that $\Pr((X^n_{(2)}, Y^n) \in T^\epsilon_n) \leq 2^{-n(I(X;Y)-3\epsilon)}$. As a result, the union bound gives us that

$$\Pr(G) \geq 1 - M \cdot 2^{-n(I(X;Y)-3\epsilon)}$$

For n large enough such that $\log_2(M) < n(I(X;Y) - 4\epsilon)$, then

$$\begin{aligned} \Pr(G) &\geq 1 - 2^{-n\epsilon} \\ &= 1 - o(1) \end{aligned}$$

Therefore, we have that

$$\begin{aligned} \Pr(\hat{m} = 1) &\geq \Pr\left(\left[(X^n_{(1)}, Y^n) \in T^\epsilon_n\right] \cap G\right) \\ &= \Pr((X^n_{(1)}, Y^n) \in T^\epsilon_n) \Pr(G) \\ &= 1 - o(1) \end{aligned}$$

[Weak Converse]:

Suppose that $R_n > C + \epsilon$. Then, we have that

$$\begin{aligned} \Pr(m \neq \hat{m}) &\geq 1 - \frac{I(m, \hat{m}) + \log(2)}{\log(M)}, \text{ by Fano's Inequality} \\ &\geq 1 - \frac{I(X^n, Y^n) + \log(2)}{\log(M)}, \text{ by the Markov Property of } m - X^n - Y^n - \hat{m} \\ &\geq 1 - \frac{\sum_{i=1}^n I(X_i, Y_i) + \log(2)}{\log(M)}, \text{ by one of the consequences of Non-negativity of Mutual Information} \\ &\geq 1 - \frac{nC + \log(2)}{\log(M)}, \text{ by the definition of } C \\ &\geq 1 - \frac{nC + \log(2)}{n(C + \epsilon)}, \text{ using the assumption of the proof} \\ &\to \frac{\epsilon}{C + \epsilon} > 0 \text{ as } n \to \infty \end{aligned}$$

[Strong Converse]:

See later when we have a larger toolkit. □

2 KL Divergence

2.1 Initial Definition and Results

In this section, we will introduce the concept of the KL Divergence between two distributions over the same space. We will also look into applications of the concept.

Definition 15 (Kullback-Leibler (KL) Divergence)**.** For two probability distributions P, Q over the same space, the

Kullback-Leibler Divergence (or the relative entropy) of P with respect to Q is[a]

$$D_{KL}(P||Q) := \begin{cases} \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}(X)\right)\right], & \textit{if } P << Q \\ \infty, & \textit{o.w.} \end{cases}$$

[a] When we say that $P << Q$ we mean that P is absolutely continuous with respect to Q. That means that for any measurable set A, $Q(A) = 0 \implies P(A) = 0$.

To make some remarks, we note that

- If p, q are pmfs, then the KL Divergence becomes $D_{KL}(P||Q) = \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right)$.
- If p, q are pdfs with respect to a measure μ, then $D_{KL}(P||Q) = \int_{\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) d\mu(x)$.
- This is a divergence and *not a distance* (ie., $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.).
- The origin of $D_{KL}(P||Q)$ is that it measures the redundancy of using Q for source coding while the true distribution is P

$$D_{KL}(P||Q) - \underbrace{\sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{q(x)}\right)}_{\text{expected code length using } Q} - \underbrace{H(p)}_{\text{optimal expected code length using } P}$$

Next, we show several properties of the KL Divergence.

Property 16 (Non-Negativity of KL Divergence)**.** $D_{KL}(P||Q) \geq 0$ with equality if and only if $P = Q$.

Proof.

$$\begin{aligned} D_{KL}(P||Q) &= \mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right] \\ &= \mathbb{E}_P\left[-\log\left(\frac{dQ}{dP}\right)\right] \\ &\geq -\log\left(\mathbb{E}_P\left[\frac{dQ}{dP}\right]\right), \text{ by Jensen's Inequality since } -\log(\cdot) \text{ is convex} \\ &= 0 \end{aligned}$$

The portion about equality is also guaranteed by Jensen's Inequality along. □

Application 17 (Non-Negativity of Mutual Information KL Divergence Proof)**.** Here, we show an alternate and simpler proof of the result of Lemma 8. To repeat the statement, we have that $I(X;Y) \geq 0$.

Proof.

$$\begin{aligned} I(X;Y) &= \mathbb{E}_{P_{XY}}\left[\log\left(\frac{P_{XY}(X,Y)}{P_X(X)P_Y(Y)}\right)\right] \\ &= D_{KL}(P_{XY}||P_XP_Y) \\ &\geq 0 \end{aligned}$$

Also, equality holds if and only if $P_{XY} = P_XP_Y$ or in other words if X and Y are independent. □

Property 18 (Joint Convexity of KL Divergence)**.** $(P, Q) \mapsto D_{KL}(P||Q)$ is jointly convex.

Proof.

First, observe that the map $(x, y) \mapsto x \log\left(\frac{x}{y}\right)$ over $\mathbb{R}^2_+$ is jointly convex. To see that, we can construct the Hessian: $H := \begin{bmatrix} \frac{1}{x} & -\frac{1}{y} \\ -\frac{1}{y} & \frac{x}{y^2} \end{bmatrix}$ which has top left entry positive and $\det(H) > 0$ as $(x, y) \in \mathbb{R}^2_+$. With this in hand, the result then follows from the fact that the sum and integral of convex functions is convex. □

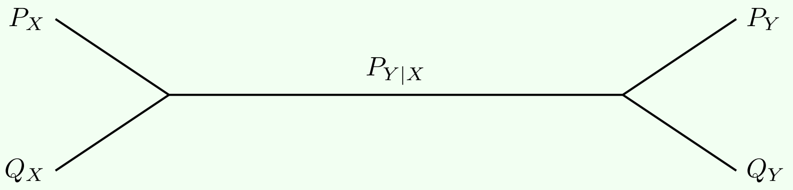
Property 19 (Chain Rule of KL Divergence)**.** $D_{KL}(P_{X^n}||Q_{X^n}) = \sum_{i=1}^n \mathbb{E}_{P_{X^{i-1}}}\left[D_{KL}\left(P_{X_i|X^{i-1}}||Q_{X_i|X^{i-1}}\right)\right]$.

Proof.

$$\begin{aligned}
D_{KL}(P_{X^n}||Q_{X^n}) &= \mathbb{E}_{P_{X^n}}\left[\log\left(\frac{P_{X^n}}{Q_{X^n}}\right)\right] \\
&= \mathbb{E}_{P_{X^n}}\left[\sum_{i=1}^n \log\left(\frac{P_{X_i|X^{i-1}}}{Q_{X_i|X^{i-1}}}\right)\right], \text{ by factoring joint probability where } P^0 := \emptyset =: Q^0 \\
&= \sum_{i=1}^n \mathbb{E}_{P_{X^n}}\left[\log\left(\frac{P_{X_i|X^{i-1}}}{Q_{X_i|X^{i-1}}}\right)\right] \\
&= \sum_{i=1}^n \mathbb{E}_{P_{X^{i-1}}}\left[\mathbb{E}_{X_i,\ldots,X_n|X^{i-1}}\left[\log\left(\frac{P_{X_i|X^{i-1}}}{Q_{X_i|X^{i-1}}}\right)|X^{i-1}\right]\right], \text{ by the tower property} \\
&= \sum_{i=1}^n \mathbb{E}_{P_{X^{i-1}}}\left[\mathbb{E}_{X_i|X^{i-1}}\left[\log\left(\frac{P_{X_i|X^{i-1}}}{Q_{X_i|X^{i-1}}}\right)\right]\right], \text{ inner } \mathbb{E}[\cdot] \text{ doesn't depend on } X_{i+1}, ..., X_n \\
&= \sum_{i=1}^n \mathbb{E}_{P_{X^{i-1}}}\left[D_{KL}\left(P_{X_i|X^{i-1}}||Q_{X_i|X^{i-1}}\right)\right]
\end{aligned}$$

□

Property 20 (Data Processing Inequality (DPI))**.** If the distributions P_X, Q_X are inputted and the distributions P_Y, Q_Y are defined as follows using the channel depicted below: $P_Y(y) = \sum_{x\in\mathcal{X}} P_X(x)P_{Y|X}(Y|X = x)$; $Q_Y(y) = \sum_{x\in\mathcal{X}} Q_X(x)P_{Y|X}(Y|X = x)$.



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

Proof.

[Method 1 – Convexity]:

First note that

$$\begin{aligned}
\mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X}\right] &= \int_{\mathcal{X}} \frac{p(x)}{q(x)} q_{x|Y} d\mu(x) \\
&= \int_{\mathcal{X}} \frac{p(x)}{q(x)} \frac{q(Y|x)q(x)}{q(Y)} d\mu(x) \\
&= \int_{\mathcal{X}} \frac{p(x)p(Y|x)}{Q(Y)} d\mu(x), \text{ since } q(Y|x) = p(Y|x) \text{ by the channel} \\
&= \int_{\mathcal{X}} \frac{p(x,Y)}{Q(Y)} d\mu(x) \\
&= \frac{P_Y}{Q_Y}
\end{aligned}$$

The same holds for discrete distributions. Then, as a result we have that

$$\begin{aligned}
D_{KL}(P_Y||Q_Y) &= \mathbb{E}_{Q_Y}\left[\frac{P_Y}{Q_Y}\log\left(\frac{P_Y}{Q_Y}\right)\right] \\
&= \mathbb{E}_{Q_Y}\left[\mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X}\right]\log\left(\mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X}\right]\right)\right], \text{ using the above} \\
&\leq \mathbb{E}_{Q_Y}\left[\mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X}\log\left(\frac{P_X}{Q_X}\right)\right]\right], \text{ by Jensen's Inequality applied to } x \mapsto x\log(x) \\
&= \mathbb{E}_{Q_X}\left[\frac{P_X}{Q_X}\log\left(\frac{P_X}{Q_X}\right)\right] \\
&= D_{KL}(P_X||Q_X)
\end{aligned}$$

□

[Method 2 – Chain Rule]:

Let $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X P_{Y|X}$.

$$\begin{aligned}
D_{KL}(P_X||Q_X) &= D_{KL}(P_X||Q_X) + \mathbb{E}_{P_X}\left[\underbrace{D_{KL}(P_{Y|X}||Q_{Y|X})}_{=0}\right] \\
&= D_{KL}(P_{XY}||Q_{XY}), \text{ by the chain rule} \\
&= D_{KL}(P_Y||Q_Y) + \mathbb{E}_{P_Y}\left[\underbrace{D_{KL}\left(P_{X|Y}||Q_{X|Y}\right)}_{\geq 0}\right], \text{ by the chain rule} \\
&\geq D_{KL}(P_Y||Q_Y)
\end{aligned}$$

□

2.2 Applications of the Data Processing Inequality

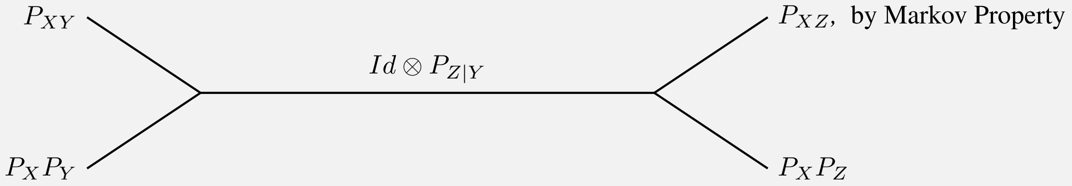
Here we discuss some applications of the data processing inequality to prove some results including some results we've already shown.

Application 21 (Mutual Information Markov Property DPI Proof)**.** Here, we show an alternate proof of the result of Lemma 12 that uses the DPI. To repeat the statement, if $X - Y - Z$ form a Markov Chain (ie., $P_{XYZ} = P_X P_{Y|X} P_{Z|Y}$), then $I(X;Y) \geq I(X;Z)$.

Proof.

Create the following channel to which we wish to apply the DPI. The channel shows that

$$\begin{aligned}
(Id \otimes P_{Z|Y})(P_{XY})(x,z) &:= \sum_{y\in\mathcal{Y}} P_{XY}(x,y)P_{Z|Y}(z|y) \\
&= \sum_{y\in\mathcal{Y}} P_{XY}(x,y)P_{Z|XY}(z|x,y), \text{ by Markov} \\
&= \sum_{y\in\mathcal{Y}} P_{XYZ}(x,y,z) \\
&= P_{XZ}(x,z) \\
(Id \otimes P_{Z|Y})(P_X P_Y)(x,z) &:= \sum_{y\in\mathcal{Y}} P_X(x)P_Y(y)P_{Z|Y}(z|y) \\
&= \sum_{y\in\mathcal{Y}} P_X(x)P_Y(y)P_{Z|Y}(z|y) \\
&= P_X(x)\sum_{y\in\mathcal{Y}} P_{YZ}(y,z) \\
&= P_X(x)P_Z(z)
\end{aligned}$$



We have that

$$\begin{aligned}
I(X;Y) &= D_{KL}(P_{XY}||P_X P_Y) \\
&\geq D_{KL}(P_{XZ}||P_X P_Z), \text{ by DPI} \\
&= I(X;Z)
\end{aligned}$$

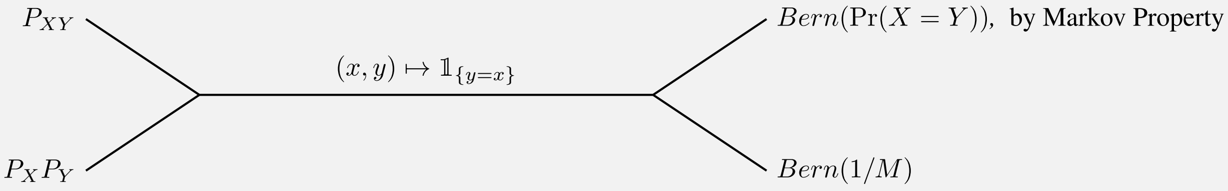
□

Application 22 (Fano's Inequality DPI Proof)**.** Here, we show an alternate proof of the result of Theorem 13 that uses the DPI. To repeat the statement, if $X \sim Unif([M])$ and Y is distributed according to some distribution with support at most $[M]$, then $\Pr(X \neq Y) \geq 1 - \frac{I(X;Y)+\log(2)}{\log(M)}$.

Proof.

Create the following channel to which we wish to apply the DPI. The channel shows that

$$\begin{aligned}
((x',y') \mapsto \mathbb{1}_{\{x'=y'\}}))(P_{XY})(1) &:= \sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\{x'=y'\}} P_{XY}(x',y') \\
&= \Pr(X=Y) \\
((x',y') \mapsto \mathbb{1}_{\{x'=y'\}}))(P_X P_Y)(1) &:= \sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\{x'=y'\}} P_X(x')P_Y(y') \\
&= \sum_{(x',y')\in\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{\{x'=y'\}} \frac{1}{M} P_Y(y') \\
&= \frac{1}{M} \sum_{y'\in\mathcal{Y}} P_Y(y') \\
&= \frac{1}{M}
\end{aligned}$$


P_{XY}
$P_X P_Y$
$(x,y) \mapsto \mathbb{1}_{\{y=x\}}$
$Bern(\Pr(X=Y))$, by Markov Property
$Bern(1/M)$

From here we see that,[a]

$$\begin{aligned}
I(X;Y) &= D_{KL}(P_{XY}||P_X P_Y) \\
&\geq D_{KL}(Bern(\Pr(X=Y))||Bern(1/M)), \text{ by DPI} \\
&= (1-\Pr(X \neq Y))\log\left(\frac{1-\Pr(X\neq Y)}{1/M}\right) + \Pr(X\neq Y)\log\left(\frac{\Pr(X\neq Y)}{1-1/M}\right) \\
&= (1-\Pr(X\neq Y))\log(M) - \Pr(X\neq Y)\underbrace{\log(1-1/M)}_{<0} - \underbrace{H_{bin}(\Pr(X\neq Y))}_{\leq \log(2)} \\
&\geq (1-\Pr(X\neq Y))\log(M) - \log(2)
\end{aligned}$$

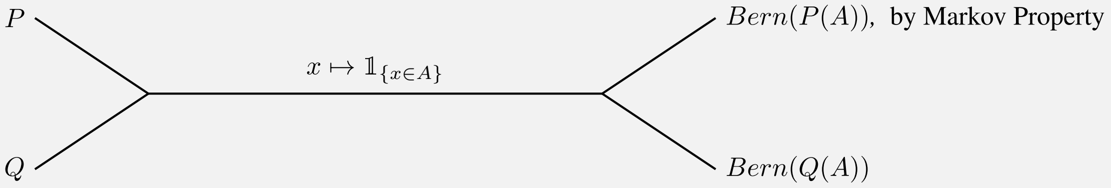
□

[a] In this derivation we use the term $H_{bin}(p) := -p\log(p) - (1-p)\log(1-p)$.

Application 23 (Contiguity). For every event A, $P(A)\log\left(\frac{P(A)}{eQ(A)}\right) \leq D_{KL}(P||Q)$. As a result, if $D_{KL}(P_n||Q_n) = O(1)$, then $Q(A_n) \to 0 \implies P(A_n) \to 0$ by a limit argument.[a]

Proof.

We create the following channel.


P
Q
$x \mapsto \mathbb{1}_{\{x\in A\}}$
$Bern(P(A))$, by Markov Property
$Bern(Q(A))$

Next note that in the proof of Property 18, we have shown that $(x, y) \mapsto x \log\left(\frac{x}{y}\right)$ is jointly convex on the domain $\mathbb{R}_+^2$ which means that it's also jointly convex on $[0,1]^2$. As a result, we find the mininum value the function attains via the first order condition. The first order condition with respect to x implies that at the optimum $y^* = x^* e$. Substituting that back into function, we get that the minimum value equals $\min_{x \in [0,1/e]} -x = -1/e > -1$. That means that

$$(1 - P(A)) \log\left(\frac{1 - P(A)}{1 - Q(A)}\right) \geq -1$$

Then,

$$\begin{aligned}
D_{KL}(P||Q) &\geq D_{KL}(Bern(P(A))||Bern(Q(A))) \\
&= P(A) \log\left(\frac{P(A)}{Q(A)}\right) + (1 - P(A)) \log\left(\frac{1 - P(A)}{1 - Q(A)}\right) \\
&\geq P(A) \log\left(\frac{P(A)}{Q(A)}\right) - 1, \text{ by above} \\
&= P(A) \log\left(\frac{P(A)}{eQ(A)}\right)
\end{aligned}$$

□

[a] By $D_{KL}(P_n||Q_n) = O(1)$, we mean that $\exists C$ such that $D_{KL}(P_n||Q_n) \leq C \ \forall n \in \mathbb{N}$.

2.3 Dual Representations of KL

Now we move to a dual representations of KL Divergence that we construct using a supremum over functions and over probability distributions.

Theorem 24 (Donsker-Varadhan)**.** For any distributions P, Q on the same space, we have that $D_{KL}(P||Q) = \sup_{\{f : \mathbb{E}_Q[\exp(f)] < \infty\}} \mathbb{E}_P[f] - \log(\mathbb{E}_Q[\exp(f)])$.

Proof.

[≤]:

Take $f(X) := \log\left(\frac{dP}{dQ}(X)\right)$, the result follows immediately.

[≥]:

Take any f such that $\mathbb{E}_Q[\exp(f(X))] < \infty$. Define the distribution $\tilde{Q}(dx) := \frac{\exp(f(x))Q(dx)}{\mathbb{E}_Q[\exp(f(X))]}$, which is well-defined since $\mathbb{E}_Q[\exp(f(X))] < \infty$. We then have that

$$\begin{aligned}
D_{KL}(P||Q) - \mathbb{E}_P[f] &= \mathbb{E}_P\left[\log\left(\frac{dP}{\exp(f) dQ}\right)\right] \\
&= \mathbb{E}_P\left[\log\left(\frac{dP}{\mathbb{E}_Q[\exp(f)] d\tilde{Q}}\right)\right] \\
&= \mathbb{E}_P\left[\log\left(\frac{dP}{d\tilde{Q}}\right)\right] - \log(\mathbb{E}_Q[\exp(f)]) \\
&= D_{KL}(P||\tilde{Q}) - \log(\mathbb{E}_Q[\exp(f)]) \\
&\geq 0
\end{aligned}$$

By transitivity, $D_{KL}(P||Q) \geq \mathbb{E}_P[f] - \log(\mathbb{E}_Q[\exp(f)]$. □

Theorem 25 (Gibbs Variational Principle)**.** For any distributions P, Q on the same space and function f such that $\mathbb{E}_Q[\exp(f(X))] < \infty$, we have that $\log(\mathbb{E}_Q[\exp(f)]) = \sup_P \mathbb{E}_P[f] - D_{KL}(P||Q)$.

Proof.

[$\leq$]:

Take $P(dx) = \frac{\exp(f(x))Q(dx)}{\mathbb{E}_Q[\exp(f)]}$, the result follows immediately.

[$\geq$]:

Apply Theorem 24 (ie., Donsker-Varadhan). □

2.3.1 Application – Transportation Inequalities

Here, we will show some applications of Theorems 24 and 25 to optimal transport.

Lemma 26 (Hoeffding's Lemma 1)**.** If X is a bounded random variable with $a \leq X \leq b$, then for any $\lambda \in \mathbb{R}$, we have that

$$\log\left(\mathbb{E}\left[\exp(\lambda(X - \mathbb{E}[X]))\right]\right) \leq \frac{\lambda^2(b-a)^2}{8}$$

Application 27 (Pinsker's Inequality Donsker-Varadhan Proof)**.** For distributions P, Q on the same space, we have that $D_{KL}(P||Q) \geq 2TV(P,Q)^2$.

Proof.

Restrict Donsker-Varadhan to functions f such that $f = \lambda g$ with $||g||_\infty \leq 1$ and $\lambda \in \mathbb{R}$. Then, by Donsker-Varadhan, we have that

$$D_{KL}(P||Q) \geq \sup_{\lambda\in\mathbb{R}, ||g||_\infty \leq 1} \lambda\mathbb{E}_P[g] - \log(\mathbb{E}_Q[\exp(\lambda g)])$$

Applying Lemma 26 (ie., Hoeffding's Lemma) to g, which is bounded, we get that $\log(\mathbb{E}_Q[\exp(\lambda g)]) \leq \lambda\mathbb{E}_Q[g] + \frac{\lambda^2}{2}$. As a result,

$$\begin{aligned} D_{KL}(P||Q) &\geq \sup_{\lambda\in\mathbb{R}, ||g||_\infty \leq 1} \lambda(\mathbb{E}_P[g] - \mathbb{E}_Q[g]) - \frac{\lambda^2}{2} \\ &= \frac{1}{2}\left(\underbrace{\sup_{||g||_\infty \leq 1} \mathbb{E}_P[g] - \mathbb{E}_Q[g]}_{=:2TV(P,Q)}\right)^2 \\ &= 2TV(P,Q)^2 \end{aligned}$$

□

Application 28 (Bobkov and Gotze)**.** The following are equivalent for a constant $C > 0$. (1) $\mathbb{E}_Q[\exp(\lambda(f - \mathbb{E}_Q[f]))] \leq \exp\left(\frac{\lambda^2}{2}C\right)$ for all $1 - Lip$ functions f and $\forall\lambda \in \mathbb{R}$. (2) $W_1(P,Q) \leq \sqrt{2CD_{KL}(P||Q)}$ holds for all distributions P.[a]

Proof.

[(1) $\implies$ (2)]:

We have that[b]

$$\begin{aligned} D_{KL}(P||Q) &\geq \sup_{\lambda\in\mathbb{R}, f:1-Lip} \lambda\mathbb{E}_P[f] - \log\left(\mathbb{E}_Q[\exp(\lambda f)]\right), \text{ by Donsker Varadhan} \\ &\geq \sup_{\lambda\in\mathbb{R}, f:1-Lip} \lambda(\mathbb{E}_P[f] - \mathbb{E}_Q[f]) - \frac{\lambda^2 C}{2}, \text{ by assumption (1)} \\ &= \frac{1}{2C}\left(\sup_{f:1-Lip} \mathbb{E}_P[f] - \mathbb{E}_Q[f]\right)^2, \text{ optimizing over } \lambda \\ &= \frac{W_1(P,Q)^2}{2C} \end{aligned}$$

[(2) $\implies$ (1)]:

$$\begin{aligned} \log\left(\mathbb{E}_Q[\exp(\lambda(f - \mathbb{E}_Q[f]))]\right) &= \sup_P \mathbb{E}_P[\lambda(f - \mathbb{E}_Q[f])] - D_{KL}(P||Q), \text{ by Gibbs} \\ &\leq \sup_P \lambda(\mathbb{E}_P[f] - \mathbb{E}_Q[f]) - \frac{(\mathbb{E}_P[f] - \mathbb{E}_Q[f])^2}{2C}, \text{ by assumption (2)} \\ &\leq \frac{\lambda^2}{2}C \end{aligned}$$

The last step follows from maximizing the concave quadratic in terms of $(\mathbb{E}_P[f] - \mathbb{E}_Q[f])$. We reach a maximal value with $(\mathbb{E}_P[f] - \mathbb{E}_Q[f]) = \lambda C$, which might be achieved or not but the quadratic will always be less than or equal to this. By transitivity and exponentiating both sides, we have that $\mathbb{E}_Q[\exp(\lambda(f - \mathbb{E}_Q[f]))] \leq \exp\left(\frac{\lambda^2}{2}C\right)$. □

[a] We write $W_1(P,Q) := \inf_{\pi\in\Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}[d(X,Y)] = \sup_{f:1-Lip} \mathbb{E}_P[f] - \mathbb{E}_Q[f]$ where $d(\cdot,\cdot)$ is a metric on the common space of X and Y that also helps characterize the space of Lipschitz functions. This is the Wassertein-1 distance. Note that the dual representation is also symmetric since if f is $1 - Lip$ then so is $-f$.

[b] We note that by the given condition on f (ie., $\mathbb{E}_Q[\exp(\lambda(f - \mathbb{E}_Q[f]))] \leq \exp\left(\frac{\lambda^2}{2}C\right)$), we know that $\mathbb{E}_Q[\exp(\lambda f)] < \infty$ so that we can apply Donsker Varadhan in the first line with the inequality.

2.3.2 *Application – Variational Inference*

Here, we will show some applications of Theorems 24 and 25 to variational inference.

Consider a family of distributions $P_\theta(x^n, y^n)$ where both $P_\theta(x^n)$ and $P_\theta(y^n|x^n)$ are tractable. The problem we have is to estimate θ given only y^n (x^n is not observable; it's either missing data or latent variables). The challenge is that $P_\theta(y^n) = \int_{\mathcal{X}\setminus} P_\theta(x^n)P_\theta(y^n|x^n)dx^n$ is often not log-concave nor tractable.

Application 29 (Evidence Lower Bound (ELBO))**.** Given the setting above, we have that $\log\left(P_\theta(y^n)\right) = \sup_q \mathbb{E}_{X^n\sim q}\left[\log\left(\frac{P_\theta(X^n,y^n)}{q(X^n)}\right)\right]$

Proof.

$$\begin{aligned}
\log\left(P_\theta(y^n)\right) &= \log\left(\mathbb{E}_{P_\theta(X^n)}\left[\exp(\log(P_\theta(y^n|X^n)))\right]\right) \\
&= \sup_q \mathbb{E}_{q(X^n)}\left[\log(P_\theta(y^n|X^n))\right] - D_{KL}(q||P_\theta), \text{ by Gibbs} \\
&= \sup_q \mathbb{E}_{q(X^n)}\left[\log(P_\theta(y^n|X^n))\right] - \mathbb{E}_{q(X^n)}\left[\log\left(\frac{q(X^n)}{P_\theta(X^n)}\right)\right] \\
&= \sup_q \mathbb{E}_{q(X^n)}\left[\log\left(\frac{P_\theta(X^n,y^n)}{q(X^n)}\right)\right] \\
&=: ELBO
\end{aligned}$$

□

Application 30 (EM Algorithm)**.** We wish to find the MLE

$$\arg\max_\theta \log(P_\theta(y^n)) = \arg\max_\theta \sup_q \mathbb{E}_{X^N\sim q}\left[\log\left(\frac{P_\theta(X^n,y^n)}{q(X^n)}\right)\right], \text{ by the ELBO proof}$$

Successive Coordinate Ascent Maximization:

- E Step: fix $\theta = \theta^{(t)}$. The maximizer of $\sup_q \mathbb{E}_{X^N\sim q}\left[\log\left(\frac{P_{\theta^{(t)}}(X^n,y^n)}{q(X^n)}\right)\right]$ is $q^{(t)}(X^n) = P_{\theta^{(t)}}(X^n|y^n)$ since $\mathbb{E}_{X^N\sim q}\left[\log\left(\frac{P_\theta(X^n,y^n)}{q(X^n)}\right)\right] = \log(P_\theta(y^n)) - D_{KL}(q||P_\theta(\cdot|y^n))$ and $D_{KL}(\cdot||\cdot) \geq 0$ with equality only when the first argument equals the second. The posterior $P_{\theta^{(t)}}(X^n|y^n)$ is factorizable in the missing data iid case.
- M Step: fix $q = q^{(t)}$. The maximizer is $\theta^{(t+1)} = \arg\max_\theta \underbrace{\mathbb{E}_{X^n\sim q^{(t)}}[\log(P_\theta(X^n,y^n))]}_{\text{Tractable; closed form sometimes}}$

For example, consider the exponential families, $P_\theta(X,Y) \propto \exp(<\theta, T(X,Y)> - A(\theta))$ where $T(x_i,y_i)$ are sufficient statistics for estimating θ, θ is the natural parameter, and $A(\theta)$ is the log-partition function. The E-Step corresponds to the computation of $\mu_i := \mathbb{E}_{X_i\sim P_{\theta^{(t)}}(\cdot|y_i)}[T(X_i,y_i)]$ and the M-Step corresponds to the usual MLE computation with first order condition $\nabla A(\theta^{(t+1)}) = \frac{1}{n}\sum_{i=1}^n \mu_i$.

Application 31 (Variational Autoencoders (VAE))**.** Given images $y_1,...,y_n$, the objective is to find a generative model $X_i \sim \mathcal{N}(0,I)$, $Y_i|X_i \sim N(\mu_\theta(X_i), \sigma^2_\theta(X_i)I)$ where μ_θ and σ^2_θ are parametrized by neural networks. As a consequence of the ELBO when using a restricted and tractable set of approximating models, $q_\phi(y_i) := \mathcal{N}(\mu_\phi(y_i), \sigma^2_\phi(y_i)I)$, we have that

$$\max_\theta \log(P_\theta(y^n)) \geq \max_\theta \max_\phi \mathbb{E}_{X^n\sim\otimes_{i=1}^n q_\phi(y_i)}\left[\log\left(\frac{P_\theta(X^N,y^n)}{q_\phi(X^n)}\right)\right]$$

The idea of the VAE is to maximize a proxy objective that results from the ELBO.

1: Replace $\mathbb{E}_{X^n\sim\otimes_{i=1}^n q_\phi(y_i)}\left[\log\left(\frac{P_\theta(X^N,y^n)}{q_\phi(X^n)}\right)\right]$ by the empirical mean of simulated samples: $x_{ij} \sim \mathcal{N}(\mu_\phi(y_i), \sigma^2_\phi(y_i)I)$ for $j \in \{1,...,M\}$.

2: Compute $\nabla_\theta \mathbb{E}_{X^n\sim\otimes_{i=1}^n q_\phi(y_i)}[\log(P_\theta(X^n,y^n))]$ by approximating this integral via Monte Carlo samples in (1).

3: Compute ∇_ϕ by the reparametrization trick below where $f(\cdot) := \log\left(\frac{P_\theta(\cdot,y^n)}{q_\phi(\cdot)}\right)$.

Reparametrization trick:

$$\begin{aligned}\nabla_\phi \mathbb{E}_{X^n \sim \otimes_{i=1}^n q_\phi(y_i)}[f(X^n)] &= \nabla_\phi \mathbb{E}_{\epsilon^n \sim \otimes_{i=1}^n \mathcal{N}(0,I)}[f(\otimes_{i=1}^n [\mu_\phi(y_i) + \sigma_\phi(y_i) \odot \epsilon_i])] \\ &\approx \frac{1}{M}\sum_{j=1}^M \nabla_\phi f(\otimes_{i=1}^n [\mu_\phi(y_i) + \sigma_\phi(y_i) \odot \epsilon_{ij}])\end{aligned}$$

The reparametrization trick is necessary so that we can change the order of the expectation and the integral. Otherwise, we will be ignoring how ϕ changes the distribution of X_n in our evaluation.

2.3.3 Application – Adaptive Data Analysis

Here, we consider the following adaptive data analysis problem. Suppose that we have data $\mathcal{D} := \{X_i : i \in [n]\}$ with $X_i \overset{iid}{\sim} P$, and a class of functions $\{\phi_t : \mathcal{X} \to \mathbb{R}\}$. For each given ϕ_t, we have

$$\begin{aligned}P_n\phi_t &:= \frac{1}{n}\sum_{i=1}^n \phi_t(x_i) \\ &\approx \mathbb{E}_P[\phi_t(X_i)] =: P\phi_t\end{aligned}$$

The question we wish to study is what happens to $P_n\phi_T$ if the index T depends on the data X^n, and therefore should be treated as a random variable. The idea here is that we can view ϕ_t is a sequence of hypothesis tests and the hypothesis test that we as an econometrician have access to is one that depends on the data (ie., the experimenter might be p-hacking).

Definition 32 (Sub-Gaussian)**.** A function $f : \mathcal{X} \to \mathbb{R}$ of a random variable $X \in \mathcal{X}$ with distribution P is σ^2-sub-Gaussian if for any $\lambda \in \mathbb{R}$, we have that

$$\mathbb{E}_P[\exp(\lambda(f(X) - \mathbb{E}_P[f(X)]))] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right)$$

One notes that this definition implies that $\mathrm{Var}_P(f(X)) \leq \sigma^2$. To see that one can differentiate both sides of the inequality with respect to λ twice and then observe the resulting inequality.[a]

[a] Also note that this definition forces $\mathbb{E}_P[f(X)] \in \mathbb{R}$ since a finite second moment implies a finite first moment by the triangle inequality and Cauchy-Schwarz: $|\mathbb{E}_P[f(X)]| \leq \mathbb{E}_P[|f(X)|] \leq \sqrt{\mathbb{E}_P[(f(X))^2]\mathbb{E}[1^2]} \leq \sqrt{\mathbb{E}_P[(f(X))^2]} < \infty$.

Application 33 (Russo and Zhou (2016))**.** If each ϕ_t is σ^2-sub-Gaussian under P, then

$$|\mathbb{E}_{P_{X^n,T}}[P_n\phi_T] - \mathbb{E}_{Q_{X,T}}[P\phi_T]| \leq \sqrt{\frac{2\sigma^2}{n}I(X^n;T)}$$

where we have defined two distributions: $P_{X^n,T}$ – the joint distribution in the problem; $Q_{X^n,T}$ – an auxiliary distribution where $X^n \perp\!\!\!\perp T$. Observe that if $I(X^n;T) = 0$, then $X^n \perp\!\!\!\perp T$ and $P_n\phi_T$ is unbiased for $P\phi_T$.

Proof.

We have that

$$\begin{aligned}
I(X^n;T) &= \mathbb{E}_{P_T}\left[D_{KL}(P_{X^n|T}||Q_{X^n})\right] \\
&\geq \mathbb{E}_{P_T}\left[\sup_{\lambda\in\mathbb{R}} \mathbb{E}_{P_{X^n|T}}\left[\frac{\lambda}{n}\sum_{i=1}^n \phi_T(X_i)|T\right] - \log\left(\mathbb{E}_{Q_{X^n}}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n \phi_T(X_i)\right)|T\right]\right)\right] \text{, by D-V} \\
&\geq \mathbb{E}_{P_T}\left[\sup_{\lambda\in\mathbb{R}} \lambda\left(\mathbb{E}_{P_{X^n|T}}\left[P_n\phi_T|T\right] - \mathbb{E}_{Q_X}[P\phi_T|T]\right) - \frac{\lambda^2\sigma^2}{2n}\right] \text{, by } \sigma^2\text{-sub-Gaussian} \\
&= \mathbb{E}_{P_T}\left[\left(\mathbb{E}_{P_{X^n|T}}\left[P_n\phi_T|T\right] - \mathbb{E}_{Q_X}[P\phi_T|T]\right)^2\frac{n}{2\sigma^2}\right] \text{, optimizing quadratic wrt } \lambda \\
&= \left(\mathbb{E}_{P_{X^n,T}}\left[P_n\phi_T\right] - \mathbb{E}_{Q_{X,T}}[P\phi_T]\right)^2\frac{n}{2\sigma^2}
\end{aligned}$$

By transitivity we get the result. I wish to validate the two weak inequality steps above in order. Define $f(X^n) = \frac{\lambda}{n}\sum_{i=1}^n \phi_T(X_i)$ for any $\lambda \in \mathbb{R}$. To apply the Donseker-Varadhan theorem, I must show that $\mathbb{E}_{Q_{X^n}}[\exp(f(X^n))|T] < \infty$.

$$\begin{aligned}
\mathbb{E}_{Q_{X^n}}[\exp(f(X^n))|T] &= \mathbb{E}_{Q_{X^n}}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^n \phi_T(X_i)\right)|T\right] \\
&= \Pi_{i=1}^n \mathbb{E}_{Q_X}\left[\exp\left(\frac{\lambda}{n}\phi_T(X_i)\right)|T\right] \text{, by } (X_i \perp\!\!\!\perp T), (X_i \perp\!\!\!\perp X_j)\ \forall i,j \\
&\leq \Pi_{i=1}^n \exp\left(\frac{\sigma^2\lambda^2}{2n^2}\right)\exp\left(\frac{\lambda}{n}\mathbb{E}_{Q_X}[\phi_T(X_i)|T]\right) \text{, by } \sigma^2\text{- sub-Gaussian} \\
&= \exp\left(\frac{\sigma^2\lambda^2}{2n}\right)\exp\left(\lambda\mathbb{E}_{Q_X}[\phi_T(X_i)|T]\right) \\
&< \infty
\end{aligned}$$

since the first term is clearly finite since $\lambda \in \mathbb{R}$ and the second term by $\lambda \in \mathbb{R}$ and the σ^2-sub-Gaussian property of ϕ_T. Specifically, as we've argued in Definition 32, a function being σ^2-sub-Gaussian means it has a finite mean. To establish the second inequality in the main proof:

$$\begin{aligned}
\log\left(\mathbb{E}_{Q_{X^n}}[\exp(f(X^n))|T]\right) &\leq \log\left(\exp\left(\frac{\sigma^2\lambda^2}{2n}\right)\exp\left(\lambda\mathbb{E}_{Q_X}[\phi_T(X_i)|T]\right)\right) \\
&= \frac{\sigma^2\lambda^2}{2n} + \lambda\mathbb{E}_{Q_X}[\phi_T(X_i)|T]
\end{aligned}$$

□

3 f-DIVERGENCE

3.1 Initial Definitions and Results

Here, we will discuss the definition of an f-divergence and see how it generalizes the concept of the KL divergence. From there, we will look at several examples of f-divergences and some applications.

Definition 34 (f-Divergence (Csiszar 1963))**.** Let $f : [0,\infty) \to \mathbb{R}$ be convex with $f(1) = 0$. The f-divergence between two distributions P and Q on the same space $\mathcal{X}$ is given by

$$D_f(P||Q) := \begin{cases} \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right], & \textit{if } P << Q \\ \int_{P_{ac}Q(\mathcal{X})} f\left(\frac{dP}{dQ}dQ\right) + f'(\infty)P_\perp Q(\mathcal{X}), & \textit{o.w.} \end{cases}$$

where $P_{ac}Q(\mathcal{X})$ is the part of the sample space of $\mathcal{X}$ where P is absolutely continuous with respect to Q and $P_\perp Q(\mathcal{X})$ is the part of the sample space where P is singular with respect to Q (eg., $Q(A) = 0$ and $P(A) \neq 0$.). We also define $f'(\infty) := \lim_{x\to\infty} \frac{f(x)}{x}$ and $f(0) := \lim_{x\to 0^+} f(x)$.

To make some remarks, we note that

- Some definitions additionally assume that $f'(1) = 0$. This is WLOG since $f(x)$ and $f(x)+c(x-1)$ have the same f-divergence for $c \in \mathbb{R}$.
- If $\mathcal{X}$ is some continuous sample space and p and q are pdfs with respect to a measure μ, we write (for the case $P << Q$), $D_f(P||Q) := \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu(x)$.
- If $\mathcal{X}$ is some discrete sample space and p and q are pmfs, we write (for the case $P << Q$), $D_f(P||Q) = \sum_{x\in\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right)$

To give examples of some well-known f-divergences, written for the case $P << Q$:

- [Total Variation Distance] If $f(x) := \frac{1}{2}|x-1|$, then $D_f(P||Q) = \frac{1}{2}\int_{\mathcal{X}} |dP - dQ| =: TV(P,Q)$
- [Squared Hellinger Distance] If $f(x) := (\sqrt{x}-1)^2$, then $D_f(P||Q) = \int_{\mathcal{X}} (\sqrt{dP} - \sqrt{dQ})^2 =: H^2(P,Q)$.
- [KL Divergence] If $f(x) := x\log(x)$, then $D_f(P||Q) = \int_{\mathcal{X}} dP \log\left(\frac{dP}{dQ}\right)$.
- [χ^2 Divergence] If $f(x) := (x-1)^2$, then $D_f(P||Q) = \int_{\mathcal{X}} \frac{(dP-dQ)^2}{dQ} =: \chi^2(P||Q)$.
- [Le Cam Distance] If $f(x) := \frac{1-x}{2(x+1)}$, then $D_f(P||Q) = \frac{1}{2}\int_{\mathcal{X}} \frac{(dP-dQ)^2}{dP+dQ} =: LC(P,Q)$.
- [Jensen-Shannon Divergence] If $f(x) := x\log(x) + (x+1)\log\left(\frac{2}{x+1}\right)$, then $D_f(P||Q) = D_{KL}\left(P||\frac{P+Q}{2}\right) + D_{KL}\left(Q||\frac{P+Q}{2}\right) =: JS(P,Q)$.

Next, we show several properties of f-Divergences.

Property 35 (Non-Negativity of f-Divergence). $D_f(P||Q) \geq 0$.

Proof.

We have that

$$\begin{aligned} D_f(P||Q) &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \\ &\geq f\left(\mathbb{E}_Q\left[\frac{dP}{dQ}\right]\right), \text{ by Jensen's on } f \text{ which is convex} \\ &= f(1) \\ &= 0 \end{aligned}$$

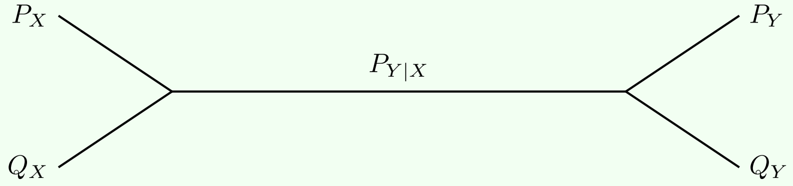
□

Property 36 (Joint Convexity of f-Divergence). $(P,Q) \mapsto D_f(P||Q)$ is jointly convex.

Proof.

First, observe that since f is convex, the map $(x, y) \mapsto yf\left(\frac{x}{y}\right)$ on the domain $(x, y) \in \mathbb{R}^2_+$ is also convex. To see that, we can construct the Hessian: $H := f''(x/y)\begin{bmatrix} \frac{1}{y} & -\frac{x}{y^2} \\ -\frac{x}{y^2} & \frac{x^2}{y^3} \end{bmatrix}$ and observe that it's PSD as the top left entry is positive and $\det(H) > 0$ for $(x, y) \in \mathbb{R}^2_+$. With this in hand, the result then follows from the fact that the sum and integral of convex functions is convex. □

Property 37 (Data Processing Inequality for f-Divergences (DPI))**.** If the distributions P_X, Q_X are inputted and the distributions P_Y, Q_Y are defined as follows using the channel depicted below: $P_Y(y) = \sum_{x\in\mathcal{X}} P_X(x)P_{Y|X}(Y|X = x)$; $Q_Y(y) = \sum_{x\in\mathcal{X}} Q_X(x)P_{Y|X}(Y|X = x)$.



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

Proof.

As we've shown in Property 20, we have that $\mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X}\right]$. We have that

$$\begin{aligned}
D_f(P_Y||Q_Y) &= \mathbb{E}_{Q_Y}\left[f\left(\frac{P_Y}{Q_Y}\right)\right] \\
&= \mathbb{E}_{Q_Y}\left[f\left(\mathbb{E}_{Q_{X|Y}}\left[\frac{P_X}{Q_X}\right]\right)\right], \text{ by above} \\
&\leq \mathbb{E}_{Q_Y}\left[\mathbb{E}_{Q_{X|Y}}\left[f\left(\frac{P_X}{Q_X}\right)\right]\right], \text{ by Jensen's inequality on } f \\
&= \mathbb{E}_{Q_X}\left[f\left(\frac{P_X}{Q_X}\right)\right] \\
&= D_f(P_X||Q_X)
\end{aligned}$$

3.2 f-Divergence and Binary Hypothesis Testing

Why do we care about f-divergence? Consider a simple binary hypothesis testing problem:

- Null $H_0 : X \sim P$ where $\text{supp}(X) = \mathcal{X}$
- Alternative $H_1 : X \sim Q$
- $T : \mathcal{X} \to \{0, 1\}$
- Type I error: $P(T(X) = 1)$
- Type II error: $Q(T(X) = 0)$

Theorem 38 (Characterization of the Binary Hypothesis Test)**.** Given the setting above, we have that

$$\inf_T [P(T(X) = 1) + Q(T(X) = 0)] = 1 - TV(P, Q)$$

Proof.

First, let's show that $TV(P,Q) = \max_A |P(A) - Q(A)|$. First, take any measurable set $A \subseteq \mathcal{X}$.

$$\begin{aligned}
TV(P,Q) &= \frac{1}{2}\int |dP - dQ| \\
&= \frac{1}{2}\int_A |dP - dQ| + \frac{1}{2}\int_{A^c} |dP - dQ| \\
&\geq \frac{1}{2}|P(A) - Q(A)| + \frac{1}{2}|P(A^c) - Q(A^c)| \\
&= \frac{1}{2}|P(A) - Q(A)| + \frac{1}{2}|(1 - P(A)) - (1 - Q(A))| \\
&= |P(A) - Q(A)|
\end{aligned}$$

Thus, $TV(P,Q) \geq \max_A |P(A) - Q(A)|$. Next, define the set $A^* = \{x \in \mathcal{X} : dP(x) > dQ(x)\}$. We have that

$$\begin{aligned}
TV(P,Q) &= \frac{1}{2}\int |dP - dQ| \\
&= \frac{1}{2}\int_{A^*} |dP - dQ| + \frac{1}{2}\int_{(A^*)^c} |dP - dQ| \\
&= \frac{1}{2}\int_{A^*} dP - dQ + \frac{1}{2}\int_{(A^*)^c} dQ - dP \\
&= \frac{1}{2}[P(A^*) - P((A^*)^c) - Q(A^*) + Q((A^*)^c) \\
&= P(A^*) - Q(A^*) \\
&= |P(A^*) - Q(A^*)| \\
&\leq \max_A |P(A) - Q(A)|
\end{aligned}$$

Thus, we have that $TV(P,Q) \leq \max_A |P(A) - Q(A)|$ and jointly $TV(P,Q) = \max_A |P(A) - Q(A)|$. Now onto the main proof.

[$\geq$]:

We see that

$$\begin{aligned}
\inf_T [P(T(X) = 1) + Q(T(X) = 0)] &= \inf_T [1 - (P(T(X) = 0) - Q(T(X) = 0))] \\
&= 1 - \sup_T (P(T(X) = 0) - Q(T(X) = 0)) \\
&= 1 - \sup_A (P(X \in A) - Q(X \in A)) \\
&= 1 - TV(P,Q)
\end{aligned}$$

□

As some remarks on the last theorem:

- If $TV(P,Q) = 0$, then $P = Q$ are totally indistinguishable.
- If $TV(P,Q)$, then $P \perp Q$ are perfectly distinguishable.
- If $TV(P,Q) \in (0,1)$, then P, Q are partially distinguishable.

3.3 Tensorization of f-Divergences

Next, I want to briefly discuss some remarks on tensorization of f-divergences. Suppose that we have distributions $P_1, ..., P_n$ and $Q_1, ..., Q_n$. We make the assumption that $P_i \perp\!\!\!\perp P_j$ for $i \neq j$ and $Q_i \perp\!\!\!\perp Q_j$ for $i \neq j$. When the P_is are independently distributed, we additionally write that $P_1, ..., P_n =: P^{\otimes n}$, and similarly for Q.

- H^2: $1 - \frac{1}{2}H^2(\Pi_i^n P_i, \Pi_i Q_i) = \Pi_i^n (1 - \frac{1}{2}H^2(P_i, Q_i))$
- KL: $D_{KL}(\Pi_i^n P_i || \Pi_{i=1}^n Q_i) = \sum_{i=1}^n D_{KL}(P_i || Q_i)$
- $\chi^2(\Pi_{i=1}^n P_i || \Pi_{i=1}^n Q_i) + 1 = \Pi_{i=1}^n (\chi^2(P_i || Q_i) + 1)$

All f-divergences locally look like the χ^2-divergence when $f''(1)$ exists and $P \approx Q$. To see that,

$$\begin{aligned} D_f(P||Q) &= \mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right) \right] \\ &\approx \mathbb{E}_Q \left[\underbrace{f(1)}_{=0} + f'(1) \underbrace{\left(\frac{dP}{dQ} - 1\right)}_{\mathbb{E}_Q[\cdot]=0} + \frac{f''(1)}{2} \left(\frac{dP}{dQ} - 1\right)^2 \right] \\ &= \frac{f''(1)}{2} \chi^2(P||Q) \end{aligned}$$

3.4 Fisher Information in Parametric Models

Suppose that $\{P_\theta : \theta \in \Theta\}$ is a "regular" parametric model with $\theta \in \mathbb{R}^d$ on $\mathcal{X}$ with measure μ. Assume that each of the models admits a density under the measure $D_\theta P =: f_\theta$. Then for $h \in \mathbb{R}^d$ and $t \approx 0$, we have by first order approximation

$$\begin{aligned} & f_{\theta+th} \approx f_\theta + t(D_\theta f_\theta)'h \\ \implies & f_{\theta+th} - f_\theta \approx t(D_\theta f_\theta)'h \end{aligned}$$

$$\begin{aligned} \chi^2(P_{\theta+th}||P_\theta) &= \int_{\mathcal{X}} \frac{(f_{\theta+th} - f_\theta)^2}{f_\theta} \mu(dx) \\ &\approx \int_{\mathcal{X}} \frac{(t(D_\theta f_\theta)'h)^2}{f_\theta} \mu(dx) \\ &= t^2 h' \underbrace{\int_{\mathcal{X}} \frac{(D_\theta f_\theta)(D_\theta f_\theta)'}{f_\theta} \mu(dx)}_{=: I(\theta) \in \mathbb{R}^{d \times d}} h \\ &= t^2 h' I(\theta) h \end{aligned}$$

From here, we can deduce that

$$\begin{aligned} I(\theta) &= \mathbb{E}_{P_\theta} \left[(D_\theta \log(f_\theta(X)))(D_\theta \log(f_\theta(X)))' \right] \\ &= \mathbb{E}_{P_\theta}[-D_{\theta\theta'} \log(f_\theta(X))] \end{aligned}$$

3.5 f-Divergence as "Average Statistical Information"

Consider our binary hypothesis testing setting of Section 3.2. Suppose that now we have a prior $\Pr(H_0) = \pi \in (0,1)$. Then, the Bayes' error is

$$\begin{aligned} B_\pi &:= \inf_T [\pi P(T(X)=1) + (1-\pi)Q(T(X)=0)] \\ &= \int_{\mathcal{X}} (\pi dP \wedge (1-\pi)dQ)\,,\ x \wedge y := \min(x,y) \end{aligned}$$

The *statistical information* is the difference between the "a priori" and the "a posteriori" Bayes' losses:

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

which is an f-divergence with

$$f_\pi(t) := \pi \wedge (1-\pi) - (\pi t) \wedge (1-\pi) \tag{1}$$

By that I mean that

$$I_\pi(P,Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \tag{2}$$

Theorem 39 (Liese and Vajda (2006))**.** For any f-Divergence, there exists a measure Γ_f on $(0,1)$ such that

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

That means that every f-divergence is an "average" statistical information with different weights on π.

Proof.

We first make a remark that's not in scope. For any $f \in C^2$ that is convex, we have that $f''(dx) := f''(x)dx$ is a measure.

Take any $f : [0,\infty) \to \mathbb{R}$ that is convex and $f(1) = 0$ and WLOG $f'(1) = 0$ that characterizes an f-divergence. For $t \in [0,\infty)$, we have that

$$\begin{aligned} f(t) &= \int_1^t (t-x)f''(dx) \\ &= \int_0^1 (x - t\wedge x)f''(dx) + \int_1^\infty (t - t\wedge x)f''(dx) \end{aligned}$$

Next define the function $\tilde{f}(t)$ as follows

$$\tilde{f}(t) = \int_0^1 (x - t\wedge x)f''(dx) + \int_1^\infty (1 - t\wedge x)f''(dx)$$

Then, we have that

$$\begin{aligned}\mathbb{E}_Q\left[(f-\tilde{f})\left(\frac{dP}{dQ}\right)\right] &= \mathbb{E}_Q\left[\int_1^\infty \left(\frac{dP}{dQ}-1\right)f''(dx)\right]\\ &= \int_1^\infty \overset{0}{\cancel{\mathbb{E}_Q\left[\frac{dP}{dQ}-1\right]}} f''(dx), \text{ under Fubini and its necessary conditions}\\ &= 0\end{aligned}$$

Next, we see that for $x \geq 0$

$$\begin{aligned}1\wedge x - t\wedge x &= (1+x)\left(\frac{1}{1+x}\wedge\frac{x}{1+x}-\frac{t}{1+x}\wedge\frac{x}{1+x}\right)\\ &=: (1+x)f_{\frac{1}{1+x}}(t), \text{ where } f_\pi \text{ is defined in Equation (1)}\end{aligned}$$

As a result,

$$\begin{aligned}\int_0^\infty (1+x)I_{\frac{1}{1+x}}(P,Q)f''(dx) &= \int_0^\infty (1+x)\mathbb{E}_Q\left[f_{\frac{1}{1+x}}\left(\frac{dP}{dQ}\right)\right]f''(dx), \text{ see Equation (2)}\\ &= \mathbb{E}_Q\left[\int_0^\infty (1+x)f_{\frac{1}{1+x}}\left(\frac{dP}{dQ}\right)f''(dx)\right], \text{ by Fubini}\\ &= \mathbb{E}_Q\left[\int_0^1 (1+x)f_{\frac{1}{1+x}}\left(\frac{dP}{dQ}\right)f''(dx)+\int_1^\infty (1+x)f_{\frac{1}{1+x}}\left(\frac{dP}{dQ}\right)f''(dx)\right]\\ &= \mathbb{E}_Q\left[\tilde{f}\left(\frac{dP}{dQ}\right)\right], \text{ by above}\\ &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right], \text{ by above}\\ &= D_f(P||Q)\end{aligned}$$

and $\Gamma_f(\pi)$ is the push-forward measure of $(1+x)f''(dx)$ by the map $x\in[0,\infty)\mapsto\frac{1}{1+x}\in(0,1)$. □

3.6 Guarantees on Contiguity

We next discuss notions of contiguity for f-Divergences.

Definition 40 (Contiguity for f-Divergence)**.** We write that $\{P_n\}$ is contiguous with respect to Q_n (written as $\{P_n\}\triangleleft\{Q_n\}$) if $Q_n(A_n)\to 0 \implies P_n(A_n)\to 0$.

Now some remarks on contiguity:

- If $TV(P,Q)\to 0$, then $\{P_n\}\triangleleft\{Q_n\}$.

- In Application 23, we've shown that $D_{KL}(P_n||Q_n)\leq C\ \forall n\in\mathbb{N}$ for some $C>0$ establish contiguity.

- If $\chi^2(P_n||Q_n) \leq C \; \forall n \in \mathbb{N}$, we have an even stronger guarantee. Make that supposition.[4]

$$\begin{aligned}\frac{(P_n(A_n) - Q_n(A_n))^2}{Q_n(A_n)(1 - Q_n(A_n))} &\leq \chi^2(P_n||Q_n) \; \forall n \in \mathbb{N} \\ &\implies P_n(A_n) \leq Q_n(A_n) + \sqrt{CQ_n(A_n)} \; \forall n \in \mathbb{N}, \text{by transitivity and since } 1 - Q_n(A_n) \leq 1\end{aligned}$$

- As we see here, different f-divergences have different powers in establishing contiguity due to different growth of $f(t)$ as $t \to \infty$.

3.7 Dual Representations of f-Divergences

Similar to KL-Divergence, f-divergences also admit dual representations.

Definition 41 (Convex Conjugate in 1 Dimension)**.** For a function $f : \mathbb{R} \to \mathbb{R}$, its convex conjugate is defined as $f^*(y) = \sup_x(xy - f(x))$.

Here are three useful properties of convex conjugates without proof:

- f^* is convex.

- $f^{**} = f$ under mild regularity conditions.

- Young's inequality: $f(x) + f^*(y) \geq xy$.

Proposition 42 (Convex Conjugate and f-Divergence)**.** $D_f(P||Q) = \sup_{g:\mathbb{E}_Q[f^* \circ g] < \infty} \mathbb{E}_P[g] - \mathbb{E}_Q[f^* \circ g]$

Proof.

$$\begin{aligned}D_f(P||Q) &= \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \\ &= \mathbb{E}_Q\left[\sup_y y\frac{dP}{dQ} - f^*(y)\right], \text{by convex congugate definition} \\ &= \mathbb{E}_Q\left[\sup_{g:\mathcal{X}\to\mathbb{R}} g\frac{dP}{dQ} - f^* \circ g\right] \\ &= \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g] - \mathbb{E}_Q[f^* \circ g]\end{aligned}$$

□

Example 43 (Convex Conjugate Characterization of TV Distance)**.** When $f(x) = \frac{1}{2}|x - 1|$, we have that $f^*(y) = \begin{cases} y, & \text{if } |y| \leq \frac{1}{2} \\ \infty, & \text{if } | > \frac{1}{2} \end{cases}$. So, $TV(P,Q) = \sup_{||g||_\infty \leq \frac{1}{2}} \mathbb{E}_P[g] - \mathbb{E}_Q[g]$.

[4] The first step in the derivation is not at all obvious and it's supported by Cauchy Schwarz. First, we note that $\frac{(P_n(A_n)-Q_n(A_n))^2}{Q_n(A_n)(1-Q_n(A_n))} = \frac{(P_n(A_n)-Q_n(A_n))^2}{Q_n(A_n)} + \frac{(P_n(A_n)-Q_n(A_n))^2}{1-Q_n(A_n)}$. Then, we note that $P_n(A_n) - Q_n(A_n) = \int_{A_n} \frac{(dP-dQ)}{\sqrt{dQ}}\sqrt{dQ}$. That implies by Cauchy-Schwarz that $(\int_{A_n} \frac{(dP-dQ)}{\sqrt{dQ}}\sqrt{dQ})^2 \leq \left(\int_{A_n} \frac{(dP-dQ)^2}{dQ}\right)\left(\int_{A_n} dQ\right) = \left(\int_{A_n} \frac{(dP-dQ)^2}{dQ}\right) Q_n(A_n)$. By transitivity, and dividing over, we get that $\frac{(P_n(A_n)-Q_n(A_n))^2}{Q_n(A_n)} \leq \int_{A_n} \frac{(dP-dQ)^2}{dQ}$. An analogous argument can be used to bound $\frac{(P_n(A_n)-Q_n(A_n))^2}{1-Q_n(A_n)}$ with A_n^c. We then sum both sides of each bound to get the result.

Example 44 (Convex Conjugate Characterization of KL Divergence)**.** When $f(x) = x\log(x)$, $f^*(y) = \exp(y-1)$, so $D_{KL}(P||Q) = \sup_g \mathbb{E}_p[g] - \mathbb{E}_Q[\exp(g-1)]$.

Example 45 (Convex Conjugate Characterization of χ^2 Divergence)**.** When $f(x) = (x-1)^2$, $f^*(y) = y + \frac{y^2}{4}$, so

$$\begin{aligned}\chi^2(P||Q) &= \sup_g \mathbb{E}_P[g] - \mathbb{E}_Q\left[g + \frac{g^2}{4}\right]\\ &= \sup_g \sup_{\lambda, c\in\mathbb{R}} \mathbb{E}_P[\lambda(g+c)] - \mathbb{E}_Q\left[\lambda(g+c) + \frac{\lambda^2(g+c)}{4}\right]\\ &= \sup_g \frac{(\mathbb{E}_P[g] - \mathbb{E}_Q[g])^2}{\mathrm{Var}_Q(g)}, \text{ optimize with respect to } c, \text{ then } \lambda\end{aligned}$$

Corollary 46 (Hammersley- Chapman- Robbins (HRC) Lower Bound)**.** In a parametric family $\{P_\theta : \theta \in \mathbb{R}\}$, if an estimatator $\hat{\theta}$ is unbiased, then by Example 45

$$\mathrm{Var}_{P_\theta}(\hat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'}||P_\theta)}$$

In particular, by taking $\theta' \to \theta$, we recover the Cramer-Rao bound

$$\mathrm{Var}(\hat{\theta}) \geq I(\theta)^{-1}$$

by our work in Section 3.4.

Example 47 (Convex Conjugate Characterization of JS Divergence)**.** When $f(x) = x\log(x) + (x+1)\log\left(\frac{2}{x+1}\right)$, we have that $f^*(y) = \begin{cases} -\log(2-\exp(y)), & \text{if } y < \log(2) \\ \infty, & \text{if } y \geq \log(2)\end{cases}$ Then, we have that

$$\begin{aligned}JS(P,Q) &= \sup_{g \leq \log(2)} \mathbb{E}_P[g] - \mathbb{E}_Q[\log(2-\exp(g))]\\ &= \sup_{0<h<1} \mathbb{E}_P[\log(h)] + \mathbb{E}_Q[\log(1-h)] + \log(2), \text{ with } h := \frac{\exp(g)}{2}\end{aligned}$$

Generative Adversarial Networks aim to minimize

$$\min_{\mathcal{G}} JS(P, P_{\mathcal{G}(Z)}) = \min_{\mathcal{G}} \sup_{0<D<1} \mathbb{E}_{X\sim P}[\log(D(X))] + \mathbb{E}_{Z\sim N}[\log(1 - D(\mathcal{G}(Z)))]$$

where $\mathcal{G}$ is the generator, P is the true data distribution, Z is noise, and D is the discriminator. The generator takes some random noise Z and via generation function $\mathcal{G}$ turns it into a data sample.

3.8 Proving Inequalities between different f-Divergences

Given two f-divergences, how can we prove inequalities between them. For instance, is there a general paradigm to prove Pinsker's inequality (ie., $2TV(P,Q)^2 \leq D_{KL}(P||Q)$).

Definition 48 (Joint Range)**.** Fix two f-divergences $D_f(P||Q)$ and $D_g(P||Q)$. Define the following two quantities:

$$R := \{(D_f(P||Q), D_g(P||Q)) : P, Q \text{ general distributions}\}$$
$$R_k := \{(D_f(P||Q), D_g(P||Q)) : P, Q \text{ general distributions on } [k]\}$$

Theorem 49 (Choquet-Bishop-de Leauw)**.** If C is a metrizable convex compact subset of a locally convex topological vector space, then $C = conv(extremal(C))$.

Theorem 50 (Caratheodory)**.** Let $S \subseteq \mathbb{R}^d$ and $x \in conv(S)$. Then, there exists $S' = \{x_1, ..., x_k\}$ such that $x \in conv(S')$ with (1) $k \leq d+1$ in general and (2) $k \leq d$ if S has at most d connected components.

Theorem 51 (Harrenmoes-Vajda (2011))**.** $R = conv(R_2) = R_4$.

Proof.

We only prove the simpler case of $P << Q$. Obviously $conv(R_2) \subseteq R$ and also obviously $R_4 \subseteq conv(R_2)$ once we've shown that $R \subseteq conv(R_2)$.

[$R \subseteq conv(R_2)$]:

Fix any point $(D_f(P||Q), D_g(P||Q)) \in \mathbb{R}$. Then $L := \frac{dP}{dQ} \in [0, \infty)$ is a random variable with $\mathbb{E}_Q[L] = 1$ and $(D_f(P||Q), D_g(P||Q)) = (\mathbb{E}_Q[f(L)], \mathbb{E}_Q[g(L)])$. Next, consider the set C of all probability measures on $[0, \infty)$ with mean 1.[a] For $\mu \in C$, we associate a point $(\mathbb{E}_\mu[f(L)], \mathbb{E}_\mu[g(L)]) \in \mathbb{R}^2$. Clearly, C is convex and

$$extremal(C) = \{\mu : \mathbb{E}_\mu[L] = 1 \text{ and } \text{supp}(\mu) \leq 2\}$$

The set $extremal(C)$ is the set of all points x that cannot be expressed as $x = \lambda y + (1-\lambda) z$ for $\lambda \in (0, 1)$ and $y, z \in C$. In fact, if A_1, A_2, A_3 form a partition of $[0, \infty)$ and $\mu = \lambda_1\mu_1 + \lambda_2\mu_2 + \lambda_3\mu_3$ for $\lambda_i > 0$ and $\text{supp}(\mu_i) \subseteq A_i$, then the probability and mean constrains of C require that (a) $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and $\lambda_1\mathbb{E}_{\mu_1}[L] + \lambda_2\mathbb{E}_{\mu_2}[L] + \lambda_3\mathbb{E}_{\mu_3}[L] = 1$, which is a line containing $(\lambda_1, \lambda_2, \lambda_3)$ so that μ cannot be an extremal point. Now, by Theorem 49 (ie., Choquet-Bishop-de Leauw[b]), any $\mu \in C$ can be written as a convex combination of extremal points of C. That implies $R \subseteq conv(R_2)$.

[$conv(R_2) \subseteq R_4$]:

By Theorem 50 (ie., Caratheodory), any point of $conv(R_2) \subseteq \mathbb{R}^2$ (which is connected), can be written as a convex combination of 2 points of R_2, which belongs to R_4. □

[a] The idea here is that that under any such measure $\mu \in C$ and for any random variable $L \in [0, \infty)$, if $\mathbb{E}_\mu[L] = 1$, then $\mathbb{E}_\mu[f(L)]$ is a possible f-divergence between two distributions P, Q with $P << Q$.

[b] We gloss over showing that C is compact.

Next, I give some examples of inequalities between various f-divergences:

- [TV vs H^2]: $\frac{H^2}{2} \leq TV \leq \sqrt{H^2\left(1 - \frac{H^2}{4}\right)}$
- [TV vs KL]: $2TV^2 \leq \frac{1}{2}KL$ and $TV \leq 1 - \frac{1}{2}\exp(-KL)$
- $KL \leq \log(1 + \chi^2)$

4 Large Deviation Theory and Hypothesis Testing

In this section, we will discuss the topic of large deviation theory from "typical" outcomes of a probability distribution.

4.1 Large Deviation Theory in Finite Alphabets

Suppose that P is a pmf on $\mathcal{X}$ with $|\mathcal{X}| < \infty$. For $X_1, ..., X_n \overset{iid}{\sim} P$.

Definition 52 (Type)**.** For an empirical distribution Q on $\mathcal{X}$, we define the type of Q to be

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

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

Lemma 53 (Types Encode all Necessary Information for $P(x^n)$)**.** For $x^n \in T_Q^n$, the $P(x^n) = \exp[-n(D_{KL}(Q||P) + H(Q))]$.

Proof.

$$\begin{aligned} P(x^n) &= \Pi_{i=1}^n P(x_i) \\ &= \Pi_{x\in\mathcal{X}} \Pi_{\{i:x_i=x\}} P(x) \\ &= \Pi_{x\in\mathcal{X}} P(x)^{nQ(x)}, \text{ by definition of } T_Q^n \\ &= \exp\left[n \sum_{x\in\mathcal{X}} Q(x)\log(P(x))\right] \\ &= \exp[-n(D_{KL}(Q||P) + H(Q))] \end{aligned}$$

□

Lemma 54 (Polynomial Number of Types)**.** The number of different type classes = $\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1} \le (n+1)^{|\mathcal{X}|-1}$.

Proof.

[=]:

Consider a formulation where we let n_x be the count of the number of observations of $x \in \mathcal{X}$. The number of types is precisely equal to the number of non-negative integer solutions to $\sum_{x\in\mathcal{X}} n_x = n$. By a "stars and bars" argument that is $\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$.

[≤]:

To see the inequality, note that any solution to the integer equation is totally determined by the assignment n_x for $|\mathcal{X}| - 1$ elements. Each of those $|\mathcal{X}| - 1$ elements can take at most one of $\{0, ..., n\}$ values, leading to the result. □

Lemma 55 (Bounding Size of Type Set)**.** We have that $\frac{\exp(nH(Q))}{(n+1)^{|\mathcal{X}|-1}} \le |T_Q^n| \le \exp(nH(Q))$.

Proof.

[Upper bound]:

We have that

$$\begin{aligned}1 &\geq Q(x^n \in T_Q^n) \\ &= |T_Q^n| \exp(-H(Q)), \text{ by Lemma 53 with } P := Q\end{aligned}$$

By transitivity and rearranging gives the desired result.

[Lower bound]:

We have that

$$\begin{aligned}1 &= \sum_P Q(x^n \in T_P^n), \text{ sum over type sets} \\ &\leq \sum_P Q(x^n \in T_Q^n), \text{ largest term in sum is given by multinom}(n; Q) \\ &\leq (n+1)^{|\mathcal{X}|-1} |T_Q^n| \exp(-H(Q)), \text{ using Lemma 53 and Lemma 54}\end{aligned}$$

By transitivity and rearranging gives the desired result. □

Corollary 56 (Bounding Probability of Event of Type)**.** We have that $\frac{\exp(-nD_{KL}(Q||P))}{(1+n)^{|\mathcal{X}|-1}} \leq P(x^n \in T_Q^n) \leq \exp(-nD_{KL}(Q||P))$.

Proof.

By Lemma 53, we have that $P(x^n \in T_Q^n) = |T_Q^n| \exp[-n(D_{KL}(Q||P) + H(Q))]$. That implies that $\frac{P(x^n \in T_Q^n)}{|T_Q^n|} = \exp[-n(D_{KL}(Q||P) + H(Q))]$. Take the result of Lemma 55 (ie., $\frac{\exp(nH(Q))}{(n+1)^{|\mathcal{X}|-1}} \leq |T_Q^n| \leq \exp(nH(Q))$), and multiply the LHS of this identity to the center term and the RHS of the identity to the outer terms of this lemma's result. That yields the desired conclusion. □

Theorem 57 (Sanov's Theorem)**.** Continue to let $|\mathcal{X}| < \infty$ and let $\hat{P}$ be the empirical distribution (type) of $X_1, ..., X_n \overset{iid}{\sim} P$ where P is strictly positive.[a] Let E be a closed set of distributions with a non-empty interior.[b] Then,

$$P(\hat{P} \in E) = \exp\left(-n \min_{Q \in E} D_{KL}(Q||P) + o(n)\right)$$

We remark that the map $P \mapsto \arg\min_{Q \in E} D_{KL}(Q||P)$ is called the *information projection*.

Proof.

[Upper bound]:

We have that

$$\begin{aligned}P(\hat{P} \in E) &= \sum_{Q \in E} P(x^n \in T_Q^n) \\ &\leq \sum_{Q \in E} \exp(-nD_{KL}(Q||P)), \text{ by Corollary 56} \\ &\leq (n+1)^{|\mathcal{X}|-1} \exp\left(-n \left[\min_{Q \in E} D_{KL}(Q||P)\right]\right), \text{ by Lemma 54}\end{aligned}$$

[Lower bound]:

We have that for any $Q \in E$, $P(\hat{P} \in E) \geq P(x^n \in T_Q^n) \geq \frac{\exp(-nD_{KL}(Q||P))}{(n+1)^{|\mathcal{X}|-1}}$. We let $Q^* = \arg\min_{Q \in E} D_{KL}(Q||P)$ (which is attained by the continuity of $Q \mapsto D_{KL}(Q||P)$ and the compactness of E). Taking $Q \to Q^*$ and applying the continuity of $Q \mapsto D_{KL}(Q||P)$ gives the desired result. □

[a] We note that the strict positivity of P gives the continuity of $Q \mapsto D_{KL}(Q||P)$ under say TV distance.

[b] we note that E is a closed subset of a compact set of $\Delta^{|\mathcal{X}|-1}$ so therefore is compact.

Corollary 58 (Sanov's Theorem Corollary). We have that

$$\lim_{n\to\infty} \frac{1}{n} \log\left(\frac{1}{P(\frac{1}{n}\sum_{i=1}^n X_i \geq \gamma)}\right) = \min_{Q:\mathbb{E}_Q[X_i]\geq\gamma} D_{KL}(Q||P)$$

Proof.

This is a straightforward application of Theorem 57 (ie., Sanov's Theorem) with $E = \{Q : \mathbb{E}_Q[X_i] \geq \gamma\}$. □

4.2 *Large Deviation Theory in General Alphabets*

Suppose that P is a pmf on $\mathcal{X}$ with $|\mathcal{X}| \in [1, \infty]$. For $X_1, ..., X_n \overset{iid}{\sim} P$.

Definition 59 (Exponential Tilt). For $\lambda \in \mathbb{R}$, the exponential tile of P along X is

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

where $\psi(\lambda) := \mathbb{E}_P[\exp(\lambda X)]$ is the cumulant generating function (CGF) of X.

We note that the family $\{P_\lambda\}$ is called an "exponential family" in statistics where $\psi(\lambda)$ is called the "log partition function". In particular, we have that:

- $\mathbb{E}_{P_\lambda}[X] = \psi(\lambda)$
- $\lambda \mapsto \psi(\lambda)$ is convex. We have that $\psi''(\lambda) = \mathrm{Var}_{P_\lambda}(X) \geq 0$.

Theorem 60 (Maximum Entropy Distribution). If $\mathbb{E}_P[X] < \gamma$ and there exists $\lambda \in \mathbb{R}$ such that $\mathbb{E}_{P_\lambda}[X] = \gamma$. Then,

$$\begin{aligned}\min_{Q:\mathbb{E}_Q[X]\geq\gamma} D_{KL}(Q||P) &\overset{(1)}{=} D_{KL}(P_\lambda||P)\\ &\overset{(2)}{=} \lambda\gamma - \psi(\lambda)\\ &\overset{(3)}{=} \psi^*(\gamma)\end{aligned}$$

where ψ^* is the convex conjugate of ψ.

Proof.

First, as a general note, we see that $\mathbb{E}_P[X] = \psi'(0) < \gamma = \psi'(\lambda)$. Thus, by the convexity of ψ, we have that $\lambda > 0$.

[(1-2)]:

We see that if $\mathbb{E}_Q[X] \geq \gamma$, then

$$\begin{aligned} D_{KL}(Q||P) &= \mathbb{E}_Q\left[\log\left(\frac{Q}{P}\right)\right] \\ &= \mathbb{E}_Q\left[\log\left(\frac{Q}{P_\lambda}\right) + \log\left(\frac{P_\lambda}{P}\right)\right] \\ &= \underbrace{D_{KL}(Q||P_\lambda)}_{\geq 0} + \mathbb{E}_Q[\lambda X - \psi(\lambda)] \\ &\geq \lambda\gamma - \psi(\lambda), \text{ since } \mathbb{E}_Q[X] \geq \gamma \text{ and } \lambda > 0 \end{aligned}$$

We see that also

$$\begin{aligned} D_{KL}(P_\lambda||P) &= \mathbb{E}_{P_\lambda}[\lambda X - \psi(\lambda)] \\ &= \lambda\gamma - \psi(\lambda) \end{aligned}$$

attains the lower bound.

[3]:

By assumption, $\gamma = \mathbb{E}_{P_\lambda}[X] = \psi'(\lambda)$. Then,

$$\begin{aligned} \psi^*(\gamma) &= \sup_{\lambda^* \in \mathbb{R}} \lambda^*\gamma - \psi(\lambda^*), \text{by Definition 41} \\ &\leq \sup_{\lambda^* \in \mathbb{R}} \lambda^*\gamma - (\psi(\lambda) + (\lambda^* - \lambda)\psi'(\lambda)), \text{ by convexity of } \psi \\ &= \lambda\gamma - \psi(\lambda) \end{aligned}$$

So, $\psi^*(\gamma) = \lambda\gamma - \psi(\lambda)$ by transitivity. □

By Theorem 60, we see that the information projection yields an exponential tilt of P and the value of the divergence is given the by the convex conjugate of the CGF of P.

Lemma 61 (Chernoff's Inequality)**.** For any random variable $X \sim P$ and $a \in \mathbb{R}$, we have that

$$P(X > a) \leq \inf_{t \geq 0} \exp(-ta)\mathbb{E}_P[\exp(tX)]$$

Theorem 62 (Cramer's Theorem)**.** For $X_1, ..., X_n \overset{iid}{\sim} P$ with $\mathbb{E}_P[X] < \gamma < ||X||_\infty$, then

$$\begin{aligned} \psi^*(\gamma) &= \lim_{n\to\infty} \frac{1}{n} \log\left(\frac{1}{P(\frac{1}{n}\sum_{i=1}^n X_i > \gamma)}\right) \\ &= \inf_{Q:\mathbb{E}_Q[X]>\gamma} D_{KL}(Q||P) \end{aligned}$$

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

Proof.

From Theorem 60, we note that $\psi(\gamma) = \inf_{Q:\mathbb{E}_Q[X]>\gamma} D_{KL}(Q||P)$. Thus, we must only show one more identity to complete the proof of this theorem.

[Method 1 – Probabilistic]:

Here, we show that $\psi^*(\gamma) = \lim_{n\to\infty} \frac{1}{n}\log\left(\frac{1}{P(\frac{1}{n}\sum_{i=1}^n X_i > \gamma)}\right)$.

[Upper Bound]:

By Chernoff's inequality (ie., Lemma 61), we have that

$$\begin{aligned} P\left(\frac{1}{n}\sum_{i=1}^n X_i > \gamma\right) &\leq \inf_{\lambda\geq 0} \exp(-\lambda n\gamma)\mathbb{E}_P\left[\exp\left(\lambda\sum_{i=1}^n X_i\right)\right] \\ &= \inf_{\lambda\geq 0} \exp\left(-n(\lambda\gamma - \psi(\lambda)\right) \\ &= \exp\left(-n\left(\sup_{\lambda\geq 0} \lambda\gamma - \psi(\lambda)\right)\right) \\ &= \exp(-n\psi^*(\gamma)), \text{ since } \mathbb{E}_P[X] < \gamma \end{aligned}$$

Rearranging and taking the limit as $n \to \infty$ yields the desired upper bound.

[Lower Bound]:

Since $\mathbb{E}_P[X] < \gamma < ||X||_\infty$, $\exists\lambda = \lambda(\epsilon)$ such that $\mathbb{E}_{P_\lambda}[X] = \gamma + \epsilon$ where P_λ is the exponential tilt of P.[a] By the LLN,

$$P_\lambda\left(\frac{1}{n}\sum_{i=1}^n X_i \in (\gamma, \gamma + 2\epsilon)\right) = 1 - o(1)$$

At the same time, for $\frac{1}{n}\sum_{i=1}^n X_i \in (\gamma, \gamma+2\epsilon)$, we have that

$$\begin{aligned} \frac{dP_\lambda}{dP}(X_1, ..., X_n) &= \exp\left(\lambda\sum_{i=1}^n X_i - n\psi(\lambda)\right) \\ &\leq \exp\left[n(\lambda(\gamma+2\epsilon) - \psi(\lambda))\right] \\ \implies P\left(\frac{1}{n}\sum_{i=1}^n X_i \in (\gamma, \gamma+2\epsilon)\right) &\geq (1 - o(1))\exp\left[-n(\lambda(\gamma+2\epsilon) - \psi(\lambda))\right] \end{aligned}$$

Taking $\epsilon \to 0^+$ and rearranging gives the desired lower bound when we also take $n \to \infty$.

[Method 2 – Information Theoretic]:

Here we show that $\lim_{n\to\infty} \frac{1}{n}\log\left(\frac{1}{P(\frac{1}{n}\sum_{i=1}^n X_i > \gamma)}\right) = \inf_{Q:\mathbb{E}_Q[X]>\gamma} D_{KL}(Q||P)$.

[Lower Bound]:

Fix any Q with $\mathbb{E}_Q[X] > \gamma$. Then, for $E_n := \{\frac{1}{n}\sum_{i=1}^n X_i > \gamma\}$, we have that

$$Q(E_n) = 1 - o(1), \text{ by LLN}$$

By Application 23, we have that

$$\begin{aligned} Q(E_n)\log\left(\frac{Q(E_n)}{eP(E_n)}\right) &\leq D_{KL}(Q_{X^n}||P_{X^n}) \\ &= nD_{KL}(Q||P), \text{ by independence} \\ \implies \frac{1}{n}\log\left(\frac{1}{P(E_n)}\right) &\leq \frac{D_{KL}(Q||P)}{Q(E_n)} - \frac{\log(Q(E_n)/e)}{n} \\ &= (1+o(1))D_{KL}(Q||P) \end{aligned}$$

showing the desired lower bound by transitivity.

[Upper Bound]:

Note that $\tilde{P}_{X^n} := P_{X^n|\frac{1}{n}\sum_{i=1}^n X_i > \gamma}$ has mean greater than γ, with

$$\frac{1}{n}\log\left(\frac{1}{P(E_n)}\right) = \frac{1}{n}D_{KL}(\tilde{P}_{X^n}||P_{X^n})$$

We argue that $\frac{1}{n}D_{KL}(\tilde{P}_{X^n}||P_{X^n}) \geq \inf_{Q:\mathbb{E}_Q[X]>\gamma} D_{KL}(Q||P)$. In fact,

$$\begin{aligned} D_{KL}(\tilde{P}_{X^n}||P_{X^n}) &= \sum_{i=1}^n \mathbb{E}_{\tilde{P}_{X^{i-1}}}[D_{KL}(\tilde{P}_{X_i|X^{i-1}}||P)], \text{ by Property 19} \\ &\geq \sum_{i=1}^n D_{KL}\left(\mathbb{E}_{\tilde{P}_{X^{i-1}}}[\tilde{P}_{X_i|X^{i-1}}]||P\right), \text{ by Property 18 (ie., convexity of KL Divergence)} \\ &\geq nD_{KL}\left(\frac{1}{n}\sum_{i=1}^n \tilde{P}_{X_i}||P\right), \text{ by Property 18 (ie., convexity of KL Divergence)} \end{aligned}$$

We see that $\bar{P} := \frac{1}{n}\sum_{i=1}^n \tilde{P}_{X_i}$ clearly satisfies $\mathbb{E}_{\bar{P}}[X] = \frac{1}{n}\mathbb{E}_{\tilde{P}}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] > \gamma$. By, transitivity, we establish the upper bound. □

[a] A proof of this argues that for suitably large λ, we can arbitrarily bound the contribution of elements to $\mathbb{E}_{P_\lambda}[X]$ by elements that have value $\leq \gamma + \epsilon$. The continuity of ψ' and the intermediate value theorem mean we can attain any intermediate value in $(\mathbb{E}_P[X], ||X||_\infty)$ for $\mathbb{E}_{P_\lambda}[X]$.

4.3 Simple Hypothesis Testing

Here, we revert to the setup in Section 3.2 and discuss Binary Hypothesis Testing in light of these new results.

Consider a simple hypothesis testing problem where:

- Null $H_0 : X \sim P$ where $\text{supp}(X) = \mathcal{X}$

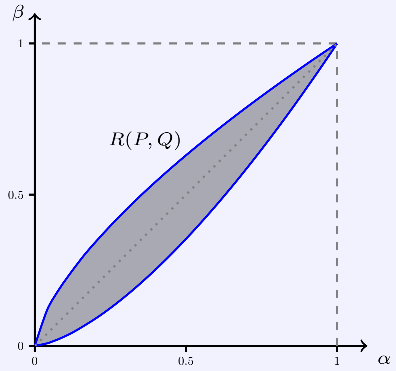
- Alternative H_1: $X \sim Q$

- Test $T : \mathcal{X} \to \{0, 1\}$

- 1 - Type I error: $\alpha(T) := P(T = 0)$

- Type II error: $\beta(T) := Q(T = 0)$

Definition 63 (Achievable Type I / Type II Error Tradeoffs)**.** We let $R(P, Q)$ denote the set of all achievable points $(\alpha, \beta) \in [0, 1]^2$ when T ranges over all possible sets.



One can show the following basic properties of $R(P,Q)$:

- $R(P,Q)$ is convex. To see that, consider a randomized combination of two tests.
- $(a,a) \in R(P,Q) \;\; \forall a \in [0,1]$. To see that, not that one can construct a test $T \sim \text{Bern}(1-a)$, which is independent of X.
- $(\alpha,\beta) \in R(P,Q) \implies (1-\alpha, 1-\beta) \in R(P,Q)$. To see that, one can consider what happens when one replaces T by $1-T$.

Property 64 (Neyman-Pearson)**.** Consider likelihood ratio tests (LRTs) of the form

$$T^*(\tau) = \begin{cases} 0 & \text{if } \log\left(\frac{P(X)}{Q(X)}\right) > \tau \\ 1 & \text{if } \log\left(\frac{P(X)}{Q(X)}\right) < \tau \\ \in \{0,1\} & \text{o.w. (and arbitrarily chosen from set)} \end{cases}$$

For any other test T, we have that $\alpha(T) \geq \alpha(T^*) \implies \beta(T) \geq \beta(T^*)$.

Proof.

If $\alpha(T) \geq \alpha(T^*)$, then we see that $\mathbb{E}_P[T - T^*] \leq 0$. Since $\mathbb{E}_P\left[\left(\frac{dQ}{dP} - \exp(-\tau)\right)(T - T^*)\right] \leq 0$ (by checking on a case by case basis when $\frac{dQ}{dP} < \exp(-\tau)$ and $\frac{dQ}{dP} > \exp(-\tau)$). Next, we see that

$$\begin{aligned} 0 &\geq \mathbb{E}_P\left[\left(\frac{dQ}{dP} - \exp(-\tau)\right)(T - T^*)\right] \\ &= \mathbb{E}_P\left[\frac{dQ}{dP}(T - T^*)\right] - \underbrace{\exp(-\tau)\mathbb{E}_P\left[(T - T^*)\right]}_{\geq 0} \\ \implies 0 &\geq \mathbb{E}_P\left[\frac{dQ}{dP}(T - T^*)\right] \\ &= \mathbb{E}_Q[T - T^*] \end{aligned}$$

With transitivity, that implies that $\beta(T) \geq \beta(T^*)$. □

4.4 Asymptotics: Chernoff Regime

Next, we will discuss asymptotics in the Chernoff regime. Consider a situation where we observe a random variable X^n where $\text{supp}(X_i) = \mathcal{X}$ for each $i \in \{1, ..., n\}$. We consider the following two hypotheses:

- $H_0 : X^n \overset{iid}{\sim} P$ (ie., $X^n \sim P^{\otimes n}$)

- $H_1 : X^n \overset{iid}{\sim} Q$ (ie., $X^n \sim Q^{\otimes n}$)

With $n \to \infty$, what are all possible values of (E_0, E_1) such that there exists T_n with (a) $1 - \alpha(T_n) \leq \exp(-nE_0)$ and (b) $\beta(T_n) \leq \exp(-nE_1)$, asymptotically? In other words, what are asymptotically the best tradeoffs between (E_0, E_1), the error exponents on Type I & II errors?

Theorem 65 ($E_0 - E_1$ Tradeoff)**.** Assume that $P << Q$ and $Q << P$. The upper boundary of all achievable (E_0, E_1) pairs is given by $E_0 = D_{KL}(P_\lambda||P)$ and $E_1 = D_{KL}(P_\lambda||Q)$ parametrized by $\lambda \in [0, 1]$ where $P_\lambda \propto P^{1-\lambda Q^\lambda}$ (ie., an interpolation between P and Q).

Proof.

[Achievability]:

A note on a non-interesting case: when $P = Q$, $D_{KL}(P_\lambda||P) = 0 = D_{KL}(P_\lambda||Q)$ $\forall\lambda \in [0, 1]$. The set of achievable $(D_{KL}(P_\lambda||P), D_{KL}(P_\lambda||Q))$ reduces to $\{(0, 0)\}$. Next, consider any set A_n so that $T_n = \mathbb{1}_{\{A_n\}}$. Then, the probability of a Type I error, is $1 - \alpha(T_n) = P(T_n = 1) = P(A_n)$. The probability of a Type II error is $\beta(T_n) = Q(T_n = 0) = Q(A_n^c) = P(A_n^c)$. That implies $\alpha(T_n) + \beta(T_n) = 1$ $\forall n \in \mathbb{N}$ and any test T_n. That means that we cannot in turn simultaneously send both error probabilities to 0 so $(E_0 = 0, E_1 = 0)$ is the best asymptotic bound we can achieve. We ignore degenerate tests where $T_n = 1$ a.s. when thinking about this asymptotic tradeoff.

Let's now focus on the more interesting case of $P \neq Q$. A sufficient statistic statistic for our hypothesis is $L := \frac{1}{n}\sum_{i=1}^n L_i$ where $L_i := \log\left(\frac{dP(X_i)}{dQ(X_i)}\right)$ so a natural test is $T_n = \mathbb{1}_{\{L \leq \gamma\}}$ for some threshold $\gamma \in \mathbb{R}$. By Theorem 60 (ie., Maximum Entropy Distribution) and Theorem 62 (Cramer's Theorem), we have that

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

where

$$\begin{aligned}
\psi_P(\lambda) &= \log\left(\mathbb{E}_P\left[\lambda L_i\right]\right) \\
&= \log\left(\int_{\mathcal{X}} (dP)^{1+\lambda}(dQ)^{-\lambda}\right) \\
\psi_Q(\lambda) &= \log\left(\mathbb{E}_Q[\lambda L_i]\right) \\
&= \log\left(\int_{\mathcal{X}} (dP)^{\lambda}(dQ)^{1-\lambda}\right)
\end{aligned}$$

and

$$\begin{aligned}
P^*(dx) &= \exp\left(\lambda_P^* \log\left(\frac{dP}{dQ}\right) - \psi_P(\lambda_P^*)\right) P(dx), \text{ with } \mathbb{E}_{P^*}[L_i] = \gamma \\
Q^*(dx) &= \exp\left(\lambda_Q^* \log\left(\frac{dP}{dQ}\right) - \psi_Q(\lambda_Q^*)\right) Q(dx), \text{ with } \mathbb{E}_{Q^*}[L_i] = \gamma
\end{aligned}$$

Since $P^* \propto P^{1+\lambda_P^*}Q^{-\lambda_Q^*}$ and $Q^* \propto P^{\lambda_Q^*}Q^{1-\lambda_Q^*}$ belong to the exponential family $(P_\lambda)_{\lambda\in[0,1]}$, and the function $\lambda \mapsto \mathbb{E}_{P_\lambda}\left[\log\left(\frac{dP}{dQ}\right)\right]$ is continuous and strictly decreasing for $\lambda \in [0, 1]$,[a] we conclude by the intermediate value theorem that $P^* = Q^* =: P_{\lambda^*}$ where λ^* is the solution to $\mathbb{E}_{P_{\lambda^*}}\left[\log\left(\frac{dP}{dQ}\right)\right] = \gamma$. Therefore, by choosing γ

appropriately between $\mathbb{E}_P\left[\log\left(\frac{dP}{dQ}\right)\right]$ and $\mathbb{E}_Q\left[\log\left(\frac{dP}{dQ}\right)\right]$, this test asymptotically achieves all pairs

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

[Strong Converse]:

Suppose that some test T_n asymptotically attains $\alpha(T_n) \geq 1 - \exp(-nE_0)$ and $\beta(T_n) \leq \exp(-nE_1)$. As an initial result, let's show that $\forall \gamma > 0$, $\alpha(T_n) - \gamma\beta(T_n) \leq P\left(\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right) > \log(\gamma)\right)$. To see this, let $L := \sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right)$. Then,

$$\begin{aligned}
\alpha(T_n) - \gamma\beta(T_n) &= P^{\otimes n}(T_n = 0) - \gamma Q^{\otimes n}(T_n = 0) \\
&= \mathbb{E}_{Q^{\otimes n}}\left[\frac{dP^{\otimes n}}{dQ^{\otimes n}} \mathbb{1}_{\{T_n = 0\}}\right] - \gamma \mathbb{E}_{Q^{\otimes n}}\left[\mathbb{1}_{\{T_n=0\}}\right] \\
&= \mathbb{E}_{Q^{\otimes n}}\left[(\exp(L) - \gamma)\left(\mathbb{1}_{\{T_n=0\}}\right)\right] \\
&\leq \mathbb{E}_{Q^{\otimes n}}\left[(\exp(L) - \gamma)\left(\mathbb{1}_{\{T_n=0, L > \log(\gamma)\}}\right)\right] \\
&\leq \mathbb{E}_{Q^{\otimes n}}\left[\exp(L)\mathbb{1}_{\{L > \log(\gamma)\}}\right] \\
&= P^{\otimes n}(L > \log(\gamma))
\end{aligned}$$

By transitivity, we have the desired result. Next, pick $\gamma = \exp(n\theta)$ for some (TBD) $\theta \in \mathbb{R}$. Then, we have that

$$\begin{aligned}
1 - \exp(-nE_0) - \exp(-n(E_1 - \theta)) &\leq \alpha(T_n) - \gamma\beta(T_n), \text{ by assumption} \\
&\leq P\left(\frac{1}{n}\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right) > \theta\right), \text{ by above} \\
\implies \exp(-nE_0) + \exp(-n(E_1 - \theta)) &\geq P\left(\frac{1}{n}\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right) \leq \theta\right) \\
\implies \frac{-1}{n}\log(\exp(-n\min\{E_0, E_1 - \theta\})(1 + e^{-n|E_0 - (E_1 - \theta)|})) &\leq -\frac{1}{n}\log\left(P\left(\frac{1}{n}\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right)\right) \leq \theta\right) \\
\implies \min\{E_0, E_1 - \theta\} - \frac{1}{n}\log(1 + e^{-n|E_0 - (E_1 - \theta)|}) &\leq -\frac{1}{n}\log\left(P\left(\frac{1}{n}\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right)\right) \leq \theta\right) \\
\implies \min\{E_0, E_1 - \theta\} &\leq \psi_P^*(\theta) \qquad (3)
\end{aligned}$$

where the last line follows from taking $n \to \infty$ and applying Theorem 62 (ie., Cramer's Theorem) to the RHS. If it's the case that $E_0 \geq D_{KL}(P_\lambda || P) + \epsilon$ and $E_1 \geq D_{KL}(P_\lambda || Q) + \epsilon$ for some $\epsilon > 0$, pick

$$\begin{aligned}
\theta &= D_{KL}(P_\lambda || Q) - D_{KL}(P_\lambda || P) \\
&= \mathbb{E}_{P_\lambda}\left[\log\left(\frac{dP}{dQ}\right)\right], \text{ as shown in the proof of Corollary 66}
\end{aligned}$$

Then, $\psi_P^*(\theta) = D_{KL}(P_\lambda || P)$ by Theorem 60 (ie., Maximum Entropy Distribution). That implies that

$$\min\{E_0, E_1 - \theta\} \geq \psi_P^*(\theta) + \epsilon$$

which is a contradiction with Equation (3).[b] □

[a] For $g(\lambda) = \mathbb{E}_{P_\lambda}\left[\log\left(\frac{dP}{dQ}\right)\right]$, we have that $g'(\lambda) = \mathrm{Var}_{P_\lambda}\left(\log\left(\frac{dP}{dQ}\right)\right) \geq 0$ with strict equality in the interesting case when $P \neq Q$.

[b] We can also consider possible improvements where $E_0 \geq D_{KL}(P_\lambda||P) + \epsilon$ and $E_1 = D_{KL}(P_\lambda||Q)$. By the continuity of $\lambda \mapsto D_{KL}(P_\lambda||P)$ and $\lambda \mapsto D_{KL}(P_\lambda||Q)$ and their monotonicity in opposite directions, we can find λ', ϵ' where $E_0 \geq D_{KL}(P_{\lambda'}||P) + \epsilon'$ and $E_1 \geq D_{KL}(P_{\lambda'}||Q) + \epsilon'$. Then we can apply this proof and invalidate these kinds of improvements. For edge cases when $E_0 = 0$ or $E_1 = 0$ – to argue that improvements aren't possible we can apply the results of Theorem 67 (ie., Stein Regime Theorem). For instance, we can use those results to assert that if $E_0 = 0$, then $E_1 = D_{KL}(P||Q) + o(1)$.

Corollary 66 ($E_0 - E_1$ Tradeoff Corollary)**.** Assume again that $P << Q$ and $Q << P$. Then,

$$\max_{(E_0,E_1)\text{ achievable}} \min\{E_0, E_1\} = -\inf_{\lambda\in[0,1]} \log\left(\int_{\mathcal{X}} (dP)^{1-\lambda}(dQ)^{\lambda}\right) \tag{4}$$

Proof.

For $P_\lambda = \frac{P^{1-\lambda}Q^{\lambda}}{Z(\lambda)}$ where $Z(\lambda) := \int_{\mathcal{X}} (dP)^{1-\lambda}(dQ)^{\lambda}$, we have that

$$\begin{aligned}
D_{KL}(P_\lambda||P) &= \mathbb{E}_{P_\lambda}\left[\log\left(\frac{dP_\lambda}{dP}\right)\right] \\
&= \mathbb{E}_{P_\lambda}\left[\lambda \log\left(\frac{dQ}{dP}\right) - \log(Z(\lambda))\right] \\
D_{KL}(P_\lambda||Q) &= \mathbb{E}_{P_\lambda}\left[\log\left(\frac{dP_\lambda}{dQ}\right)\right] \\
&= \mathbb{E}_{P_\lambda}\left[(1-\lambda) \log\left(\frac{dP}{dQ}\right) - \log(Z(\lambda))\right] \\
\implies \mathbb{E}_{P_\lambda}\left[\log\left(\frac{dQ}{dP}\right)\right] &= D_{KL}(P_\lambda||P) - D_{KL}(P_\lambda||Q)
\end{aligned}$$

Define the function $f : [0,1] \to \mathbb{R}$ by

$$\begin{aligned}
f(\lambda) &:= \log\left(Z(\lambda)\right) \\
&= \log\left(\int_{\mathcal{X}} (dP)^{1-\lambda}(dQ)^{\lambda}\right) \\
&= \log\left(\mathbb{E}_P\left[\exp\left(\lambda \log\left(\frac{dQ}{dP}\right)\right)\right]\right)
\end{aligned}$$

We note that

$$\begin{aligned}
f'(\lambda) &= \frac{d}{d\lambda}\left[\log\left(\int_{\mathcal{X}} (dP)^{1-\lambda}(dQ)^{\lambda}\right)\right]_{\lambda=\lambda^*} \\
&= \frac{1}{Z}\int_{\mathcal{X}} (dP)^{1-\lambda^*}(dQ)^{\lambda^*} \log\left(\frac{dQ}{dP}\right) \\
&= \mathbb{E}_{P_{\lambda^*}}\left[\log\left(\frac{dQ}{dP}\right)\right]
\end{aligned}$$

We also note that

$$\begin{aligned}
D_{KL}(P_\lambda||P) &= \lambda f'(\lambda) - f(\lambda) \\
D_{KL}(P_\lambda||Q) &= -(1-\lambda) f'(\lambda) - f(\lambda)
\end{aligned}$$

Let λ^* denote the minimizer of the convex function f on $[0,1]$.[a] Then, we have that the first order condition that determines λ^* is

$$\begin{aligned} 0 &= f'(\lambda) \\ &= \mathbb{E}_{P_{\lambda^*}}\left[\log\left(\frac{dQ}{dP}\right)\right], \text{ by above} \end{aligned}$$

For this λ^*, we have that $D_{KL}(P_{\lambda^*}||P) = D_{KL}(P_{\lambda^*}||Q)$ and

$$\begin{aligned} D_{KL}(P_{\lambda^*}||P) &= -\log(Z) \\ &= -\log\left(\int_{\mathcal{X}}(dP)^{1-\lambda^*}(dQ)^{\lambda^*}\right) \\ &= -\inf_{\lambda\in[0,1]}\log\left(\int_{\mathcal{X}}(dP)^{1-\lambda}(dQ)^{\lambda}\right), \text{ by definition of } \lambda^* \end{aligned}$$

By Theorem 65, the optimal tradeoff is characterized by $(E_0(\lambda), E_1(\lambda)) = (D_{KL}(P_\lambda||P), D_{KL}(P_\lambda||Q))$. We next can compute the following derivatives

$$\begin{aligned} \frac{d}{d\lambda}\left(D_{KL}(P_\lambda||P)\right) &= \frac{d}{d\lambda}\left(\lambda f'(\lambda) - f(\lambda)\right) \\ &= \lambda f''(\lambda) \\ &\geq 0, \text{ by convexity of } f \\ \frac{d}{d\lambda}\left(D_{KL}(P_\lambda||Q)\right) &= \frac{d}{d\lambda}\left(-(1-\lambda)f'(\lambda) - f(\lambda)\right) \\ &= -(1-\lambda)f''(\lambda) \\ &\leq 0, \text{ by convexity of } f \end{aligned}$$

In particular, the inequalities are strict whenever $P \neq Q$ so that $-\inf_{\lambda\in[0,1]}\log\left(\int_{\mathcal{X}}(dP)^{1-\lambda}(dQ)^{\lambda}\right)$ characterizes the optimal value of the LHS of Equation (4). □

[a] The function f is convex in $\lambda \in [0,1]$ since $f''(\lambda) = \text{Var}_{P_\lambda}\left(\log\left(\frac{dQ}{dP}\right)\right) \geq 0$ where P_λ is the exponential tilt of P along $X = \log\left(\frac{dQ}{dP}\right)$ as in Definition 59 and defined at the start of the proof.

4.5 Asymptotics: Stein Regime

Next, we will discuss the Stein regime. Consider the same situation where we observe a random variable X^n where $\text{supp}(X_i) = \mathcal{X}$ for each $i \in \{1, ..., n\}$. We consider the following two hypotheses:

- $H_0 : X^n \overset{iid}{\sim} P$ (ie., $X^n \sim P^{\otimes n}$)
- $H_1 : X^n \overset{iid}{\sim} Q$ (ie., $X^n \sim Q^{\otimes n}$)

For a fixed $\epsilon \in (0,1)$, for any test T_n if we force $\lim_{n\to\infty}\alpha(T_n) = 1-\epsilon$, what's the largest value of E_n^* such $\beta(T_n^*) = \exp(-nE_n^*)$ for some test T_n^*?

From Theorem 65 (ie., $E_0 - E_1$ Tradeoff), we already know that with $E_0 = 0$, we already know that $E_n^* = D_{KL}(P||Q) + o(1)$. Can we also get the next order term?

Theorem 67 (Stein Regime Theorem)**.** Given the setup above, we have that,

$$E_n^* = D_{KL}(P||Q) - \sqrt{\frac{V(P||Q)}{n}}\,\text{ercf}^{-1}(\epsilon) + o\left(\frac{1}{\sqrt{n}}\right)$$

where

$$
\begin{aligned}
\operatorname{ercf}(z) &= \Pr(\mathcal{N}(0,1) > z) \\
&= \int_z^\infty \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) dx \\
V(P||Q) &= \operatorname{Var}_P\left(\log\left(\frac{dP}{dQ}\right)\right)
\end{aligned}
$$

and we assume that $V(P||Q) < \infty$.

Proof.

[Achievability]:

Consider the test $T_n := \mathbb{1}_{\{\frac{1}{n}\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right) \leq \gamma\}}$. By the CLT, we have that under P,

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\log\left(\frac{dP}{dQ}(X_i)\right) - D_{KL}(P||Q)\right) \xrightarrow{d} \mathcal{N}(0, V(P||Q))
$$

As a result, by the continuous mapping theorem, we note that

$$
\lim_{n\to\infty} P(T_n = 1) = \Phi\left(\frac{\gamma - D_{KL}(P||Q)}{\sqrt{V(P||Q)/n}}\right)
$$

We wish for $\lim_{n\to\infty} P(T_n = 1) = 1 - \epsilon$, which implies that

$$
\begin{aligned}
&\Phi\left(\frac{\gamma - D_{KL}(P||Q)}{\sqrt{V(P||Q)/n}}\right) = 1 - \epsilon \\
&\qquad\qquad \implies \gamma = D_{KL}(P||Q) + \sqrt{\frac{V(P||Q)}{n}}\Phi^{-1}(1-\epsilon) \\
&\qquad\qquad\qquad\quad = D_{KL}(P||Q) - \sqrt{\frac{V(P||Q)}{n}}\operatorname{ercf}^{-1}(\epsilon)
\end{aligned}
$$

For $\beta(T_n)$,

$$
\begin{aligned}
\beta(T_n) &= Q\left(\frac{1}{n}\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right) > \gamma\right) \\
&\leq \exp(-n\gamma)\mathbb{E}_Q\left[\exp\left(\sum_{i=1}^n \log\left(\frac{dP}{dQ}(X_i)\right)\right)\right], \text{ by Markov inequality} \\
&= \exp(-n\gamma)
\end{aligned}
$$

[Converse]:

Suppose for the sake of contradiction that $E_n \geq D_{KL}(P||Q) + \frac{c}{\sqrt{n}}$ for some $c > -\sqrt{V(P||Q)}\operatorname{ercf}^{-1}(\epsilon)$ and some test T_n. Then, we have that for any $\delta > 0$ with $\lim_{n\to\infty} \alpha(T_n) = 1-\epsilon$, we get that

$$\begin{aligned}
1-\epsilon-o(1) &= \alpha(T_n) - \exp\left[n\left(D_{KL}(P||Q)+\frac{c-\delta}{\sqrt{n}}\right)\right]\beta(T_n), \text{ by assumption}\\
&\leq P\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{dP}{dQ}(X_i)\right) > D_{KL}(P||Q)+\frac{c-\delta}{\sqrt{n}}\right), \text{ by (a) in proof of Theorem 65}\\
&\xrightarrow{\mathbb{P}} \operatorname{ercf}\left(\frac{c-\delta}{\sqrt{V(P||Q)}}\right), \text{ under P and by CLT as } n\to\infty
\end{aligned}$$

With the initial LHS, taking $n \to \infty$ as well yields that

$$\begin{aligned}
1-\epsilon &\leq \operatorname{ercf}\left(\frac{c-\delta}{\sqrt{V(P||Q)}}\right)\\
\implies c &\leq -\sqrt{V(P||Q)}\operatorname{ercf}^{-1}(\epsilon)+\delta
\end{aligned}$$

Taking $\delta \to 0^+$ yields that

$$c \leq -\sqrt{V(P||Q)}\operatorname{ercf}^{-1}(\epsilon)$$

which contradicts the initial assumption. □

5 Functional (In)equalities

Recall from Section 1.3 that inequalities that can be shown via (a) monotonicity (ie., $H(X|Y) \leq H(X)$) and (b) submodularity (ie., $H(X_A) + H(X_B) \geq H(X_{A\cup B} + H(X_{A\cap B})))$ are called *Shannon-type inequalities*. We now focus on non-*Shannon-type inequalities*.

Definition 68 (Differential Entropy)**.** For a random variable X with density f on $\mathbb{R}^d$, its differential entropy is defined as

$$\begin{aligned}
h(X) &:= h(f)\\
&= \int_{\mathbb{R}^d} -f(x)\log(f(X))dx
\end{aligned}$$

To make some remarks, we note that:

- $h(X) \in \mathbb{R}\cup\{-\infty,+\infty\}$.
- $h(aX) = h(X) + \log(a)$ for $a \in \mathbb{R}$
- $h(X) \leq h(X,Y)$ no longer holds.
- It is still true that $I(X;Y) = h(X) + h(Y) - h(X,Y) \geq 0$.

Example 69 (Normal Random Variable Differential Entropy)**.** If $X \sim \mathcal{N}(\mu,\Sigma)$ and

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d\det(\Sigma)}}\exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

so

$$\begin{aligned} h(X) &= \mathbb{E}_X\left[\frac{1}{2}\log\left((2\pi)^d\det(\Sigma)\right) + \frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right] \\ &= \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\left(\det(\Sigma)\right) \\ &= \frac{1}{2}\log\left((2\pi e)^d\det(\Sigma)\right) \end{aligned}$$

One can use the fact that the Trace$(\cdot)$ is invariant to circular rotations to the second equality.

Definition 70 (Fisher Information of a Random Variable)**.** For a random variable X on $\mathcal{X}$ with density f with respect to the measure μ, the Fisher information is

$$J(X) := \int_{\mathcal{X}} \frac{(f_x)(f_x)'}{f}\mu(dx)$$

Remark 71 (Connecting Fisher Information of Random Variable to Fisher Information of Parametric Model)**.** Recall from Section 3.4, we defined that for $Y \sim P_\theta$ on $\mathcal{X}$ with density $D_\theta P =: f_\theta$ with respect to μ, $I(\theta) := I^Y(\theta) := \int_{\mathcal{X}} \frac{(D_\theta f_\theta)(D_\theta f_\theta)'}{f_\theta}\mu(dy)$.

They are connected via $Y = \theta + X$ when $\mathcal{X} = \mathbb{R}^d$ which is the same space where θ resides. In that case $f_\theta(y) = f(y-\theta)$. Then, $D_\theta f_\theta(y) = -f_x(y-\theta)$. Thus,

$$\begin{aligned} I^Y(\theta) &= \int_{\mathcal{X}} \frac{(D_\theta f_\theta)(D_\theta f_\theta)'}{f_\theta}\mu(dy) \\ &= \int_{\mathcal{X}} \frac{(-f_x)(-f_x)'}{f_x}\mu(dx) \\ &= J(X) \end{aligned}$$

5.1 Towards the Entropy Power Inequality

Here, we will proceed towards proving the Entropy Power Inequality (ie., Theorem 76).

Property 72 (Scalar Factor on Fisher Information of a Random Variable)**.** We have that for $a \in \mathbb{R}\setminus\{0\}$, $J(aX) = \frac{1}{a^2}J(X)$.

Proof.

For the derivation, suppose that X on $\mathbb{R}^d$ has density $f_X(x)$ with respect to the Lesbegue measure. Then aX has density $\frac{1}{|a|}f_X(x/a)$ with respect to the same measure. For notational ease here, we will let f'_X be the derivative of the density.

$$\begin{aligned} J(aX) &= \int_{\mathcal{X}} \frac{\frac{1}{a^4}(f'_X(x/a))(f'_x(x/a))'}{\frac{1}{|a|}f_X(x/a)}dx \\ &= \int_{\mathcal{X}} \frac{\frac{1}{a^4}(f'_X(x))(f'_x(x))'}{\frac{1}{|a|}f_X(x)}|a|dx \\ &= \frac{1}{a^2}\int_{\mathcal{X}} \frac{(f'_X(x))(f'_x(x))'}{f_X(x)}dx \\ &= \frac{1}{a^2}J(X) \end{aligned}$$

□

Property 73 (Fisher Information of Random Variable Markov Chain)**.** If $\theta - X - Y$ is a Markov chain for a scalar θ, then $I^Y(\theta) \leq I^X(\theta)$.

Proof.

By the data processing inequality of f-divergences, we have that for any Δ,

$$\begin{aligned} \frac{1}{\Delta^2}\chi^2\left(P_{Y|\theta+\Delta}||P_{Y|\theta}\right) &\leq \frac{1}{\Delta^2}\chi^2\left(P_{X|\theta+\Delta}||P_{X|\theta}\right) \\ \implies I^Y(\theta) &= \lim_{\Delta\to 0}\frac{1}{\Delta^2}\chi^2\left(P_{Y|\theta+\Delta}||P_{Y|\theta}\right) \\ &\leq \lim_{\Delta\to 0}\frac{1}{\Delta^2}\chi^2\left(P_{X|\theta+\Delta}||P_{X|\theta}\right) \\ &= I^X(\theta) \end{aligned}$$

By transitivity, we have the result.[a] □

[a] We use the factorization of χ^2 divergence in Section 3.4 for a scalar θ.

Theorem 74 (Stam Theorem of Fisher Information of Random Variables)**.** For independent random variables $X_1, X_2 \in \mathbb{R}^d$, we have that

$$\begin{aligned} \frac{1}{J(X_1+X_2)} &\geq \frac{1}{J(X_1)} + \frac{1}{J(X_2)} \\ \iff a^2J(X_1) + b^2J(X_2) &\geq (a+b)^2J(X_1+X_2)\ \forall a,b>0 \end{aligned}$$

Proof.

Take any $a, b > 0$ and $\theta \in \mathbb{R}^d$. Define $Y_1 = a\theta + X_1$ and $Y_2 = b\theta + X_2$. Then, we have that

$$\begin{aligned} I^{Y_1}(\theta) &= I^{Y_1/a}(\theta), \text{ by Remark 71} \\ &= J(X_1/a), \text{ by Remark 71} \\ &= a^2J(X_1), \text{ by Property 72} \end{aligned}$$

Therefore, we have that

$$\begin{aligned} (a+b)^2J(X_1+X_2) &= I^{Y_1+Y_2}(\theta) \\ &\leq I^{Y_1,Y_2}(\theta), \text{ by data processing inequality of Fisher Information (unproven here)} \\ &= a^2J(X_1) + b^2J(X_2) \end{aligned}$$

The equivalence with the other characterization is just cheeky algebra and so will be omitted here. As a hint, begin by dividing both sides of the known inequality by ab and defining $\lambda = \frac{a}{b}$ and optimizing the LHS with respect to λ to make the bound as tight as possible. □

Theorem 75 (de Bruijn)**.** Suppose X is some real valued random variable with a "well-behaved" density p and $Z \sim \mathcal{N}(0,1)$ with $X \perp\!\!\!\perp Z$. Then for any $a > 0$, we have that

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

where $h(\cdot)$ is the differential entropy operator.

Proof.

Let $p_a = p * \mathcal{N}(0,a)$ be the density of $X + Z\sqrt{a}$ where $*$ is the convolution operator. Then, we have that

$$\frac{\partial p_a}{\partial a} = \frac{1}{2} p_a''$$

To see that, note that for any twice differentiable and integrable testing function f,[a]

$$\begin{aligned}
\frac{\partial}{\partial a}\mathbb{E}_{p_a}[f] &= \lim_{\Delta \to 0} \frac{1}{\Delta}\mathbb{E}_{p_a}\left[f(X + Z\sqrt{a+\Delta})\right] \\
&= \lim_{\Delta \to 0} \frac{1}{\Delta}\mathbb{E}_{p_a, Z'}\left[f(X + Z\sqrt{a} + Z'\sqrt{\Delta}) - f(X + Z\sqrt{a})\right],\ Z' \overset{indep}{\sim} \mathcal{N}(0,1) \\
&= \lim_{\Delta \to 0} \mathbb{E}_{p_a}\left[f'(X + Z\sqrt{a})\left(Z'\sqrt{\Delta}\right) + \frac{1}{2} f''\left(X + Z\sqrt{a}\right)\Delta\left(Z'\right)^2 + o(\Delta)\right] \\
&= \frac{1}{2}\mathbb{E}_{p_a}\left[f''\right] \\
&= \frac{1}{2}\int_{\mathbb{R}} f p_a'', \text{ using integration by parts}
\end{aligned}$$

As a result, we have that for any suitable function f

$$\begin{aligned}
\frac{\partial}{\partial a}\mathbb{E}_{p_a}[f] &= \int_{\mathbb{R}} f(y) \frac{\partial p_a(y)}{\partial a} dy \\
&= \frac{1}{2}\int_{\mathbb{R}} f(y) \frac{\partial^2 p_a(y)}{\partial a^2} dy, \text{ by above} \\
\implies \frac{\partial p_a(y)}{\partial a} &= \frac{1}{2}\frac{\partial^2 p_a(y)}{\partial a^2}, \text{ by taking } f(y) = \frac{\partial p_a(y)}{\partial a} - \frac{\partial^2 p_a(y)}{\partial a^2}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{d}{da} h(X + Z\sqrt{a}) &= -\int_{\mathbb{R}} (1 + \log(p_a)) \frac{\partial p_a}{\partial a} \\
&= -\frac{1}{2}\int_{\mathbb{R}} (1 + \log(p_a)) p_a'' \\
&= \frac{1}{2}\int_{\mathbb{R}} \frac{(p_a)^2}{p_a}, \text{ using integration by parts} \\
&= \frac{1}{2} J(X + Z\sqrt{a})
\end{aligned}$$

By transitivity, we conclude the proof.[b] □

[a] The second step below follows from the fact that $X + Z\sqrt{a+\Delta} =_D X + Z\sqrt{a} + Z'\sqrt{\Delta}$ where $Z' \overset{indep}{\sim} \mathcal{N}(0,1)$. The third step follows from the Taylor expansion of the second line on $\sqrt{\Delta}$ around $X + Z\sqrt{a}$ – it uses the fact that $\mathbb{E}'_Z[Z'] = 0$ and $\mathbb{E}'_Z[(Z')^2] = 1$.

[b] Intentionally, we left vague what it means for p to be "well-behaved". We need some well-behaved tails to support the last integration by parts in the proof.

Theorem 76 (Entropy Power Inequality (EPI))**.** For independent random variables X, Y on $\mathbb{R}^d$

$$\exp\left(\frac{2}{d}h(X+Y)\right) \geq \exp\left(\frac{2}{d}h(X)\right) + \exp\left(\frac{2}{d}h(Y)\right)$$

That equality holds if and only if X, Y are Gaussian and $\Sigma_X = c\Sigma_Y$ for $c > 0$. Thus, the EPI shows that for given values of $h(X)$ and $h(Y)$, $h(X+Y)$ is minimized when X, Y are Gaussian.

Proof.

[General Inequality]:

We will proceed by induction on d.

Base Case ($d=1$):

Let $X_\lambda := X * \mathcal{N}(0, f(\lambda))$ and $Y_\lambda := Y * \mathcal{N}(0, g(\lambda))$ for some functions f, g that are TBD. Since $\frac{d}{d\lambda} \exp(2h(X_\lambda)) = \exp(2h(X_\lambda))J(X_\lambda)f'(\lambda)$ by Theorem 75 (ie., de Bruijn), we have that

$$\frac{d}{d\lambda}\left[\frac{\exp(2h(X_\lambda)) + \exp(2h(Y_\lambda))}{\exp(2h(X_\lambda + Y_\lambda))}\right] = \frac{[\exp(2h(X_\lambda))J(X_\lambda)f'(\lambda) + \exp(2h(Y_\lambda))J(Y_\lambda)g'(\lambda)]}{\exp(2h(X_\lambda + Y_\lambda))} - \frac{[(\exp(2h(X_\lambda)) + 2\exp(2h(Y_\lambda)))\,J(X_\lambda + Y_\lambda)(f'(\lambda) + g'(\lambda))]}{\exp(2h(X_\lambda + Y_\lambda))}$$

after a little simplification. Choosing $f'(\lambda) = \exp(2h(X_\lambda))$ and $g'(\lambda) = \exp(2h(Y_\lambda))$ yields that

$$\frac{d}{d\lambda}\left[\frac{\exp(2h(X_\lambda)) + \exp(2h(Y_\lambda))}{\exp(2h(X_\lambda + Y_\lambda))}\right] \geq 0, \text{ invoking Property 74}$$

As $\lambda \to \infty$, both X_λ and Y_λ are "more and more Gaussian" and "more and more independent" as $f(\lambda), g(\lambda) \to \infty$. The ratio

$$\frac{\exp(2h(X_\lambda)) + \exp(2h(Y_\lambda))}{\exp(2h(X_\lambda + Y_\lambda))}$$

approaches 1 from below by the condition on the derivative as $\lambda \to \infty$.[a] Therefore, at $\lambda = 0$, the ratio must be ≤ 1, which implies the EPI for $d = 1$.

Inductive Step:

Let $X, Y \in \mathbb{R}^d$ where $d \geq 2$. We let X^c denote the first c components of X (same for Y).

$$\begin{aligned}
h(X^d + Y^d) &= h(X^{d-1} + Y^{d-1} \mid h(X_d + Y_d \mid X^{d-1} + Y^{d-1})) \\
&\geq h(X^{d-1} + Y^{d-1}) + h(X_d + Y_y \mid X^{d-1}, Y^{d-1}), \text{ conditioning reduces entropy} \\
&\geq \frac{d-1}{2} \log\left[e^{\frac{2}{d-1}h(X^{d-1}) + \frac{2}{d-1}h(Y^{d-1})}\right], \text{ by inductive hypothesis} \\
&\quad + \frac{1}{2}\mathbb{E}_{X^{d-1}, Y^{d-1}}\left[\log\left(e^{2h(X_d \mid X^{d-1} = x^{d-1})}) + e^{2h(Y_d \mid Y^{d-1} = y^{d-1})}\right)\right], \text{ since } X \perp\!\!\!\perp Y \\
&\geq \frac{d-1}{2} \log\left[e^{\frac{2}{d-1}h(X^{d-1}) + \frac{2}{d-1}h(Y^{d-1})}\right] \\
&\quad + \frac{1}{2}\log\left(e^{2h(X_d \mid X^{d-1})}) + e^{2h(Y_d \mid Y^{d-1})}\right), (x, y) \mapsto \log\left(e^x + e^y\right) \text{ convex} \\
&\geq \frac{d}{2}\log\left(e^{\frac{2}{d}h(X^{d-1}) + \frac{2}{d}h(X_d \mid X^{d-1})} + e^{\frac{2}{d}h(Y^{d-1}) + \frac{2}{d}h(Y_d \mid Y^{d-1})}\right), (x, y) \mapsto \log\left(e^x + e^y\right) \text{ convex} \\
&= \frac{d}{2}\log\left(e^{\frac{d}{2}h(X^d)} + e^{\frac{2}{d}h(Y^d)}\right)
\end{aligned}$$

By transitivity, we have the shown the inductive step concluding our proof by induction.[b]

Equality when Independent Gaussians with Scalar Apart Covariance Matrix:

We will prove just one direction here. The other direction is omitted (for now).

Independent Gaussians with Scalar Apart Covariance Matrix $\implies$ Equality:

Suppose that $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \Sigma_Y)$ are independent Gaussian random variables on $\mathbb{R}^d$ such that $\Sigma_X = c\Sigma_Y$ for $c > 0$. We have that $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, (1 + c)\Sigma_X)$ since X and Y are independent. We also have that, by Example 69,

$$h(X) = \frac{1}{2} \log \left((2\pi e)^d \det(\Sigma_X) \right)$$

We then have that

$$\begin{aligned}
\exp\left(\frac{2}{d} h(X+Y)\right) &= \exp\left(\frac{2}{d}\left(\frac{1}{2}\log\left((2\pi e)^d \det((1+c)\Sigma_X)\right)\right)\right) \\
&= (2\pi e)(1+c)\det(\Sigma_X)^{1/d} \\
&= 2\pi e \det(\Sigma_X)^{1/d} + 2\pi e \det(c\Sigma_X)^{1/d} \\
&= \exp\left(\frac{2}{d} h(X)\right) + \exp\left(\frac{2}{d} h(Y)\right)
\end{aligned}$$

Equality $\implies$ Independent Gaussians with Scalar Apart Covariance Matrix:

Omitted. □

[a] We note that $\exp(2h(A + B)) = \exp(2h(A)) + \exp(2h(B))$ for univariate independent Gaussians A, B.

[b] The last step that uses convexity with weights $(\frac{d-1}{d}, \frac{1}{d})$ on each of the terms respectively.

Example 77 (Entropic CLT)**.** Let $X_1, ..., X_n$ be iid with $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}(X_1) = 1$ and $h(X_1) > -\infty$. Let $T_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ by the standardized sum. Then by Theorem 76 (ie., EPI), we have that

$$\begin{aligned}
h(T_{n+m}) &= h\left(\sqrt{\frac{m}{n+m}} \frac{1}{\sqrt{m}} \sum_{i=1}^m X_i + \sqrt{\frac{n}{n+m}} \frac{1}{\sqrt{n}} \sum_{i=m+1}^{n+m} X_i\right) \\
&\geq \frac{1}{2} \log\left(e^{2h\left(T_m \sqrt{\frac{m}{n+m}}\right)} + e^{2h\left(T_n \sqrt{\frac{n}{n+m}}\right)}\right) \\
&= \frac{1}{2} \log\left(\frac{m}{n+m} e^{2h(T_m)} + \frac{n}{n+m} e^{h(T_n)}\right), \text{ since for } a \in \mathbb{R}, h(aZ) = h(Z) + \log(|a|)
\end{aligned}$$

In other words, the sequence $a_n := n \exp(2h(T_n))$ is super-additive (ie., $a_{n+m} \geq a_n + a_m \; \forall n, m$). Moreover, since $\mathrm{Var}(T_n) = 1$, the maximum entropy principle (exercise) implies that

$$h(T_n) \leq \frac{1}{2} \log (2\pi e)$$

so that $\frac{a_n}{n} \leq 2\pi e$. Therefore, $\frac{a_n}{n}$ must have a limit (ie., $h(T_n) \to h^*$) since it's bounded from above and non-decreasing and

$$\begin{aligned}
D_{KL}(P_{T_n}||\mathcal{N}(0,1)) &= -h(T_n) - \int_{\mathbb{R}} f_{T_n}(x)\log(\phi(x))dx, \text{ where } \phi \text{ is the pdf of } \mathcal{N}(0,1) \\
&= -h(T_n) + \int_{\mathbb{R}} f_{T_n}(x)\left[\frac{1}{2}\log(2\pi) + \frac{1}{2}x^2\right]dx \\
&= -h(T_n) + \frac{1}{2}\log(2\pi) + \frac{1}{2}\cancelto{1}{\mathbb{E}_{X_1}[X_1^2]} \\
&= -h(T_n) + \frac{1}{2}\log(2\pi e) \\
&\to D^*
\end{aligned}$$

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

Theorem 78 (Maximum Entropy Principle)**.** Let X be any random variable with $\text{Cov}(X) = \Sigma$, $\text{mean}(X) = \mu$, and support $\mathbb{R}^d$. We define $Y \sim \mathcal{N}(0,\Sigma)$, $Z \sim \mathcal{N}(\mu,\Sigma)$, and denote the distribution law of random variable $A \in \{X, Y, Z\}$ by P_A. We have that

$$h(X) \leq h(Y)$$

Proof.

we have that

$$\begin{aligned}
0 &\leq D_{KL}(P_X||P_Z), \text{ by Property 16} \\
&= \int_{\mathbb{R}^d} \log\left(\frac{dP_X(x)}{dP_Z(x)}\right)dP_X(x) \\
&= -h(X) + \int_{\mathbb{R}^d} \log\left(\frac{1}{dP_Z(x)}\right)dP_X(x) \\
&= -h(X) - \mathbb{E}_{P_X}\left[\log(dP_Z(X))\right] \\
&= -h(X) + \frac{1}{2}\log\left((2\pi e)^d \det(\Sigma)\right), \text{ using a derivation like in Example 69} \\
&= -h(X) + h(Y)
\end{aligned}$$

By transitivity, we have the result. □

5.2 *Information and Estimation in the Gaussian Model*

Let X be a general random variable and $Z \overset{indep}{\sim} \mathcal{N}(0,1)$. We next define the random variable $Y_\gamma = X\sqrt{\gamma} + Z$ where $\gamma > 0$ is a signal-to-noise (SNR) parameter. Our aim of this section is to provide an information theoretic characterization of the minimum mean squared error (MMSE) for estimating X based on Y_γ. We will provide a proof of a more general result using filtering theory for Brownian motions.

Definition 79 (Minimum Mean Squared Error (MMSE))**.** For random variables X, Y we define the MMSE of X given Y as

$$\text{mmse}(X \mid Y_\gamma) := \mathbb{E}\left[(X - \mathbb{E}[X \mid Y_\gamma])^2\right]$$

Sometimes, when the context is clear, we will abbreivate $\text{mmse}(X \mid Y_\gamma) = \text{mmse}(\gamma)$.

Definition 80 (Stochastic Process Adapted to Filtration)**.** A stochastic process X_t is adapted to a filtration $\mathcal{F}_t$ if for every t, the random variable X_t is measurable with respect to $\mathcal{F}_t$. A random variable X_t is measurable with respect to $\mathcal{F}_t$ if $X_t^{-1}(B) \in \mathcal{F}_t$ for every Borel set B.

Lemma 81 (Brownian Motion Filtering Lemma)**.** For $dY_t = f(t)dt + dB_t$ with $f(t)$ adapted to the filtration $\mathcal{F}^Y := \sigma\left(\{Y_s : s \leq t\}\right)$, then

$$\log\left(\frac{dP_{Y^T}}{dP_{B^T}}\right) = \int_0^T f(t)d\xi_t - \frac{1}{2}\int_0^T f(t)^2 dt$$

Proof (Sketch).

For $t \geq 0$ and small $\Delta > 0$, the conditional distribution of $\xi_{t+\Delta} - \xi_t \mid \xi^t$ (where ξ_t is the sample path of the process) is

$$\begin{cases} \mathcal{N}\left(\int_t^{t+\Delta} f(s)ds, \Delta\right) & \text{under } P_{Y^T} \\ \mathcal{N}(0, \Delta) & \text{under } P_{B^T} \end{cases}$$

by the adaptedness of $f(t)$ to $\mathcal{F}^Y$. So the log-likelihood ratio is (letting $\mu = \int_t^{t+\Delta} f(s)ds$)

$$\begin{aligned} \log\left(\frac{\mathcal{N}(\int_t^{t+\Delta} f(s)ds, \Delta)}{\mathcal{N}(0, \Delta)}\right) &= \frac{1}{\Delta}\int_t^{t+\Delta} f(s)ds(\xi_{t+\Delta} - \xi_t) - \frac{1}{2\Delta}\left(\int_t^{t+\Delta} f(s)\right)^2 \\ &\approx f(t)(\xi_{t+\Delta} - \xi_t) - \frac{\Delta}{2}f(t)^2 \end{aligned}$$

Next, imagine dividing $[0, T]$ into N small chunks. We can write

$$\begin{aligned} \log\left(\frac{dP_{Y^T}}{dP_{B^T}}\right) &= \lim_{\Delta \to 0}\sum_{i=1}^N \log\left(\frac{dP_{Y^{t_i+\Delta}|Y^{t_i}}}{dP_{B^{t_i+\Delta}|B^{t_i}}}\right) \\ &\approx \lim_{\Delta \to 0}\sum_{i=1}^N f(t_i)(\xi_{t_i+\Delta} - \xi_{t_i}) - \frac{\Delta}{2}f(t_i)^2 \\ &= \int_0^T f(t)d\xi_t - \frac{1}{2}\int_0^T f(t)^2 dt \end{aligned}$$

□

Lemma 82 (Brownian Motion Conditional Mean Adaptation Lemma)**.** For $dY_t = X_t dt + dB_t$, then $\tilde{B}_t := Y_t - \int_0^t \mathbb{E}[X_s \mid Y^s]ds$ is a Brownian motion adapted to $\mathcal{F}^Y := \sigma\left(\{Y_s : s \leq t\}\right)$.[a]

Proof.

Clearly $\tilde{B}_t$ is adapted to $\mathcal{F}^Y$ since it's a predictable function of Y^t. In addition

$$\begin{aligned} \tilde{B}_t &= Y_t - \int_0^t \mathbb{E}[X_s \mid Y^s]ds \\ &= \int_0^t X_s ds + B_t - \int_0^t \mathbb{E}[X_s \mid Y^s]ds \\ &= \int_0^t (X_s - \mathbb{E}[X_s \mid Y^s])ds + B_t \end{aligned}$$

is an $\mathcal{F}^Y$ adapted martingale that satisfied $\tilde{B}_0 = 0$ and has quadratic variation t. Thus, by the Levy criterion,[b] $\tilde{B}_t$ is a Brownian motion.

[a] A remark 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 $\mathcal{F}^Y$.

[b] The technical details of this criterion are omitted.

Theorem 83 (SDE I-MMSE). If $dY_t = X_t dt + dB_t$ for $t \in [0, T]$, then

$$I(X^T; Y^T) = \frac{1}{2} \int_0^T \mathbb{E}\left[(X_t - \mathbb{E}[X_t \mid Y^t])^2\right] dt$$

Proof.

We have that

$$\begin{aligned} I(X^T; Y^T) &= \mathbb{E}_{P_{X^T, Y^T}}\left[\log\left(\frac{dP_{Y^T|X^T}}{dP_{Y^T}}\right)\right] \\ &= \mathbb{E}_{P_{X^T, Y^T}}\left[\log\left(\frac{dP_{Y^T|X^T}}{dP_{B^T}}\right)\right] - \mathbb{E}_{P_{X^T, Y^T}}\left[\log\left(\frac{dP_{Y^T}}{dP_{B^T}}\right)\right] \end{aligned}$$

For the first term, since X^T is given, Lemma 81 gives that

$$\mathbb{E}_{P_{X^T, Y^T}}\left[\log\left(\frac{dP_{Y^T|X^T}}{dP_{B^T}}\right)\right] = \mathbb{E}\left[\int_0^T X_t dY_t - \frac{1}{2}\int_0^T X_t^2 dt\right]$$

For the second term, Lemma 82 tells us that

$$\tilde{B}_t = Y_t - \int_0^t \mathbb{E}[X_s \mid Y^s] ds$$

is a $\mathcal{F}^Y$ adapted Brownian motion. So,

$$\begin{aligned} \log\left(\frac{dP_{Y^T}}{dP_{B^T}}(\xi^T)\right) &= \log\left(\frac{dP_{Y^T}}{dP_{\tilde{B}^T}}(\xi^T)\right) \\ &= \int_0^T \mathbb{E}[X_t \mid Y^t] d\xi_t - \frac{1}{2}\int_0^T \left(\mathbb{E}[X_t \mid Y^t]\right)^2 dt, \text{ by Lemma 81} \\ \implies \mathbb{E}\left[\log\left(\frac{dP_{Y^T}}{dP_{B^T}}\right)\right] &= \mathbb{E}\left[\int_0^T \mathbb{E}[X_t \mid Y^t] dY_t - \frac{1}{2}\int_0^T (\mathbb{E}[X_t \mid Y^t])^2 dt\right] \end{aligned}$$

In the application of Lemma 81, we use the fact that $dY_t = \mathbb{E}[X_t \mid Y^t]dt + d\tilde{B}_t$. Therefore,

$$\begin{aligned} I(X^T; Y^T) &= \mathbb{E}\left[\int_0^T (X_t - \mathbb{E}[X_t \mid Y^t]) dY_t + \frac{1}{2}\int_0^T ((\mathbb{E}[X_t \mid Y^t])^2 - X_t^2) dt\right] \\ &= \mathbb{E}\left[\int_0^T \left(X_t - \mathbb{E}[X_t \mid Y^t]\right) X_t + \frac{1}{2}((\mathbb{E}[X_t \mid Y^t])^2 - X_t^2) dt\right] \\ &= \int_0^T \frac{1}{2}\mathbb{E}\left[(X_t - \mathbb{E}[X_t \mid Y^t])^2\right] dt \end{aligned}$$

□

Corollary 84 (I-MMSE)**.** Recall the definitions at the start of this section. Let X be a general random variable and $Z \overset{indep}{\sim} \mathcal{N}(0,1)$. We next define the random variable $Y_\gamma = X\sqrt{\gamma} + Z$ where $\gamma > 0$ is a signal-to-noise (SNR) parameter. We have that

$$\frac{d}{d\gamma} I(X;Y_\gamma) = \frac{1}{2}\,\mathrm{mmse}(X \mid Y_\gamma)$$

Proof.

In Theorem 83, take $X_t := X$. Then Y_T is a sufficient statistic of Y^T for estimating X. In other words, $I(X^T;Y^T) = I(X;Y_T)$ and $\mathbb{E}[X_t \mid Y^t] = \mathbb{E}[X \mid Y_t]\ \forall t \in [0,T]$. Therefore, applying Theorem 83,

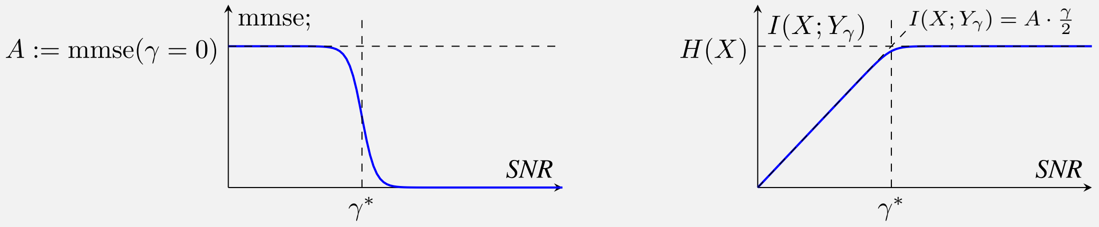
$$\begin{aligned} I(X;Y_T) &= I(X^T;Y^T) \\ &= \frac{1}{2}\int_0^T \mathbb{E}\left[(X_t - \mathbb{E}[X_t \mid Y^t])^2\right] dt \\ &= \frac{1}{2}\int_0^T \mathbb{E}[(X - \mathbb{E}[X \mid Y_t])^2] dt \\ &= \frac{1}{2}\int_0^T \mathrm{mmse}(X \mid Y_t) dt \\ \implies \frac{dI(X;Y_T)}{dT} &= \frac{1}{2}\,\mathrm{mmse}(X \mid Y_T),\ \text{by the Fundamental Theorem of Calculus} \end{aligned}$$

Moreover, we have that

$$\begin{aligned} Y_T &= XT + B_T \\ \implies \underbrace{\frac{Y_T}{\sqrt{T}}}_{=:Y_\gamma} &= X\underbrace{\sqrt{T}}_{=:\sqrt{\gamma}} + \mathcal{N}(0,1) \end{aligned}$$

so that the SNR parameter is T. □

Application 85 (I-MMSE for Sharp Phase Transition)**.** Suppose we expect a problem to have a sharp phase transition at SNR $= \gamma^*$ meaning that for $\gamma < \gamma^* - \delta$ we don't expect to recover X from Y_γ but at $\gamma > \gamma^* + \delta$ we do expect to be able.



In this case for some small $\epsilon > 0$ characterizing the "sharp" phase transition,

$$\begin{aligned}
\frac{(1-\epsilon)\gamma^*}{2}\,\mathrm{mmse}(0)(1-o(1)) &\leq I(X;Y_{(1-\epsilon)\gamma^*}) \\
&= \frac{1}{2}\int_0^{(1-\epsilon)\gamma^*} \mathrm{mmse}(\gamma)d\gamma, \text{ by Corollary 84} \\
&= \frac{1}{2}\int_0^{(1-2\epsilon)\gamma^*} \mathrm{mmse}(\gamma)d\gamma + \frac{1}{2}\int_{(1-2\epsilon)\gamma^*}^{(1-\epsilon)\gamma^*} \mathrm{mmse}(\gamma)d\gamma \\
&\leq \frac{(1-2\epsilon)\gamma^*}{2}\,\mathrm{mmse}(0) + \frac{\epsilon\gamma^*}{2}\,\mathrm{mmse}((1-2\epsilon)\gamma), \text{ since } \mathrm{mmse}(\cdot) \text{ is nonincreasing} \\
\implies \frac{\epsilon\gamma^*}{2}\,\mathrm{mmse}((1-2\epsilon)\gamma^*) &\geq \frac{(1-\epsilon)\gamma^*}{2}\,\mathrm{mmse}(0)(1-o(1)) - \frac{(1-2\epsilon)\gamma^*}{2}\,\mathrm{mmse}(0), \text{ by transitivity} \\
&= \frac{\gamma^*}{2}\left[(1-\epsilon)(1-o(1)) - (1-2\epsilon)\right]\mathrm{mmse}(0) \\
&= \frac{\epsilon\gamma^*}{2}(1-o(1))\,\mathrm{mmse}(0) \\
\implies \mathrm{mmse}((1-2\epsilon)\gamma^*) &\geq (1-o(1))\,\mathrm{mmse}(0)
\end{aligned}$$

In other words, the $\mathrm{mmse}(\cdot)$ does not really drop before $\gamma = \gamma^*$

Recall that a high-level, Theorem 13 (ie., Fano's inequality) shows that the estimation error is large when the information $I(X;Y)$ is small. Surprisingly, the I-MMSE formula in Application 85 shows that this is also possible when $I(X;Y)$ is large if it's berfore a sharp phase transition.

Example 86 (Sparse Mean Estimation Problem)**.** Consider the "sparse" mean estimation problem $Y \sim \mathcal{N}(\theta, 1)$ with $\theta \sim (1-\rho)\delta_0 + \rho\delta_\mu$ with $\rho = o(1)$. If $\mu \leq \sqrt{2(1-\epsilon)\log\left(\frac{1}{\rho}\right)}$ for some $\epsilon > 0$, then

$$\begin{aligned}
\mathrm{mmse}(\theta \mid Y) &\geq (1-o(1))\mathbb{E}[\theta^2] \\
&= (1-o(1))\rho\mu^2
\end{aligned}$$

In other words, the mmse is essentially attained by the best estimator $\hat{\theta} = \rho\mu$ without seeing Y.

Proof (Sketch).

Let $X \sim (1-\rho)\delta_0 + \rho\delta_1$ so that $\theta = X\mu$ and define $\mu =: \sqrt{\gamma}$. Then, $Y =_d Y_\gamma := X\sqrt{\gamma} + \mathcal{N}(0,1)$. The mutual information can be computed as (for any Q)

$$\begin{aligned}
I(X;Y_\gamma) &= \mathbb{E}\left[\log\left(\frac{dP_{Y_\gamma|X}}{dP_{Y_\gamma}}\right)\right] \\
&= \mathbb{E}\left[\log\left(\frac{dP_{Y_\gamma|X}}{dQ_{Y\gamma}}\right)\right] - D_{KL}(P_{Y_\gamma}||Q_{Y_\gamma})
\end{aligned}$$

Choosing $Q_{Y_\gamma} = \mathcal{N}(\rho\sqrt{\gamma}, 1)$, then

$$\begin{aligned}
\mathbb{E}\left[\log\left(\frac{dP_{Y_\gamma|X}}{dQ_{Y_\gamma}}\right)\right] &= \mathbb{E}[D_{KL}(P_{Y_\gamma|X}||Q_{Y_\gamma})] \\
&= \mathbb{E}\left[\frac{\left(X\sqrt{\gamma} - \rho\sqrt{\gamma}\right)^2}{2}\right], \text{ since } D_{KL}(\mathcal{N}(\mu_1,\sigma^2)||\mathcal{N}(\mu_2,\sigma^2)) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2} \\
&= \frac{\rho(1-\rho)}{2}\gamma
\end{aligned}$$

We also see that

$$D_{KL}(P_{Y_\gamma}||Q_{Y_\gamma}) = o(\rho\gamma)$$

after some algebra if $\gamma < 2(1-\epsilon)\log\left(\frac{1}{\rho}\right)$ so as to bound the tail behavior of Y_γ. That implies that

$$I(X, Y_\gamma) \geq \frac{\rho(1-\rho)}{2}\gamma(1-o(1)) \; \forall \gamma > 0$$

so long as $\gamma < 2(1-\epsilon)\log\left(\frac{1}{\rho}\right)$. Now, the concavity of $\gamma \mapsto I(X;Y_\gamma)$ that arises from the nonincreasingness of $\gamma \mapsto \operatorname{mmse}(\gamma)$ means that for $\Delta > 0$

$$\begin{aligned}
I(X, Y_\gamma) + \frac{d}{d\gamma}I(X;Y_\gamma) &\geq I(X;Y_{\gamma+\Delta}) \\
\implies \frac{d}{d\gamma}I(X;Y_\gamma) &\geq \frac{I(X;Y_{\gamma+\Delta}) - I(X;Y_\gamma)}{\Delta} \\
&= \frac{\rho(1-\rho)}{2}(1-o(1)) \\
&= (1-o(1))\frac{\rho}{2}, \text{ again if } \gamma < 2(1-\epsilon)\log\left(\frac{1}{\rho}\right) \\
\implies \operatorname{mmse}(X \mid Y_\gamma) &= (1-o(1))\rho, \text{ by Corollary 84 again if } \gamma < 2(1-\epsilon)\log\left(\frac{1}{\rho}\right) \\
\implies \operatorname{mmse}(\theta|Y) &= \gamma \operatorname{mmse}(X \mid Y_\gamma) \\
&\geq (1-o(1))\rho\gamma \\
&= (1-o(1))\rho\mu^2, \text{ if } \mu < \sqrt{2(1-\epsilon)\log\left(\frac{1}{\rho}\right)}
\end{aligned}$$

By transitivity, we have the result. □

Theorem 87 (Tensorization of I-MMSE)**.** If $Y_\gamma = X\sqrt{\gamma} + \mathcal{N}(0, I_n)$. Then,

$$\begin{aligned}
\frac{d}{d\gamma}I(X;Y_\gamma) &= \frac{1}{2}\operatorname{mmse}(X|Y_\gamma) \\
&:= \frac{1}{2}\mathbb{E}\left[||X - \mathbb{E}[X|Y_\gamma]||_2^2\right]
\end{aligned}$$

Proof.

Consider the model where $Y_i = X_i\sqrt{\gamma_i} + \mathcal{N}(0,1)$ for possibly different $(\gamma_1, ..., \gamma_n)$. Then,

$$\begin{aligned}
\frac{\partial}{\partial\gamma_i}I(X^n;Y^n) &= \frac{\partial}{\partial\gamma_i}I(X_i;Y^n) + \frac{\partial}{\partial\gamma_i}I(X_n;Y^n \mid X_i) \\
&= \frac{\partial}{\partial\gamma_i}I(X_i;Y^n) + \overset{0}{\cancel{\frac{\partial}{\partial\gamma_i}h(Y^n \mid X_i)}} - \overset{0}{\cancel{\frac{\partial}{\partial\gamma_i}h(Y^n \mid X^n)}} \\
&= \frac{\partial}{\partial\gamma_i}I(X_i;Y^n), \text{ since conditional on } X_i, \gamma_i \text{ adds no randomness to } Y^n \\
&= \frac{1}{2}\operatorname{mmse}(X_i|Y^n), \text{ by Corollary 84}
\end{aligned}$$

Then, we have that

$$\begin{aligned}\frac{d}{d\gamma}I(X;Y_\gamma) &= \sum_{i=1}^{n}\frac{\partial}{\partial\gamma_i}I(X;Y_\gamma)|_{\gamma_i=\gamma}, \text{ by total differentiation}\\ &= \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[(X_i-\mathbb{E}[X_i \mid Y_\gamma])^2\right], \text{ by above}\\ &= \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}(X_i-\mathbb{E}[X_i \mid Y_\gamma])^2\right]\\ &= \frac{1}{2}\mathbb{E}[||X-\mathbb{E}[X \mid Y_\gamma]||_2^2]\\ &= \frac{1}{2}\,\mathrm{mmse}(X \mid Y_\gamma)\end{aligned}$$

□

6 Statistical Decision Theory and Classical Asymptotics

We consider the following framework for statistical decision theory:

- [Statistical Model]: a family of distributions $\{P_\theta : \theta \in \Theta\}$.[5]
- [Observation]: $X \sim P_\theta$ with an unknown $\theta \in \Theta$. We denote the support of X, $\mathrm{supp}(X) =: \mathcal{X}$.
- [Decision Rule / Estimator]: A (possibly random) map $\hat{\theta} : \mathcal{X} \to \mathcal{A}$ with $\mathcal{A}$ called the action space.
- [Loss]: A loss function $\Theta \times \mathcal{A} : \mathbb{R}_+$.
- [Risk / Expected Loss] The risk of an estimator $\hat{\theta}$ under L and some parameter θ^* is $r(\hat{\theta};\theta^*) = \mathbb{E}_{X\sim P_{\theta^*}}[L(\theta^*,\hat{\theta}(X))]$ where we often abbreviate $\mathbb{E}_{X\sim P_{\theta^*}}[\cdot] =: \mathbb{E}_{\theta^*}[\cdot]$.

Example 88 (Density Estimation)**.** Suppose that $X_1, ..., X_n \sim f_\theta$ so that $P_\theta = f^{\otimes n}$. Different losses capture different goals such as

$$\begin{aligned}L_1(f_\theta, a) &= |a - f_\theta(0)|, \text{ density at a point}\\ L_2(f_\theta, a) &= \int_{\mathcal{X}} |f_\theta - a|^2, \text{ global estimation}\\ L_3(f_\theta, a) &= \left|a - \int_{\mathcal{X}} h \circ f_\theta\right|, \text{ functional estimation}\end{aligned}$$

Example 89 (Linear Regression)**.** Suppose that $X_1, ..., X_n$ either fixed or from a random design. Suppose that $\mathbb{E}[Y \mid X] =< \theta, X >$. Different losses capture different goals such as

$$\begin{aligned}L_1(\theta,\hat{\theta}) &= ||\hat{\theta}-\theta||_2^2, \text{ estimation error}\\ L_2(\theta,\hat{\theta}) &= \mathbb{E}_{X\sim P_X}\left[(<\theta, X> - <\hat{\theta}, X>)^2\right], \text{ prediction error}\end{aligned}$$

Example 90 (Learning Theory)**.** Suppose $(X_1, Y_1), ..., (X_n, Y_n) \overset{iid}{\sim} P_{XY}$. For a class of functions $\mathcal{F}$, we might

[5] We say that this model is parametric if $\dim(\Theta) < \infty$, non-parametric if $\dim(\Theta) = \infty$, and semi-parametric if $\Theta = \Theta_1 \times \Theta_2$ with $\dim(\Theta_1) < \infty$ and $\dim(\Theta_2) = \infty$.

wish loss to capture the excess risk with respect to $\mathcal{F}$. That is

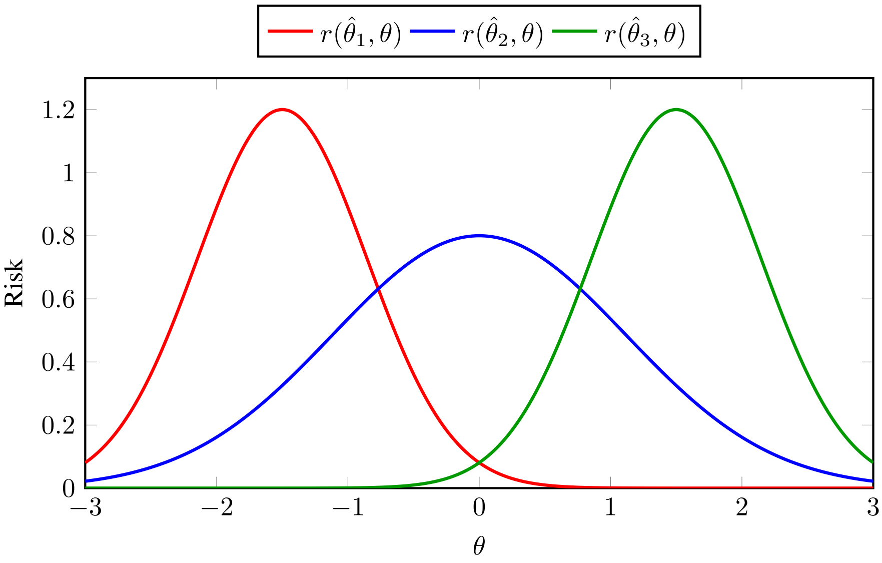
$$L(P_{XY}, \hat{f}) = \mathbb{E}_{P_{XY}}\left[(Y - \hat{f}(X))^2\right] - \inf_{f \in \mathcal{F}} \mathbb{E}_{P_{XY}}\left[(Y - f(X))^2\right]$$

Example 91 (Optimization)**.** Suppose that we're given an unknown function $f : \mathcal{X} \to \mathbb{R}$ that we can query (eg., input x_t and observe y_t from $y_t = f(x_t) + \epsilon_t$). We wish to find the minimum of f. We devise a querying strategy $x_{t+1} = \phi(x^t, y^t)$ for $t \in \{1, ..., T\}$. Our loss is how far off we are from the true minimum at $T + 1$

$$L(f, X_{T+1}) = f(X_{T+1}) - \min_{x \in \mathcal{X}} f(x)$$

6.1 *Comparison of Estimators*

For an estimator $\hat{\theta}$, recall that its risk $r(\hat{\theta}; \theta)$ is a function of θ, not just of the estimator. How can we compare two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$?



- [Option 1]: We say that $\hat{\theta}_2$ is inferior to $\hat{\theta}_1$ if $r(\hat{\theta}_2; \theta) \geq r(\hat{\theta}_1; \theta)$ $\forall \theta \in \Theta$ and $r(\hat{\theta}_2; \theta) > r(\hat{\theta}_1; \theta)$ for some $\theta \in \Theta$.[6] In this case, $\hat{\theta}_2$ is inadmissible. Note that inadmissibility is a weak function since the constant estimator $\hat{\theta} = \theta_0$ is trivially admissible.

- [Option 2]: Given a (prior) probability distribution $\pi(\theta)$ on Θ, look at the weighted average $r_\pi(\hat{\theta}) = \int_\Theta \pi(\theta) r(\hat{\theta}; \theta) d\theta$. The minimizer of $\hat{\theta} \mapsto r_\pi(\theta)$ is called the *Bayes estimator* under π.

- [Option 3]: Look at the *worse-case* risk $r^*(\hat{\theta}) = \max_{\theta \in \Theta} r(\hat{\theta}; \theta)$. The minimizer of $\hat{\theta} \mapsto r^*(\hat{\theta})$ is called the *minimax estimator*.

Definition 92 (Bayes Estimator)**.** For a given prior $\pi(\theta)$ on Θ, Bayes estimator under π is defined as

$$\hat{\theta}_\pi := \arg\min_a \underbrace{\int_\Theta \pi(\theta) r(a; \theta) d\theta}_{=: r_\pi(a)}$$

We denote the minimum by r^*_π.

[6] In practice, this evaluation criterion is not used much since it's quite weak.

Definition 93 (Minimax Estimator)**.** The minimax estimator is defined as

$$\hat{\theta}_{mm} := \arg\min_a \underbrace{\max_{\theta\in\Theta} r(a;\theta)}_{=:r^*(a)}$$

We denote the minimum by r^*.

Theorem 94 (Minimax Theorem)**.** Define the Bayes risk $r^*_\pi := \inf_{\hat{\theta}} r_\pi(\hat{\theta})$. Also define the minimax risk $r^* := \inf_{\hat{\theta}} r^*(\hat{\theta})$. We have that

$$\begin{aligned} r^* &\geq r^*_\pi \;\forall\pi \\ r^* &= \sup_\pi r^*_\pi, \text{ under regularity conditions} \end{aligned}$$

The maximizer of $\sup_\pi r^*_\pi$ is called the *least favorable prior*.

Proof.

[≥]:

We know that for any π and any estimator $\hat{\theta}$, $\sup_\theta r(\hat{\theta};\theta) \geq \mathbb{E}_{\theta\sim\pi}[r(\hat{\theta};\theta)]$ since the maximum is at least the average. Taking the infimum over $\hat{\theta}$ on both sides, that implies that for any π, we have that $r^* \geq r^*_\pi$.

[=]:

Recall that a randomized estimator $\hat{\theta}$ is a probability distribution $p(A \mid X)$ over actions $A \in \mathcal{A}$. As a result, we have that

$$\begin{aligned} \sup_\pi r^*_\pi &= \sup_\pi \inf_p \mathbb{E}_{\theta\sim\pi}[\mathbb{E}_{X\sim P_\theta}[\mathbb{E}_{a\sim p(\cdot|X)}[L(\theta,a)]]] \\ &= \inf_p \sup_\pi \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}[\mathbb{E}_{a\sim p(\cdot|X)}[L(\theta,a)]]\right], \text{ under regularity assumptions} \\ &= \inf_p \sup_\theta \mathbb{E}_{X\sim P_\theta}[\mathbb{E}_{a\sim p(\cdot|X)}[L(\theta,a)]] \\ &= r^* \end{aligned}$$

Finding the Bayes estimator is statistically easy: the prior $\pi(\theta)$ induces a joint distribution $\pi(\theta)P_\theta(X)$ on $\Theta \times \mathcal{X}$, which therefore admits a posterior $\pi(\theta \mid X) \propto \pi(\theta)p_\theta(X)$. Then, the Bayes estimator is the barycenter of $\pi(\theta \mid X)$ under L, ie.,

$$\hat{\theta}_\pi = \arg\min_a \mathbb{E}_{\theta\sim\pi(\cdot|X)}[L(\theta,a)]$$

However, finding the Bayes estimator can be computationally hard. Finding the minimax estimator is often statistically hard and is only feasible in a few examples (see below). One often pivots to express interest in asymptotically minimax estimators or rate-optimal results (ie., find $\hat{\theta}$ such that $r^*(\hat{\theta}) \leq Cr^*$ for some $C > 0$).

Example 95 (Binomial Distribution Minimax Risk)**.** Let $X \sim \text{Bin}(n,\theta)$ and $L(\theta,a) = (\theta-a)^2$. To find the least favorable prior, try $\pi(\theta) \propto \theta^{b-1}(1-\theta)^{b-1}$ (ie., $\theta \sim \text{Beta}(b,b)$). Then, the posterior is $\pi(\theta \mid X)) \propto \pi(\theta)\theta^X(1-$

$\theta)^{n-X} = \theta^{b+X-1}(1-\theta)^{b+n-X-1}$ (ie., $\theta \mid X \sim \text{Beta}(b+X, b+n-X)$). The Bayes estimator is

$$\begin{aligned}\hat{\theta}_\pi(X) &= \arg\min_a \mathbb{E}_{\theta\sim\pi(\cdot|X)}[L(\theta, a(X))] \\ &= \arg\min_a \mathbb{E}_{\theta\sim\pi(\cdot|X)}[(\theta - a(X))^2] \\ &= \mathbb{E}_{\theta\sim\pi(\cdot|X)}[\theta \mid X] \\ &= \frac{X+b}{n+2b}\end{aligned}$$

The risk function is

$$\begin{aligned}r(\hat{\theta}_\pi;\theta) &= \mathbb{E}_{X\sim P_\theta}\left[L(\theta, \hat{\theta}_\pi(X))\right] \\ &= (\text{bias}(\hat{\theta}_\pi(X)))^2 + \text{Var}(\hat{\theta}_\pi(X)) \\ &= \left(\frac{n\theta+b}{n+2b} - \theta\right)^2 + \frac{n\theta(1-\theta)}{(n+2b)^2} \\ &= \frac{1}{(n+2b)^2}\left[b^2 + (n-4b^2)\theta(1-\theta)\right]\end{aligned}$$

By choosing $b = \frac{\sqrt{n}}{2}$, we have that $r(\hat{\theta}_\pi;\theta) = \frac{1}{4(\sqrt{n}+1)^2}$. Therefore, $\hat{\theta}_\pi(X) = \frac{X+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}$ attains the worst case risk $r^*(\hat{\theta}_\pi) = \frac{1}{4(\sqrt{n}+1)^2}$ and

$$\begin{aligned}r^* &\le r^*(\hat{\theta}_\pi), \text{ maximizer of minimax can guarantee this worst case loss} \\ &= r_\pi(\hat{\theta}_\pi) \\ &\le r^*, \text{ minimizer can select } \theta \text{ for which } \hat{\theta}_\pi \text{ performs worse} \\ \implies r^* &= \frac{1}{4(\sqrt{n}+1)^2}\end{aligned}$$

□

Definition 96 (Bowl-Shaped Function). *A function $\rho : \mathbb{R}^n \to \mathbb{R}_+$ is bowl-shaped if $\rho(x) = \rho(-x)$ and ρ is quasi-convex meaning that $\{x : \rho(x) \le a\}$ is convex for any $a \in \mathbb{R}$.*

Lemma 97 (Anderson's Lemma). If $X \sim \mathcal{N}(0, \Sigma)$ and ρ is bowl-shaped, then $\min_{a\in\mathbb{R}^n} \mathbb{E}_X[\rho(X+a)] = \mathbb{E}_X[\rho(X)]$.

Proof.

Let $K_c := \{x : \rho(x) \le c\}$. Since ρ is bowl-shaped, we know that K_c is convex and $K_c = -K_c$. Then,

$$\begin{aligned}\mathbb{E}_X[\rho(X+a)] &= \int_0^\infty \Pr(\rho(X+a) > c)dc, \text{ since } \rho \text{ is bowl-shaped} \\ &= \int_0^\infty (1 - \Pr(X+a \in K_c))dc \\ &\overset{(*)}{\ge} \int_0^\infty (1 - \Pr(X \in K_c))dc \\ &= \int_0^\infty \Pr(X > c)dc \\ &= \mathbb{E}_X[\rho(X)]\end{aligned}$$

To see step $(*)$, note that[a]

$$\begin{aligned}
\Pr(X \in K_c) &= \Pr\left(X \in \frac{1}{2}(K_c + a) + \frac{1}{2}(K_c - a)\right), \text{ since } K_c = \frac{K_c}{2} + \frac{K_c}{2} \text{ by convexity} \\
&\geq \sqrt{\Pr(X \in K_c + a)\Pr(X \in K_c - a)}, \text{ since } X \text{ has log-concave distribution} \\
&= \sqrt{\Pr(X \in K_c + a)\Pr(X \in -K_c - a)}, \text{ since } K_c = -K_c \\
&= \Pr(X \in K_c + a), \text{ distribution of } X \text{ is symmetric around } 0
\end{aligned}$$

□

[a] A function f is log-concave if $f(\lambda x + (1-\lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}$ for any $x, y \in \text{domain}(f)$ and $\lambda \in [0,1]$.

Example 98 (GLM). Let $X \sim \mathcal{N}(\theta, I_n)$ and define $L(\theta, a) = \rho(\theta - a)$ where $\rho : \mathbb{R}^n \to \mathbb{R}_+$ is a continuous bowl-shaped loss function. We claim that $\hat{\theta} = X$ is the minimax estimator with risk $r^* = \mathbb{E}[\rho(Z)]$ where $Z \overset{indep}{\sim} \mathcal{N}(0, I_n)$.

Proof.

$[r^* \geq \mathbb{E}_Z[\rho(Z)]]$:

Let's try the prior $\pi = \mathcal{N}(0, \tau^2 I_n)$, then, one can show that the posterior $\theta \mid X \sim \mathcal{N}\left(\frac{\tau^2 X}{1+\tau^2}, \frac{\tau^2}{1+\tau^2} I_n\right)$. Thus, we have that

$$\begin{aligned}
r^* &\geq r^*_\pi \\
&= \mathbb{E}_X\left[\min_{a \in \mathbb{R}^n} \mathbb{E}_{\theta|X}\left[\rho(\theta - a)\right]\right] \\
&= \mathbb{E}_Z\left[\rho\left(Z\sqrt{\frac{\tau^2}{1+\tau^2}}\right)\right], \text{ by Lemma 97 (ie., Anderson's Lemma)}
\end{aligned}$$

Letting $\tau \to \infty$ gives that $r^* \geq \mathbb{E}[\rho(Z)]$ by continuity of ρ.

$[r^* \leq \mathbb{E}_Z[\rho(Z)]]$:

Take $\hat{\theta}(X) = X$. Then,

$$\begin{aligned}
r(\hat{\theta}; \theta) &= \mathbb{E}_X\left[\rho(\theta - X)\right] \\
&= \mathbb{E}_Z\left[\rho(Z)\right]
\end{aligned}$$

□

6.2 Classical Asymptotics

We consider $X_1, ..., X_n \overset{iid}{\sim} P_\theta$ with $n \to \infty$ to enter the Hajek-Le Cam classical asymptotic regime. We let $\mathcal{X} := \text{supp}(X_i)$.

Definition 99 (Differentiable in Quadratic Mean (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

$$\int_{\mathcal{X}} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2}h' s_\theta(x)\sqrt{p_\theta(x)}\right]^2 \mu(dx) = p\left(||h||^2\right)$$

where μ is any dominating measure for $(P_\theta)_{\theta\in\Theta}$ and $p_\theta = \frac{dP_\theta}{d\mu}$ (ie., p_θ is the density of P_θ under μ).

We make the following two observations about Definition 99:

First, assume that $h \mapsto \sqrt{p_{\theta+h}(x)}$ is differentiable almost everywhere. Then, we have that

$$\begin{aligned}
\sqrt{p_{\theta+h}(x)} &= \sqrt{p_\theta(x)} + h'\frac{\partial}{\partial\theta}\sqrt{p_\theta(x)} + o(||h||) \\
\implies \int_{\mathcal{X}} &\left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2}h's_\theta(x)\sqrt{p_\theta(x)}\right]^2 \mu(dx) \\
&= \int_{\mathcal{X}} \left[h'\left(\frac{\partial}{\partial\theta}\sqrt{p_\theta(x)} - \frac{1}{2}s_\theta(x)\sqrt{p_\theta(x)}\right) + o(||h||)\right]^2 \mu(dx)
\end{aligned}$$

For this expression to equal $o(||h||^2)$, we need that $\frac{\partial}{\partial\theta}\sqrt{p_\theta(x)} - \frac{1}{2}s_\theta(x)\sqrt{p_\theta(x)} = 0\ a.e.$ which means that

$$\begin{aligned}
s_\theta(x) &= \frac{2}{\sqrt{p_\theta(x)}}\frac{\partial}{\partial\theta}\sqrt{p_\theta(x)} \\
&= \frac{\frac{\partial}{\partial\theta}p_\theta(x)}{p_\theta(x)} \\
&= \frac{\partial}{\partial\theta}\log\left(p_\theta(x)\right)
\end{aligned}$$

Second, define $r_h(x) := \sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2}h's_\theta(x)\sqrt{p_\theta(x)}$. Then,

$$\begin{aligned}
2 &\geq H^2(P_{\theta+h}, P_\theta) \\
&= \int_{\mathcal{X}} \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)}\right]^2 \mu(dx) \\
&= \int_{\mathcal{X}} \left[\frac{1}{2}h's_\theta(x)\sqrt{p_\theta(x)} + r_h(x)\right]^2 d\mu(x), \text{ with QMD} \\
&= \frac{1}{4}\int_{\mathcal{X}} \left(h's_\theta(x)\sqrt{p_\theta(x)}\right)^2 d\mu(x) + \int_{\mathcal{X}} h's_\theta(x)\sqrt{p_\theta(x)}r_h(x)\mu(dx) + \int_{\mathcal{X}} (r_h(x))^2\mu(dx) \\
&= \frac{1}{4}h' \underbrace{\int_{\mathcal{X}} s_\theta(x)s_\theta(x)'\mu(dx)}_{=:I(\theta)} h + o(||h||^2), \text{ by Cauchy-Schwarz} \\
&= \frac{1}{4}h'I(\theta)h + o(||h||^2)
\end{aligned}$$

by transitivity and the fact that h can go in any direction, we have that $I(\theta)$ exists.

6.2.1 History of Asymptotic Theorems

Consider a situation where $X_1, ..., X_N \overset{iid}{\sim} P_\theta$. We assume that $P_\theta \in \{P_{\theta'} : \theta' \in \Theta\}$ and that for some dominating measure μ, any distribution in this model parametrized by θ has density f_θ. We define the MLE estimator for θ from $\mathcal{D} := \{X_1, ..., X_N\}$ by

$$\hat{\theta}_N^{MLE} := \arg\max_{\theta'} \Pi_{i=1}^n f_{\theta'}(X_i)$$

Fisher claimed the following program:

(1) For any true $\theta \in \Theta$, the MLE $\hat{\theta}_N^{MLE}$ satisfies $\sqrt{N}(\hat{\theta}_N^{MLE} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$ where $I(\theta)$ is the Fisher information matrix of P_θ.

(2) For any true $\theta \in \Theta$, and for any other sequence of estimators $\{T_n\}$ with $\sqrt{N}(T_N - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma_\theta)$, then $\Sigma_\theta \succeq I(\theta)^{-1}$.

While (1) is true under mild regularity conditions, (2) is unfortunately not true as observed by Hodges (1951).

Example 100 (Hodges (1951))**.** Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\theta, 1)$. Construct

$$\hat{\theta}_N = \begin{cases} \frac{1}{N}\sum_{i=1}^N X_i & \text{if } \frac{1}{N}\sum_{i=1}^N X_i \geq N^{-1/4} \\ 0 & \text{if } \frac{1}{N}\sum_{i=1}^N X_i < N^{-1/4} \end{cases}$$

It's easy to see that

$$\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} \begin{cases} \mathcal{N}(0,1) & \textit{if } \theta \neq 0 \\ 0 & \textit{if } \theta = 0 \end{cases}$$

so that (2) does not hold in Fisher's program when $\theta = 0$.

Example 100 (ie., Hodges (1951)) shows that cautions need to be taken when defining the "optimality" of the MLE or the inverse Fisher information matrix. It then took statisticians 20 years to find the right definitions, through the following angles:

- Hodges' estimator is not "regular" (restricting the class of estimators).
- The set of violations has Lebesgue measure 0 ("superefficiency" occurs rarely).
- The performance of Hodges' estimator is bad when $\theta \approx N^{-1/4}$ – it has a large asymptotic local risk.

6.3 A Collection of Asymptotic Theorems

Here we present a few key asymptotic theorems and some examples. Unless we state otherwise we assume that we're operating in a finite dimensional parameter space.

Theorem 101 (Convolution Theorem)**.** Let $(P_\theta)_{\theta \in \Theta}$ be QMD at every $\theta \in \Theta$. If $\sqrt{N}(T_N - \psi(\theta)) \xrightarrow{d} L_\theta$ under $P_\theta^{\otimes N}$ $\forall \theta \in \Theta$ (with $\dim(\Theta) = d$) and $\{T_n\}$ is *regular* in the sense that

$$\sqrt{N}\left(T_N - \psi\left(\theta + \frac{h}{\sqrt{N}}\right)\right) \xrightarrow{d} L_\theta$$

under $P_{\theta+\frac{h}{\sqrt{N}}}^{\otimes N}$ $\forall h \in \mathbb{R}^d$ and $\forall \theta \in \Theta$. Then, for any $\theta \in \Theta$, there exists a probability measure $\mathcal{M}_\theta$ such that

$$L_\theta = \mathcal{N}(0, (D_\theta\psi(\theta))' I(\theta)^{-1} D_\theta\psi(\theta)) * \mathcal{M}_\theta$$

where $*$ is the convolution operator.[a]

[a] Convolution makes a distribution more "noisy" (by spreading out probability mass) so that the limiting distribution of our estimator can be written as a convolution between the efficient Gaussian part and some measure.

Theorem 102 (Almost Everywhere Convolution Theorem)**.** Under all of the conditions and definitions of Theorem 101 except for the regularity of $\{T_n\}$, then

$$L_\theta = \mathcal{N}(0, (D_\theta\psi(\theta))' I(\theta)^{-1} D_\theta\psi(\theta)) * \mathcal{M}_\theta$$

for Lebesgue almost every θ.

Theorem 103 (Local Asymptotic Minimax Theorem (LAM))**.** For every continuous and bowl-shaped loss ρ and any sequence of estimators $\{T_n\}$,

$$\lim_{c\to\infty} \liminf_{N\to\infty} \sup_{||h||\le c} \mathbb{E}_{\theta+\frac{h}{\sqrt{N}}} \left[\rho\left(\sqrt{N}(T_N) - \psi\left(\theta + \frac{h}{\sqrt{N}}\right)\right)\right] \ge \mathbb{E}_Z[\rho(Z)]$$

where $Z \overset{indep}{\sim} \mathcal{N}(0, (D_\theta\psi(\theta))' I(\theta)^{-1} D_\theta\psi(\theta))$.[a]

[a] This theorem provides a lower bound on the minimax (asymptotic) risk of the local family $\left(P_{\theta+\frac{h}{\sqrt{N}}}\right)_{||h||\le c}$ under the loss $L(\theta, a) = \rho\left(\sqrt{N}\left(a - \psi(\theta)\right)\right)$.

The general proofs of these theorems rely on the asymptotic equivalence between the models $\left(P_{\theta+\frac{h}{\sqrt{N}}}\right)_{||h||\le c}$ and the GLM $\left(\mathcal{N}(h, I(\theta)^{-1})\right)_{||h||\le c}$, which is out of scope of these notes (but can be found here).

6.4 Proving some Special Cases of LAM

Here, we will prove some more accessible versions of Theorem 103 (ie., LAM).

Theorem 104 (Bayesian Cramer-Rao (BCR) in 1D (van-Trees Inequality))**.** Let $\theta \in [a, b]$ and $\pi(\cdot)$ be a differentiable prior on $[a.b]$ with $\pi(a) = \pi(b) = 0$ and suppose that $J(\pi) = \int_a^b \frac{\pi'(\theta)^2}{\pi(\theta)} d\theta < \infty$ ($J(\cdot)$ defined in Definition 70). Also suppose $X \sim P_\theta$ with support $\mathcal{X}$ and density p_θ with respect to some dominating measure μ. Then, for any estimator $\hat{\theta}$,[a]

$$\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[(\hat{\theta}(X) - \theta)^2\right]\right] \ge \frac{1}{\mathbb{E}_{\theta\sim\pi}[I(\theta)] + J(\pi)}$$

Proof.

First, we see that

$$\begin{aligned} \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(\hat{\theta}(X) - \theta\right) D_\theta\left(\log(\pi(\theta)p_\theta(X))\right)\right]\right] &= \int_{\mathcal{X}}\int_a^b (\hat{\theta}(X) - \theta) D_\theta(\log(\pi(\theta)p_\theta(X))) d\theta\mu(dx) \\ &= \int_{\mathcal{X}}\int_a^b \pi(\theta)p_\theta(X) d\theta\mu(dx), \text{ integration by parts} \\ &= 1 \end{aligned}$$

Next, we see that

$$\begin{aligned}\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[(D_\theta(\log(\pi(\theta)p_\theta(X))))^2\right]\right] &= \mathbb{E}_{\theta\sim\pi}\left[\left(\frac{\pi'(\theta)}{\pi(\theta)}\right)^2\right] + \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(\frac{D_\theta p_\theta(X)}{p_\theta(X)}\right)^2\right]\right] \\ &\quad + 2\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\frac{\pi'(\theta)}{\pi(\theta)}\cdot\frac{D_\theta p_\theta(X)}{p_\theta(X)}\right]\right] \\ &= J(\pi) + \mathbb{E}_{\theta\sim\pi}\left[I(\theta)^{-1}\right]\end{aligned}$$

The right most term in the second to last line equals 0 under some regularity conditions that implies

$$\begin{aligned}\int_{\mathcal{X}} D_\theta p_\theta(X)\mu(dx) &= D_\theta\int_{\mathcal{X}} p_\theta(X)\mu(dx) \\ &= 0\end{aligned}$$

Finally, by Cauchy-Schwarz, we see that

$$\begin{aligned}&\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(\hat{\theta}(X)-\theta\right)D_\theta\left(\log(\pi(\theta)p_\theta(X))\right)\right]\right] \\ &\leq \sqrt{\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(\hat{\theta}(X)-\theta\right)^2\right]\right]\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(D_\theta\left(\log(\pi(\theta)p_\theta(X))\right)\right)^2\right]\right]} \\ \implies\ & 1 \leq \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(\hat{\theta}(X)-\theta\right)^2\right]\right]\left(\mathbb{E}_{\theta\sim\pi}\left[I(\theta)^{-1}\right]+J(\pi)\right)\end{aligned}$$

That implies that

$$\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(\hat{\theta}(X)-\theta\right)^2\right]\right] \geq \frac{1}{\mathbb{E}_{\theta\sim\pi}\left[I(\theta)^{-1}\right]+J(\pi)}$$

□

[a] This statement can be compared to the usual Cramer-Rao lower bound $\mathbb{E}_{X\sim P_\theta}\left[(\hat{\theta}-\theta)^2\right] \geq \frac{1}{I(\theta)}$ for unbiased $\hat{\theta}$. The lower bound is adjusted slightly to allow for unbiased estimators.

Theorem 105 (Multivariate Bayesian Cramer-Rao (BCR))**.** Let $\pi = \Pi_{i=1}^d \pi_i$ be a differentiable prior on $\Pi_{i=1}^d[a_i, b_i]$ vanishing on the boundary and $J(\pi) = diag(J(\pi_1), ..., J(\pi_d))$. Also suppose $X \sim P_\theta$ with support $\mathcal{X}$ and density p_θ with respect to some dominating measure μ. Then, for any estimator $\hat{\theta}$ we have that

$$\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left|\left|\hat{\theta}-\theta\right|\right|^2\right]\right] \geq \text{Tr}\left[\left((\mathbb{E}_{\theta\sim\pi}[I(\theta)]+J(\pi)\right)^2\right]$$

Proof.

Similar to the proof of Theorem 104 (ie., 1D Bayesian Cramer-Rao), we can show that $\forall k \in \{1, ..., d\}$, we have that

$$\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[(\hat{\theta}_k-\theta_k)D_\theta\left(\log(\pi(\theta)p_\theta(X))\right)\right]\right] = e_k$$

where e_k is the k-th basis vector. That implies that

$$\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[(\hat{\theta}-\theta)\left(D_\theta\left(\log(\pi(\theta)p_\theta(X))\right)\right)'\right]\right] = I_d \tag{5}$$

Next, let

$$\begin{aligned}\Sigma &:= \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_\theta}\left[\left(D_\theta(\log\left(\pi(\theta)p_\theta(X)\right)\right)\left(D_\theta(\log\left(\pi(\theta)p_\theta(X)\right)\right)'\right]\right]\\ &= \mathbb{E}_{\theta\sim\pi}[I(\theta)] + J(\pi)\end{aligned}$$

again like in the proof of Theorem 104 (ie., 1D Bayesian Cramer-Rao). Then, by Cauchy-Schwarz, we have that for any $u, v \in \mathbb{R}^d$

$$\begin{aligned}u'\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_X\left[(\hat{\theta}-\theta)\left(D_\theta\left(\log(\pi(\theta)p_\theta(X))\right)\right)'\right]\right]v &= \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_X\left[((\hat{\theta}-\theta)'u)\left(D_\theta\left(\log(\pi(\theta)p_\theta(X))\right)\right)'v\right]\right]\\ &\leq \sqrt{\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_X\left[(\hat{\theta}-\theta)'uu'(\hat{\theta}-\theta)\right]\right]v'\Sigma v}\end{aligned}$$

If we take $u = e_k, v = \Sigma^{-1}u$ and use the identity of Equation (5), we get that

$$\begin{aligned}(u'I_dv)^2 &\leq \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_X\left[(\hat{\theta}_k-\theta_k)^2\right]\right]v'\Sigma v\\ \implies \frac{\left(e_k\Sigma^{-1}e_k\right)^2}{(\Sigma^{-1}e_k)'\Sigma(\Sigma^{-1}e_k)} &\leq \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_X\left[(\hat{\theta}_k-\theta_k)^2\right]\right]\\ \implies e_k\Sigma^{-1}e_k &\leq \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_X\left[(\hat{\theta}_k-\theta_k)^2\right]\right]\\ \implies \left(\Sigma^{-1}\right)_{kk} &\leq \mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_X\left[(\hat{\theta}_k-\theta_k)^2\right]\right]\end{aligned}$$

Summing over $k \in \{1, ..., d\}$ gives the result. □

Application 106 (Deriving LAM from BCR when $\psi(\theta) = \theta, \rho(x) = ||x||^2$)**.** Initially, let $\theta \in [a, b]$. If,

$$\pi(\theta) = \frac{2}{b-a}\cos^2\left(\frac{\pi}{2}\cdot\frac{2\theta-(a+b)}{b-a}\right)$$

then we have that $\pi(a) = \pi(b) = 0$, $\pi(\cdot)$ is differentiable, and[a]

$$\begin{aligned}J(\pi) &= \int_a^b \frac{8\pi^2}{(b-a)^3}\sin^2\left(\frac{\pi}{2}\cdot\frac{2\theta-(a+b)}{b-a}\right)d\theta\\ &= \frac{4\pi^2}{(b-a)^2}\\ &< \infty\end{aligned}$$

Now, fix any $\theta \in \Theta \subseteq \mathbb{R}^d$. Also suppose $X \sim P_\theta$ with support $\mathcal{X}$ and density p_θ with respect to some dominating measure μ. For each $i \in \{1, ..., d\}$, choose π_i as defined earlier in this application on $\left[\theta_i - \frac{c}{\sqrt{n}}, \theta_i + \frac{c}{\sqrt{n}}\right]$. Theorem 105 (ie., Multivariate BCR) gives

$$\begin{aligned}\inf_{\hat{\theta}}\sup_{||h||_\infty\leq c}\mathbb{E}_{X\sim P_{\theta+\frac{h}{\sqrt{n}}}}\left[\left|\left|\hat{\theta}-\left(\theta+\frac{h}{\sqrt{n}}\right)\right|\right|^2\right] &\geq \inf_{\hat{\theta}}\mathbb{E}_{\theta\sim\pi}\left[\mathbb{E}_{X\sim P_{\theta+\frac{h}{\sqrt{n}}}}\left[\left|\left|\hat{\theta}-\left(\theta+\frac{h}{\sqrt{n}}\right)\right|\right|^2\right]\right]\\ &\geq \operatorname{Tr}\left(\left(n\mathbb{E}_{\theta\sim\pi}\left[I(\theta)\right]+\frac{n\pi^2}{c^2}I_d\right)^{-1}\right)\\ &= \frac{1-o(1)}{n}\operatorname{Tr}\left(I(\theta)^{-1}\right),\ \text{as } n\to\infty \text{ and } c\to\infty\end{aligned}$$

The first inequality is a consequence of Theorem 94 (ie., Minimax Theorem). The second inequality in this last sequence uses (a) Theorem 105, (b) the derivation of $J(\pi)$ above, and (c) the fact that for $X^n \sim P_\theta^{\otimes n}$, the Fisher Information of $P_\theta^{\otimes n}$ is $nI(\theta)$ where $I(\theta)$ is the Fisher Information of P_θ. The final equality assumes that $\theta \mapsto I(\theta)$ is continuous at θ and $c = o(\sqrt{n})$ so that the average Fisher information collapses onto the Fisher Information at θ.

[a] One can show that this choice of π minimizes the value of $J(\pi)$.

Since the global minimax risk is always lower bounded by the local minimax risk, the LAM gives us *asymptotic* lower bounds on the minimax risk.

Example 107 (Revisiting Binomial Distribution Minimax Risk)**.** Recall the setting of Example 95. We have that

$$\begin{aligned} r_n^* &= \inf_{\hat{\theta}} \sup_{\theta \in [0,1]} \mathbb{E}_{X \sim Bin(\theta, n)} \left[\left(\hat{\theta}(X) - \theta \right)^2 \right] \\ &\geq \inf_{\hat{\theta}} \sup_{||h|| \leq c} \mathbb{E}_{X \sim Bin\left(\frac{h}{\sqrt{n}} + \frac{1}{2}, n\right)} \left[\left(\hat{\theta} - \left(\frac{1}{2} + \frac{h}{\sqrt{n}} \right) \right)^2 \right], \text{ for some small } c \\ &\geq \frac{1 - o(1)}{nI(1/2)}, \text{ by Application 106} \\ &= \frac{1 - o(1)}{4n} \end{aligned}$$

This expression is consistent with the exact expression for $r_n^* = \frac{1}{4(\sqrt{n}+1)^2}$ computed in Example 95.[a]

[a] There we labeled the quantity r^* as we didn't have any sort of asymptotics occurring anywhere in the derivation.

Example 108 (Non-Parametric Entropy Estimation)**.** Let $X_1, ..., X_n \overset{iid}{\sim} f$, which is a density on $[0, 1]$. The goal is to estimate the differential entropy $h(f) = \int_0^1 -f(x) \log(f(x)) dx$ under the squared loss. The challenge is that this is *not* a finite-dimensional model so that LAM doesn't apply directly. To apply the LAM, we consider a one-parameter sub-family $(f_0 + tg)_{|t| \leq \epsilon}$. Then, we have that

$$\begin{aligned} I(0) &= \int_0^1 \frac{g(x)^2}{f_0(x)} dx \\ \frac{d}{dt} h(f_0 + tg)\Big|_{t=0} &= -\int_0^1 (1 + \log(f_0(x))) g(x) dx \end{aligned}$$

Then, we can apply Application 106 with $n \to \infty$ and the Delta method to get that

$$\begin{aligned} r_n^* &\geq \frac{1 - o(1)}{n} \left(\int_0^1 \frac{g(x)^2}{f_0(x)} dx \right)^{-1} \left(\int_0^1 (1 + \log(f_0(x))) g(x) dx \right)^2 \\ &=: \frac{1 - o(1)}{n} V(f_0, g) \end{aligned}$$

We can maximize this lower bound with respect to g. Since $\int_0^1 g(x) dx = 0$ (so that $f_0 + tg$ is a density $\forall |t| \leq \epsilon$), Cauchy-Schwarz[a] gives us that

$$\begin{aligned} V(f_0, g) &= \left(\int_0^1 \frac{g(x)^2}{g(x)} dx \right)^{-1} \left(\int_0^1 (\log(f_0(x) + h(f_0))\, g(x) dx \right)^2, \text{ since } \int_0^1 g(x) dx = 0 \\ &\leq \int_0^1 f_0(x) \left(\log(f_0(x)) + h(f_0) \right)^2 dx \\ &= \int_0^1 f_0(x) \log^2 \left(f_0(x) \right) dx - h(f_0)^2 \end{aligned}$$

where equality holds when $g(x) = f_0(x)\left(\log(f_0(x)) + h(f_0)\right)$. Therefore, we have that

$$r_n^* \geq \frac{1 - o(1)}{n} \sup_{f_0} \left(\int_0^1 f_0(x) \log^2\left(f_0(x)\right) dx - h(f_0)^2 \right)$$

[a] The application of Cauchy-Schwarz is $\left(\int_0^1 \left(\log(f_0(x)) + h(f_0)\right) g(x) dx\right)^2 \leq \left(\int_0^1 f_0(x)\left(\log(f_0(x)) + h(f_0)\right)^2 dx\right)\left(\int_0^1 \frac{g(x)^2}{f_0(x)} dx\right)$, which then we can rearrange into $V(f_0, g) \leq \int_0^1 f_0(x)\left(\log(f_0(x)) + h(f_0)\right) dx$.

There are some pros and cons to the asymptotic theorems:

- [Pro 1]: We can plug-and-play for essentially all statistical models.
- [Pro 2]: We get an exact constant for the asymptotic risk.
- [Con 1]: Bounds are asymptotic, assuming $n \to \infty$ and $d = \dim(\theta)$ is fixed.
- [Con 2]: The bounds are for asymptotic variance and when using these values for non-asymptotic cases with a high-dimensional parameter vector, high-dimensional scenarios *bias* can be the dominating factor.

This motivates study of non-asymptotic lower-bounds for minimax risk in the next few lectures.

7 Minimax lower bounds: Le Cam, Assouad, and Fano

This section will cover non-asymptotic strategies to show a rate-optimal lower-bound on the performance of an estimator.

7.1 Leveraging Two Hypotheses

First, we focus on a technique (ie., Le Cam's Two Point Method) to come up with asymptotic lower bounds on the performance of an estimator by looking at just two hypotheses.

Theorem 109 (Le Cam's Two Point Method)**.** Consider a statistical model $\{P_\theta : \theta \in \Theta\}$ and we wish to construct an estimator $\hat{\theta}(X)$ for θ from an observation $X \sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta, a)$.[a] Suppose that $\theta_0, \theta_1 \in \Theta$ satisfy the separation condition

$$\inf_a \left(L(\theta_0, a) + L(\theta_1, a)\right) \geq \Delta$$

for some $\Delta > 0$,[b] then

$$r^* \geq \inf_{\hat{\theta}} \frac{1}{2} \left[\mathbb{E}_{X \sim P_{\theta_0}}[L(\theta_0, \hat{\theta}(X))] + \mathbb{E}_{X \sim P_{\theta_1}}[L(\theta_1, \hat{\theta}(X))]\right] \geq \frac{\Delta}{2}\left(1 - TV(P_{\theta_0}, P_{\theta_1})\right)$$

where r^* is defined in Definition 93 and the TV distance is defined in Section 3.1.

Proof.

The first inequality is obvious by Theorem 94 so that we focus on the second. For any estimator $\hat{\theta} := \hat{\theta}(X)$, we have

that

$$\begin{aligned}\mathbb{E}_{X\sim P_{\theta_0}}[L(\theta_0,\hat{\theta})]+\mathbb{E}_{X\sim P_{\theta_1}}[L(\theta_1,\hat{\theta})] &= \int_{x\in\mathcal{X}} L(\theta_0,\hat{\theta}(x))p_{\theta_0}(x)dx+\int_{x\in\mathcal{X}} L(\theta_1,\hat{\theta}(x))p_{\theta_1}(x)dx\\ &\geq \Delta\int_{x\in\mathcal{X}}\min(p_{\theta_0}(x),p_{\theta_1}(x))dx, \text{ by the separation condition}\\ &= \Delta\int_{x\in\mathcal{X}}\frac{1}{2}[p_{\theta_0}(x)+p_{\theta_1}(x)-|p_{\theta_0}(x)-p_{\theta_1}(x)|]dx\\ &= \Delta(1-TV(P_{\theta_0},P_{\theta_1})), \text{ by definition of TV distance}\\ &\geq \frac{\Delta}{2}(1-TV(P_{\theta_0},P_{\theta_1}))\end{aligned}$$

□

[a] Assume that the models admit a density with respect to a measure on $\mathcal{X}$ though the measure-theoretic details are skirted here.

[b] Intuitively, the separation condition says that no single action can do *very well* when the truth is θ_0 and when the truth is θ_1.

The general paradigm for applying Theorem 109 is that one finds two points $\theta_0,\theta_1\in\Theta$ satisfying

- The separation condition: $\inf_a(L(\theta_0,a)+L(\theta_1,a))\geq\Delta$.
- The indistinguishability condition: $TV(P_{\theta_0},P_{\theta_1})\leq 1-\Omega(1)$

and then one has a lower bound on the minimax risk r^*. The second condition can be implied by some of the inequalities in Section 3.8. Despite the simplicity of the two-point method, it has numerous applications.

Example 110 (Normal Mean Estimation)**.** Suppose that $X\sim\mathcal{N}(\theta,\sigma^2)$ for unknown $\theta\in\mathbb{R}$ and known $\sigma^2>0$. The target is establishing bounds on $r^*:=\inf_{\hat{\theta}}\sup_\theta\mathbb{E}_{X\sim\mathcal{N}(\theta,\sigma^2)}[(\hat{\theta}(X)-\theta)^2]$ so that $L(\theta,a)=(\theta-a)^2$. Clearly, by choosing $\hat{\theta}(X)=X$, we achieve $r^*\leq\sigma^2$.

To lower bound r^*, apply the two-point method with $\theta_0=0$ and $\theta_1=\delta$ for any $\delta\in\mathbb{R}\setminus\{0\}$. We have that the separation condition of Theorem 109 holds with $\Delta:=\frac{\delta^2}{2}$, which can be seen by minimizing the polynomial $f(a):=L(\theta_0,a)+L(\theta_1,a)$. Next, we can show that the indistinguishability condition holds with $TV(\mathcal{N}(0,\sigma^2),\mathcal{N}(\delta,\sigma^2))=2\Phi\left(\frac{|\delta|}{2\sigma}\right)-1$, where $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0,1)$.

Therefore, applying Theorem 109 (ie., Le Cam's Two Point Method), we have that

$$\begin{aligned}r^* &\geq \sup_{\delta\in\mathbb{R}\setminus\{0\}}\frac{\delta^2}{2}\left(1-\Phi\left(\frac{|\delta|}{2\sigma}\right)\right)\\ &\approx 0.332\sigma^2\end{aligned}$$

This is a fairly weak lower bound as $r^*=\sigma^2$ by a trivial application of 97.

Example 111 (Binomial Model)**.** Let $X\sim Bin(n,\theta)$ with unknown $\theta\in[0,1]$. The target is computing $r^*:=\inf_{\hat{\theta}}\sup_\theta\mathbb{E}_{X\sim\mathcal{N}(\theta,\sigma^2)}[(\hat{\theta}(X)-\theta)^2]$ so that $L(\theta,a)=(\theta-a)^2$.

For an upper bound on r^*, one can choose $\hat{\theta}(X)=\frac{X}{n}$ so that $\mathbb{E}_{X\sim Bin(\theta,n)}\left[\left(\frac{X}{n}-\theta\right)^2\right]=\frac{\theta(1-\theta)}{n}\leq\frac{1}{4n}=O\left(\frac{1}{n}\right)$.

For a lower bound on r^*, one can apply the two point method with $\theta_0=\frac{1}{2}$ and $\theta_1=\frac{1}{2}+\frac{1}{2\sqrt{n}}$. As for the separation condition, one can see that it holds with $\Delta=\frac{1}{2}\left(\frac{1}{2\sqrt{n}}\right)^2=\Omega\left(\frac{1}{n}\right)$. As for the indistinguishability condition, one has that

$$\begin{aligned}D_{KL}(Bin(n,\theta_1)||Bin(n,\theta_0)) &= nD_{KL}(Bern(\theta_1)||Bern(\theta_0)), \text{ by Property 19}\\ &= \frac{n}{2}\left[\left(1+\frac{1}{\sqrt{n}}\right)\log\left(1+\frac{1}{\sqrt{n}}\right)+\left(1-\frac{1}{\sqrt{n}}\right)\log\left(1-\frac{1}{\sqrt{n}}\right)\right]\\ &= \frac{n}{2}O\left(\frac{1}{n}\right)\\ &= O(1)\end{aligned}$$

By Section 3.8, since we know that $TV(Bin(n,\theta_1), Bin(n,\theta_0)) \leq 1 - \frac{1}{2}\exp(-D_{KL}(Bin(n,\theta_1)||Bin(n,\theta_0)))$, then we have that

$$TV(Bin(n,\theta_1), Bin(n,\theta_0)) \leq 1 - \Omega(1)$$

Then, using the result of Theorem 109 (ie., Le Cam's Two Point Method), we have that

$$\begin{aligned}r^* &\geq \frac{\Delta}{2}\left(1 - TV(P_{\theta_0}, P_{\theta_1})\right)\\ &\geq \Omega\left(\frac{1}{n}\right)(1-(1-\Omega(1)))\\ &= \Omega\left(\frac{1}{n}\right)\end{aligned}$$

Intuitively, this result makes sense: we can't have an estimator that has loss that decays at faster than a $\frac{1}{n}$ rate, which is the rate at which the variance of the standard sample fraction estimator decays.

Example 112 (Functional Estimation). Let $X_i \overset{iid}{\sim} P$ for $i \in [N]$ where $P \in \Delta^{K-1}$ for $K < \infty$ so that P is a probability mass function. For simplicity, we define $X = (X_1, ..., X_n)$. We define the loss $L(P, a) := |a - H(p)|$ so that we seek to estimate the entropy of P. Jia et al. (2015) and Win and Yang (2016) show that[a]

$$\begin{aligned}r^* &= \inf_{\hat{\theta}}\sup_{P}\mathbb{E}_{X\sim P^{\otimes n}}[|\hat{\theta}(X) - H(P)|]\\ &\asymp \frac{k}{\log(n)} + \frac{\log(k)}{\sqrt{n}} \text{ if } k \leq n\log(n)\end{aligned}$$

Proof.

I focus here on the proof that $r^* = \Omega\left(\frac{\log(k)}{\sqrt{n}}\right)$. The other lower bound is proven using a more involved two point method in the next section.

Define $P_0 := (\frac{1}{2}, \frac{1}{2(k-1)}, ..., \frac{1}{2(k-1)})$ and $P_1 := (\frac{1+\epsilon}{2}, \frac{1-\epsilon}{2(k-1)}, ..., \frac{1-\epsilon}{2(k-1)})$ for $\epsilon \in (0,1)$ and small. Then, we have that

$$\begin{aligned}D_{KL}(P_0||P_1) &= \frac{1}{2}\log\left(\frac{1}{1-\epsilon^2}\right)\\ &= O\left(\frac{1}{n}\right), \text{ using a Taylor expansion}\end{aligned}$$

if $\epsilon = O\left(\frac{1}{\sqrt{n}}\right)$. Then, by Property 19, we have that $D_{KL}(P_0^{\otimes n}||P_1^{\otimes n}) = nD_{KL}(P_0||P_1)$, which implies that $D_{KL}(P_0^{\otimes n}||P_1^{\otimes n}) = O(1)$. Then, using the result of Section 3.8, we see that $TV(P_0^{\otimes n}, P_1^{\otimes n}) = 1 - \Omega(1)$.

Also if $\epsilon = O\left(\frac{1}{\sqrt{n}}\right)$, we have that

$$\begin{aligned}|H(P_0) - H(P_1)| &= \left|\frac{1}{2}\log(2) + \frac{1}{2}\log(2(k-1)) - \frac{1-\epsilon}{2}\log\left(\frac{2}{1-\epsilon}\right) - \frac{1+\epsilon}{2}\log\left(\frac{2(k-1)}{1+\epsilon}\right)\right| \\ &\asymp \epsilon\log(k), \text{ by using a Taylor expansion on } \log(1-\epsilon) \text{ and } \log(1+\epsilon)\end{aligned}$$

Next, we see that

$$\begin{aligned}\inf_a (L(P_0,a) + L(P_1,a)) &= \inf_a (|a - H(P_0)| + |a - H(P_1)|) \\ &\geq |H(P_0) - H(P_1)|, \text{ using the reverse triangle inequality} \\ &\asymp \epsilon\log(k), \text{ if } \epsilon = O\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

There, choosing $\epsilon \asymp \frac{1}{\sqrt{n}}$ we can invole Theorem 109 (ie., Le Cam's Two Point Method) to say that

$$\begin{aligned}r^* &\geq \Theta(\epsilon\log(k))(1 - TV(P_0,P_1)) \\ &= \Theta(\epsilon\log(k))(1 - (1 - \Omega(1))) \\ &= \Omega\left(\frac{\log(k)}{\sqrt{n}}\right)\end{aligned}$$

[a] We use the notation that $g(n) \asymp f(n) \iff g(n) = \Theta(f(n))$.

Example 113 (Two-Armed Bandit). Suppose that $\theta = (\mu_1, \mu_2) \in [0,1]^2$. For $t \in [T]$, a learner pulls an arm $\pi_t \in \{1,2\}$ based on (π^{t-1}, r^{t-1}) and observes reward $r_t \overset{iid}{\sim} \mathcal{N}(\mu_{\pi_t}, 1)$. The learner aims to minimize the regret

$$R_T^\theta(\pi) := T\max(\mu_1,\mu_2) - \sum_{t=1}^T \mu_{\pi_t}$$

We wish to show that for $\Delta > 0$[a]

$$\begin{aligned}r^* &= \inf_\pi \sup_{\{\mu_1,\mu_2 : |\mu_1 - \mu_2| > \Delta\}} \mathbb{E}_\theta\left[R_T^\theta(\pi)\right] \\ &= \Omega\left(\frac{1 \vee \log(T\Delta^2)}{\Delta} \wedge T\Delta\right)\end{aligned}$$

In particular, choosing $\Delta \asymp \frac{1}{\sqrt{T}}$ gives the usual lower bound $\Omega(\sqrt{T})$ for two-armed bandits.

Proof.

First, by Property 19, we have that

$$\begin{aligned}D_{KL}(P_\theta || P_{\theta'}) &= \sum_{t=1}^T \mathbb{E}_\theta\left[\frac{(\mu_1 - \mu_1')^2}{2}\mathbb{1}_{\{\pi_t = 1\}} + \frac{(\mu_2 - \mu_2')^2}{2}\mathbb{1}_{\{\pi_t = 2\}}\right] \\ &= \frac{(\mu_1 - \mu_1')^2}{2}\mathbb{E}_\theta[T_1] + \frac{(\mu_2 - \mu_2')^2}{2}\mathbb{E}_\theta[T_2]\end{aligned}$$

where $T_i := \sum_{t=1}^T \mathbb{1}_{\{\pi_t = 1\}}$ for $i \in \{1,2\}$.

Next, choose two points $\theta = (\Delta, 0)$ and $\theta' = (\Delta, 2\Delta)$ for $\Delta > 0$. In this case, we have that $D_{KL}(P_\theta || P_{\theta'}) = 2\Delta^2\mathbb{E}_\theta[T_2]$. Notice that $\mathbb{E}_\theta[T_2]$ depends on the policy π that's chosen. From Section 3.8, recall that we have that $1 - TV(P_\theta, P_{\theta'}) \geq \frac{1}{2}\exp(-D_{KL}(P_\theta || P_{\theta'}))$.

Next, notice that

$$\begin{aligned}\inf_{\pi}\left(R_T^{\theta}(\pi)+R_T^{\theta'}(\pi)\right)&=T\max(\mu_1,\mu_2)-\sum_{t=1}^{T}\mu_{\pi_t}+T\max(\mu_1',\mu_2')-\sum_{t=1}^{T}\mu_{\pi_t}'\\&=\Delta T\end{aligned}$$

As a result, applying Theorem 109 (ie., Le Cam's Two Point Method), we have that

$$\begin{aligned}r^*&\geq\frac{T\Delta}{2}\left(1-TV(P_\theta,P_{\theta'})\right)\\&\geq\frac{T\Delta}{2}\left(\frac{1}{2}\exp\left(-D_{KL}(P_\theta||P_{\theta'})\right)\right)\\&\geq\Omega\left(T\Delta\exp(-2\Delta^2\mathbb{E}_\theta[T_2])\right)\end{aligned}$$

Notice that also using Theorem 94 (ie., Minimax Theorem), we have that

$$\begin{aligned}r^*&\geq\mathbb{E}_\theta[R_T^\theta(\pi)]\\&=\Delta\mathbb{E}_\theta[T_2]\end{aligned}$$

Thus, we have that

$$\begin{aligned}r^*&=\Omega\left(\max\left(\Delta\mathbb{E}_\theta[T_2],T\Delta\exp(-2\Delta^2\mathbb{E}_\theta[T_2])\right)\right)\\&=\Omega\left(\min_{T_0\in[0,T]}\max(\Delta T_0,T\Delta\exp(-2\Delta^2T_0))\right),\text{ since }\max\left(\Delta\mathbb{E}_\theta[T_2],...\right)\geq\min_{T_0\in[0,T]}\max(\Delta T_0,...)\\&=\Omega\left(\frac{1\vee\log(T\Delta^2)}{\Delta}\wedge T\Delta\right)\end{aligned}$$

The last step is justified since the minimax occurs either at ΔT, which is the value of the maximization problem on the boundary or when[b]

$$\begin{aligned}&\Delta T_0\asymp T\Delta\exp(-2\Delta^2T_0)\\\implies&T_0+\frac{1}{\Delta^2}\log(T_0)\asymp\frac{\log(T)\vee1}{\Delta^2}\\\implies&T_0\asymp\frac{\log(T\Delta^2)\vee1}{\Delta^2},\text{ assuming }\Delta=O(poly(T))\end{aligned}$$

which is when the increasing function and decreasing function inside the maximum intersect (the minimizer can force an interior solution). The presence of the $\vee 1$ is for the asymptotic regime where $1\gtrsim T\Delta^2$. Plugging the asymptotic equivalence of the terms back into the minimax, we get that the terms are asymptotically equal to

$$\frac{\log\left(T\Delta^2\right)\vee1}{\Delta}$$

Hence, we have justified the last step above that

$$\Omega\left(\min_{T_0\in[0,T]}\max(\Delta T_0,T\Delta\exp(-2\Delta^2T_0))\right)=\Omega\left(\frac{1\vee\log(T\Delta^2)}{\Delta}\wedge T\Delta\right)$$

□

[a] We use the notation that $x\wedge y:=\min(x,y)$ and $x\vee y:=\max(x,y)$.

[b] Note that we use the $\asymp$ notation and not $=$ when establishing the minimax since we're operating inside an $\Omega(\cdot)$.

Example 114 (Multi-Armed Bandit)**.** Assume the same observation model as in Example 113 but now with K arms so that $\theta = (\mu_1, ..., \mu_K) \in [0,1]^K$. We now analogously define the regret as

$$R_T^\theta(\pi) := T \max_{i \in [K]}(\mu_i) - \sum_{i=1}^{T} \mu_{\pi_t}$$

We wish to show that

$$\begin{aligned} r^* &= \inf_\pi \sup_\theta \mathbb{E}_\theta[R_T^\theta(\pi)] \\ &= \Omega\left(\sqrt{KT}\right) \end{aligned}$$

Proof.

Define $\theta_1 := (\Delta, 0, ..., 0)$ and $\theta_{2,i} := (\Delta, 0, ..., 0, 2\Delta, 0, ..., 0)$ for $i \in \{2, ..., K\}$. For each $i \in \{2, ..., K\}$, as in Example 113, we can show that $\inf_\pi \left(R_T^{\theta_1}(\pi) + R_T^{\theta_{2,i}}(\pi)\right) = T\Delta$. Similarly, as in Example 113, we can show that $D_{KL}(P_{\theta_1}||P_{\theta_{2,i}}) = 2\Delta^2 \mathbb{E}_{\theta_1}[T_i]$ where $T_i = \sum_{t=1}^{T} \mathbb{1}_{\{\pi_t = 1\}}$.

We next make the key observation that since $\sum_{i=2}^{K} \mathbb{E}_{\theta_1}[T_i] \leq T$, there $\exists i_0$ such that $\mathbb{E}_{\theta_1}[T_{i_0}] \leq \frac{T}{K-1}$, by the pigeonhole principle. Next, decide that $\Delta \asymp \sqrt{\frac{K}{T}}$ so that $D_{KL}(P_{\theta_1}||P_{\theta_{2,i_0}}) = O(1)$ since K is fixed and finite. Since $D_{KL}(P_{\theta_1}||P_{\theta_{2,i_0}}) = O(1)$, from Section 3.8, we have that $1 - TV(P_{\theta_1}, P_{\theta_{2,i_0}}) \geq \frac{1}{2}\exp(-D_{KL}(P_{\theta_1}||P_{\theta_{2,i_0}})) \implies 1 - TV(P_{\theta_1}, P_{\theta_{2,i_0}}) = \Omega(1)$. Therefore, applying Theorem 109 (ie., Le Cam's Two Point Method), we get that

$$\begin{aligned} r^* &\geq \frac{T\Delta}{2}(1 - TV(P_{\theta_1}, P_{\theta_{2,i_0}})) \\ \implies r^* &= \Omega\left(\sqrt{KT}\right) \end{aligned}$$

7.2 Leveraging Multiple Hypotheses

We now focus on techniques (ie., Assouad and Fano) to investigate the asymptotic performance of estimators using multiple hypotheses. Using multiple hypotheses to determine bounds on the asymptotic performance will sometimes perform better than just using two hypotheses as we will have more to work with.

Example 115 (Normal Mean Estimation in High Dimensions with just Two Hypotheses)**.** Suppose that $X \sim \mathcal{N}(\theta, \sigma^2 I_n)$ for unknown θ and known $\sigma^2 > 0$. The target is establishing bounds on $r^* := \inf_{\hat{\theta}} \sup_\theta \mathbb{E}_{X \sim \mathcal{N}(\theta, \sigma^2 I_n)}$ $[||\hat{\theta}(X) - \theta||^2]$ so that $L(\theta, a) = ||\theta - a||^2$. We can follow a procedure analogous to that in Example 95 and apply Theorem 109 to arrive at

$$\begin{aligned} r^* &\geq \sup_{\theta_1 \neq \theta_2} \frac{||\theta_0 - \theta_1||^2}{2}\left(1 - TV(\mathcal{N}(\theta_1, \sigma^2 I_n), \mathcal{N}(\theta_2, \sigma^2 I_n))\right) \\ &= \sup_{\theta_1 \neq \theta_2} \frac{||\theta_0 - \theta_1||^2}{2}\left(1 - \Phi\left(\frac{||\theta_1 - \theta_2||}{2\sigma}\right)\right) \\ &= \Omega(\sigma^2) \end{aligned}$$

Recall from Anderson's Lemma that $r^* = n\sigma^2$ so that the two point method doesn't capture the dependence on n.

Lemma 116 (Golden Formula for Mutual Information)**.** We have that

$$\begin{aligned} I(X;Y) &= \min_{Q_Y} D_{KL}(P_{XY}||P_X Q_Y) \\ &= \min_{Q_Y} \mathbb{E}_{P_X}\left[D_{KL}(P_{Y|X}||Q_Y)\right] \end{aligned}$$

Proof.

We have that

$$\begin{aligned} I(X;Y) &= D_{KL}(P_{XY}||P_X P_Y) \\ &= \mathbb{E}_{P_{XY}}\left[\log\left(\frac{P_{XY}}{P_X P_Y}\right)\right] \\ &= \mathbb{E}_{P_{XY}}\left[\log\left(\frac{P_{XY}}{P_X Q_Y}\right)\right] - \mathbb{E}_{P_{XY}}\left[\log\left(\frac{P_Y}{Q_Y}\right)\right], \text{ for any } Q_Y \\ &= D_{KL}(P_{XY}||P_X Q_Y) - D_{KL}(P_Y||Q_Y), \text{ for any } Q_Y \\ &\le D_{KL}(P_{XY}||P_X Q_Y)\ \forall Q_Y \end{aligned}$$

Next, since $I(X;Y) = D_{KL}(P_{XY}||P_X P_Y)$, we get that

$$I(X;Y) = \min_{Q_Y} D_{KL}(P_{XY}||P_X Q_Y)$$

The second line in the statement is by simple probability manipulations. □

Theorem 117 (Fano's Inequality)**.** Consider a statistical model $\{P_\theta : \theta \in \Theta\}$ and we wish to construct an estimator $\hat{\theta}(X)$ for θ from an observation $X \sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta, a)$. Let $\theta_1, ..., \theta_m \in \Theta$ satisfy the separation condition for $\Delta > 0$

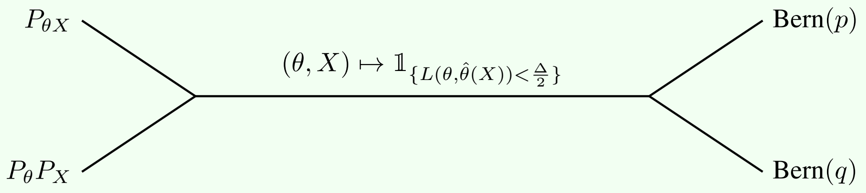
$$\min_{i \neq j} \inf_a \left(L(\theta_i, a) + L(\theta_j, a)\right) \ge \Delta$$

Then for $\pi = \text{Unif}(\{\theta_1, ..., \theta_m\})$, we have that

$$r_\pi^* \ge \frac{\Delta}{2}\left(1 - \frac{I(\theta; X) + \log(2)}{\log(m)}\right)$$

Proof.

Define $p := \Pr_{P_{\theta X}}\left(L(\theta, \hat{\theta}(X)) < \frac{\Delta}{2}\right)$ and $q := \Pr_{P_\theta P_X}\left(L(\theta, \hat{\theta}(X)) < \frac{\Delta}{2}\right)$. Construct the following channel.


$P_{\theta X}$
$P_\theta P_X$
$(\theta, X) \mapsto \mathbb{1}_{\{L(\theta,\hat{\theta}(X))<\frac{\Delta}{2}\}}$
Bern(p)
Bern(q)

Note that $q \le \frac{1}{m}$ by the separation condition and the uniform-ness of π. Thus Theorem 20 (ie., Data Processing

Inequality (DPI)) gives that

$$\begin{aligned} I(\theta; X) &= D_{KL}(P_{\theta X}||P_X P_\theta) \\ &\geq D_{KL}\left(\text{Bern}\,(p)\,||\text{Bern}(q)\right) \end{aligned}$$

Next, since $q \leq \frac{1}{m}$

$$\begin{aligned} D_{KL}(\text{Bern}(p)||\text{Bern}(q)) &= p\log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \\ &= \underbrace{p\log(p) + (1-p)\log(1-p)}_{\geq -\log(2)} + \underbrace{p\log\left(\frac{1}{q}\right)}_{\geq p\log\left(\frac{1}{m}\right)} + \underbrace{[-(1-p)\log(1-q)]}_{\geq 0} \\ &\geq -\log(2) + p\log(m) \end{aligned}$$

This bound with the above conclusion of the DPI and some rearrangement gives that

$$\begin{aligned} p &\leq \frac{I(\theta;X)+\log(2)}{\log(m)} \\ \implies 1-p &\geq 1 - \frac{I(\theta;X)+\log(2)}{\log(m)} \end{aligned}$$

By Markov's Inequality, we have that

$$\begin{aligned} \frac{\mathbb{E}_{P_\theta}\left[L(\theta,\hat{\theta}(X))\right]}{\Delta/2} &\geq 1-p \\ \implies \mathbb{E}_{P_{\theta X}}\left[L(\theta,\hat{\theta}(X))\right] &\geq \frac{\Delta}{2}\left(1 - \frac{I(\theta;X)+\log(2)}{\log(m)}\right) \\ \implies r_\pi^* &\geq \frac{\Delta}{2}\left(1 - \frac{I(\theta;X)+\log(2)}{\log(m)}\right), \text{ by definition} \end{aligned}$$

□

Theorem 118 (Generalized Fano)**.** Consider a statistical model $\{P_\theta : \theta \in \Theta\}$ and we wish to construct an estimator $\hat{\theta}(X)$ for θ from an observation $X \sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta, a)$. Then for any prior π on Θ, we have that

$$r_\pi^* \geq \frac{\Delta}{2}\left(1 - \frac{I(\theta;X)+\log(2)}{\log\left(\frac{1}{P_\Delta}\right)}\right)$$

where $P_\Delta := \sup_a \pi(L(\theta,a) < \Delta)$.

Proof.

There's a proof that mirrors that of Theorem 117 (ie., Fano's Inequality) and so is omitted.

Lemma 119 (Assouad's Lemma)**.** Consider a statistical model $\{P_\theta : \theta \in \Theta\}$ and we wish to construct an estimator $\hat{\theta}(X)$ for θ from an observation $X \sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta, a)$. For a hypercube parametrization $u \in \{-1,1\}^d$, associate $\theta_u \in \Theta$. Suppose that

$$\inf_a (L(\theta_u, a) + L(\theta_{u'}, a)) \geq \Delta \sum_{i=1}^{d} \mathbb{1}_{\{u_i \neq u'_i\}},\ \forall u, u' \in \{-1,1\}^d$$

then for the prior $\pi = \text{Unif}(\{\theta_u : u : \{-1,1\}^d\})$, we have that

$$r^*_\pi \geq \frac{\Delta}{4} \sum_{i=1}^{d} (1 - TV(P_{i,+}, P_{i,-}))$$

where $P_{i,+} := \frac{1}{2^{d-1}} \sum_{\{u:u_i=1\}} P_{\theta_u}$ and $P_{i,-} := \frac{1}{2^{d-1}} \sum_{\{u:u_i=-1\}} P_{\theta_u}$.

Proof.

For any estimator $\hat{\theta} = \hat{\theta}(X)$, construct $\hat{u} \in \{-1,1\}^d$ as follows:

$$\hat{u} := \operatorname*{arg\,min}_{u \in \{-1,1\}^d} L(\theta_u, \theta)$$

Then, $\forall u \in \{-1,1\}^d$, we have that

$$\begin{aligned}
L(\theta_u, \hat{\theta}) &\geq \frac{L(\theta_u, \hat{\theta}) + L(\theta_{\hat{u}}, \hat{\theta})}{2}, \text{ by definition of } \hat{u} \\
&\geq \frac{\Delta}{2} \sum_{i=1}^{d} \mathbb{1}_{\{u_i \neq \hat{u}_i\}}, \text{ by assumption} \\
\implies r^*_\pi &= \mathbb{E}_\pi[\mathbb{E}_{\theta_u}[L(\theta_u, \hat{\theta})]] \\
&= \frac{1}{2^d} \sum_{u \in \{-1,1\}^d} \mathbb{E}_{\theta_u}[L(\theta_u, \hat{\theta})] \\
&\geq \frac{1}{2^d} \sum_{u \in \{-1,1\}^d} \frac{\Delta}{2} \sum_{i=1}^{d} \Pr_{\theta_u}(u_i \neq \hat{u}_i) \\
&= \frac{\Delta}{4} \sum_{i=1}^{d} P_{i,+}(\hat{u}_i \neq 1) + P_{i,-}(\hat{u}_i \neq -1) \\
&\geq \frac{\Delta}{4} \sum_{i=1}^{d} (1 - TV(P_{i,+}, P_{i,-})), \text{ applying Theorem 38}
\end{aligned}$$

In the inequality of the last line, we can imagine that we're conducting a binary hypothesis test where $H_0 : P_{i,+}$ and $H_1 : P_{i,-}$ and we're summing the likelihood of Type I and Type II errors for which we have a characterization in Theorem 38. □

Corollary 120 (Classical Assouad)**.** Take the context of Lemma 119. We have that

$$r^*_\pi \geq \frac{d\Delta}{4} \left(1 - \max_{\{u,u' : u,u' \text{ are neighbors}\}} TV(P_{\theta_u}, P_{\theta_{u'}})\right)$$

Proof.

This result follows trivially from Lemma 119 since $d \max_i(\text{summand}_i) \geq \sum_{i=1}^{d} \text{summand}_i$. □

Corollary 121 (Other Assouad Corollary)**.** Take the context of Lemma 119. We have that

$$r_\pi^* \geq \frac{d\Delta}{4}\left(1 - \mathbb{E}_{u\sim\text{Unif}(\{-1,1\}^d)}\left[\mathbb{E}_{i\in\text{Unif}([d])}\left[TV(P_{\theta_u}, P_{\theta_u,-i})\right]\right]\right)$$

where $P_{\theta_u,-i}$ is P_{θ_u} with the ith bit flipped.

Proof.

We have by Lemma 119,

$$\begin{aligned}
r_\pi^* &\geq \frac{\Delta}{4}\sum_{i=1}^d (1 - TV(P_{i,+}, P_{i,-})) \\
&= \frac{\Delta}{4}\sum_{i=1}^d \left(1 - TV\left(\frac{1}{2^{d-1}}\sum_{\{u:u_i=1\}} P_{\theta_u}, \frac{1}{2^{d-1}}\sum_{\{u:u_i=-1\}} P_{\theta_u}\right)\right) \\
&= \frac{\Delta}{4}\sum_{i=1}^d \left(1 - TV\left(\mathbb{E}_{\mathit{Unif}(\{u:u_i=1\})}[P_{\theta_u}], \mathbb{E}_{\text{Unif}(\{u:u_i=1\})}[P_{\theta_u,-i}]\right)\right) \\
&= \frac{\Delta}{4}\sum_{i=1}^d \left(1 - \mathbb{E}_{\text{Unif}(\{u:u_i=1\})}\left[TV(P_{\theta_u}, P_{\theta_u,-i})\right]\right) \\
&= \frac{d\Delta}{4}\left(1 - \mathbb{E}_{\text{Unif}(\{u:u_i=1\})}\left[\frac{1}{d}\sum_{i=1}^d TV(P_{\theta_u}, P_{\theta_u,-i})\right]\right) \\
&= \frac{d\Delta}{4}\left(1 - \mathbb{E}_{\text{Unif}(\{u:u_i=1\})}\left[\mathbb{E}_{i\in\text{Unif}([d])}\left[TV(P_{\theta_u}, P_{\theta_u,-i})\right]\right]\right)
\end{aligned}$$

□

Lemma 122 (Gilbert-Varshamov)**.** There exists $A \subseteq \{-1,1\}^n$ with $m := |A|$ such that $\min_{u\neq u'\in A}\sum_{i=1}^n \mathbb{1}_{\{u_i\neq u_i'\}} \geq d$ and

$$\begin{aligned}
m &\geq \frac{2^n}{\sum_{j=0}^{d-1}\binom{n}{j}} \\
&= 2^{n\left(1-h_2\left(\frac{d}{n}\right)\right)+o(n)}
\end{aligned}$$

where as a reminder $h_2(x) = x\log\left(\frac{1}{x}\right) + (1-x)\log\left(\frac{1}{1-x}\right)$ and $\frac{d}{n}$ is constant along an asymptotic sequence. We report a slightly more general version of this lemma in Example 148.

Proof.

We will use a volume argument to show the result. We have that $\forall u \in \{-1,1\}^n$,

$$\sum_{j=1}^{d-1}\binom{n}{j} = \left|\{u' \in \{-1,1\}^n : \sum_{i=1}^n \mathbb{1}_{\{u_i\neq u_i'\}} \leq d-1\}\right|$$

Thus, if $m < \frac{2^n}{\sum_{j=0}^{n-1}\binom{n}{j}}$, there must exist some $u \in \{-1,1\}^n \setminus A$ with distance $\geq d$ to all existing points in A. Thus, we can iteratively construct A in this way satisfying the given inequality. The last equality follows from Stirling's approximation. From Stirling's approximation, We have that

$$\begin{aligned}
n! &= \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1+o(1)) \\
\implies \binom{n}{j} &= \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi(n-j)}\left(\frac{n-j}{e}\right)^{n-j}\sqrt{2\pi j}\left(\frac{j}{e}\right)^j}(1+o(1)), \text{ for } j \in [0:n] \\
&= \frac{1}{\sqrt{2\pi\frac{j(n-j)}{n}}}\left(\frac{n^n}{(n-j)^{n-j}j^j}\right)(1+o(1)) \\
&= \frac{1}{\sqrt{2\pi\frac{j(n-j)}{n}}}\left(\frac{1}{(1-\frac{j}{n})^{1-\frac{j}{n}}(\frac{j}{n})^{\frac{j}{n}}}\right)^n (1+o(1)) \\
&= \frac{1}{\sqrt{2\pi\frac{j(n-j)}{n}}}\left(2^{nh_2(\frac{j}{n})}\right)(1+o(1)) \\
&= \left(2^{nh_2(\frac{j}{n})-o(\log_2(n))+O(1)}\right)(1+o(1)) \\
&= 2^{nh_2(\frac{j}{n})+o(n)}, \text{ absorbing } (1+o(1)) \text{ into the exponent}
\end{aligned}$$

WLOG, suppose that n is odd and $d \leq \frac{n+1}{2}$ (otherwise relabel $d := n-d$). Then we have that

$$\begin{aligned}
\sum_{j=0}^{d-1}\binom{n}{j} &\leq d\binom{n}{d} \\
\implies \sum_{j=0}^{d-1}\binom{n}{j} &= 2^{nh_2(\frac{d}{n})+o(n)+\log_2(d)} \\
&= 2^{nh_2(\frac{d}{n})+o(n)}
\end{aligned}$$

The second equality in the statement now follows trivially. □

Example 123 (Normal Mean Estimation in High Dimensions with Multiple Hypotheses)**.** Suppose that $X \sim \mathcal{N}(\theta, \sigma^2 I_n)$ for unknown θ and known $\sigma^2 > 0$. The target is establishing bounds on $r^* := \inf_{\hat{\theta}}\sup_\theta \mathbb{E}_{X\sim\mathcal{N}(\theta,\sigma^2 I_n)}$ $[||\hat{\theta}(X)-\theta||^2]$ so that $L(\theta,a) = ||\theta-a||^2$. We will show using a few different approaches to show that $r^* = \Omega(n\sigma^2)$.

Proof.

[Method 1 – Fano]:

Construct a subset $\Theta_0 \subseteq \{-\delta,\delta\}^n$ (with δ TBD) such that $m := |\Theta_0|$ is large enough and

$$\begin{aligned}
\min_{\{\theta\neq\theta':\theta'\in\Theta_0\}} ||\theta-\theta'||^2 &= \min_{\{\theta\neq\theta':\theta'\in\Theta\}} 4\delta^2 \underbrace{\sum_{i=1}^n \mathbb{1}_{\{\theta_i\neq\theta'_i\}}}_{\geq\frac{n}{4}} \\
&\geq \delta^2 n
\end{aligned}$$

which is possible by Lemma 122 (ie., Gilbert-Varshamov) with $m := \exp(\Omega(n))$. Then, by Lemma 116 (ie., Golden

Formula for Mutual Information), we have that

$$\begin{aligned}
I(\theta; X) &= \min_{Q} \mathbb{E}_{\theta \sim \text{Unif}(\Theta_0)}[D_{KL}(\mathcal{N}(\theta, \sigma^2 I_n)||Q)] \\
&\leq \mathbb{E}_{\theta \sim \text{Unif}(\Theta_0)}[D_{KL}(\mathcal{N}(\theta, \sigma^2 I_n)||\mathcal{N}(0, \sigma^2 I_n))] \\
&\leq \max_{\theta \in \Theta_0} D_{KL}(\mathcal{N}(\theta, \sigma^2 I_n)||\mathcal{N}(0, \sigma^2 I_n)) \\
&= \max_{\theta \in \Theta_0} \frac{||\theta||^2}{2\sigma^2} \\
&= \frac{n\delta^2}{2\sigma^2}, \text{ since } \theta \in \{-\delta, \delta\}^n
\end{aligned}$$

In turn, Theorem 117 (ie., Fano's Inequality) with $\Delta := \delta^2 n$ and Theorem 94 (ie., Minimax Theorem) imply that

$$\begin{aligned}
r^* &\geq r^*_\pi \text{ where } \pi = \text{Unif}(\Theta_0) \\
&= \Omega\left(\Delta\left(1 - \frac{I(\theta; X) + \log(2)}{\log(m)}\right)\right) \\
&= \Omega\left(\delta^2 n\left(1 - \frac{n\frac{\delta^2}{2\sigma^2} + \log(2)}{\Omega(n)}\right)\right) \\
\implies r^* &= \Omega(n\sigma^2)
\end{aligned}$$

if we pick $\delta \asymp \sigma$.

[Method 2 – Generalized Fano]:

Let $\pi := \text{Unif}(\{-1, 1\}^n)$ and $\theta \sim \pi$. Then, by a sequence of steps in the Method 1 proof, we have that $I(\theta; X) \leq \frac{n\delta^2}{2\sigma^2}$. Choose $\Delta := \frac{n\delta^2}{12}$ for δ TBD. Then, we have that as defined in Theorem 118 (ie., Generalized Fano),

$$\begin{aligned}
P_\Delta &:= \sup_{a \in \mathbb{R}^n} \pi\left(||\theta - a||^2 < \Delta\right) \\
&\leq \sup_{\hat{\theta} \in \{-\delta, \delta\}^n} \pi\left(||\theta - \hat{\theta}||^2 < 4\Delta\right) \\
&\leq \frac{1}{2^n} \sum_{j=0}^{\lceil n/12 \rceil} \binom{n}{j} \\
&= 2^{-\Omega(n)}, \text{ using Stirling's approximation like in Lemma 122}
\end{aligned}$$

The first inequality follows from the fact that we can define $b_i(a) := \delta \cdot \text{sign}(a)$ so that $b(a) \in \{-\delta, \delta\}^n \; \forall a \in \mathbb{R}^n$. We in turn see that $\forall \theta \in \{-\delta, \delta\}^n$, we have that

$$\begin{aligned}
&|b_i(a) - \theta_i| \leq 2|a_i - \theta_i| \\
\implies &||b(a) - \theta||^2 \leq 4||a - \theta||^2 \\
\implies &\left(a \in \{\tilde{a} \in \mathbb{R}^n : ||\theta - \tilde{a}||^2 < \Delta\} \implies b(a) \in \{\tilde{b} \in \{-\delta, \delta\}^n : ||\tilde{b} - \theta||^2 < 4\Delta\}\right)
\end{aligned}$$

which implies the first inequality. The second inequality follows from the fact that for any $\theta, \hat{\theta} \in \{-\delta, \delta\}$, $||\hat{\theta} - \theta|| =$

$4\delta^2 \sum_{i=1}^n \mathbb{1}_{\{\theta_i \neq \hat{\theta}_i\}}$ so that

$$\begin{aligned}\pi\left(||\theta - \hat{\theta}||^2 < 4\Delta\right) &= \pi\left(4\delta^2 \sum_{i=1}^n \mathbb{1}_{\{\theta_i \neq \hat{\theta}_i\}} < 4\Delta\right) \\ &= \pi\left(\sum_{i=1}^n \mathbb{1}_{\{\theta_i \neq \hat{\theta}_i\}} < \frac{n}{12}\right) \\ &\leq \frac{1}{2^n} \sum_{j=0}^{\lceil n/12 \rceil} \binom{n}{j}\end{aligned}$$

Therefore, invoking Theorem 118 and Theorem 94 (ie., Minimax Theorem), we get that

$$\begin{aligned}r^* &\geq r^*_\pi \\ &= \Omega\left(\Delta\left(1 - \frac{I(\theta; X) + \log(2)}{\log\left(1/P_\Delta\right)}\right)\right) \\ &= \Omega\left(n\delta^2\left(1 - \frac{\frac{n\delta^2}{2\sigma^2} + \log(2)}{\Omega(n)}\right)\right) \\ &= \Omega(n\sigma^2)\end{aligned}$$

when we pick $\delta \asymp \sigma$.

[Method 3 – Classical Assouad]:

For $\delta > 0$, let $\theta_u := \delta u$ with $u \in \{-1, 1\}^d$. Then, for $\Delta := 2\delta^2$ and using Theorem 94 (ie., Minimax Theorem) and Corollary 120 (ie., Classical Assouad), we get that

$$\begin{aligned}r^* &\geq r^*_\pi \\ &\geq \frac{n\Delta}{4}\left(1 - \max_{\{u,u':u,u' \text{ are neighbors}\}} TV(P_{\theta_u}, P_{\theta_{u'}})\right) \\ &= \frac{n\delta^2}{2}\left(1 - \max_{\{u,u':u,u' \text{ are neighbors}\}} TV(\mathcal{N}(\theta_u, \sigma^2 I_n), \mathcal{N}(\theta_{u'}, \sigma^2 I_n))\right) \\ &= \frac{n\delta^2}{2}\left(1 - \max_{\{u,u':u,u' \text{ are neighbors}\}} \Phi\left(\frac{||\theta_u - \theta_{u'}||}{2\sigma}\right)\right) \\ &= \frac{n\delta^2}{2}\left(1 - \max_{\{u,u':u,u' \text{ are neighbors}\}} \Phi\left(\frac{2\delta}{2\sigma}\right)\right) \\ &= \frac{n\delta^2}{2}\left(1 - \Phi\left(\frac{\delta}{\sigma}\right)\right) \\ &= \Omega(n\sigma^2)\end{aligned}$$

if we pick $\delta = \sigma$. □

Definition 124 (VC Dimension)**.** A class of functions $\mathcal{F} \subseteq \{f : \mathcal{X} \to \{0, 1\}\}$ has VC-dimension at least d if $\exists x_1, ..., x_d \in \mathcal{X}$ so that $\forall u \in \{-1, 1\}^d$, there exists $f_u \in \mathcal{F}$ such that $f_u(x_i) = u_i \ \forall i \in [d]$. The VC dimension of $\mathcal{F}$ is the largest $d \in \mathbb{N}$ so that $\mathcal{F}$ has VC dimension at least d. This concept captures the expressive power of the function class $\mathcal{F}$ (ie., how many points can this class of functions label arbitrarily).

Example 125 (Learning Theory)**.** Let $X_i, Y_i \overset{iid}{\sim} P_{XY}$ for $i \in [n]$ with P_{XY} unknown, $\text{supp}(X_i) = \mathcal{X}$, and $\text{supp}(Y_i) = \{-1, 1\}$. Let $\mathcal{F}$ be a given class of functions $\mathcal{X} \to \{0, 1\}$ with VC dimension d. We define the excess risk for a classifier $\hat{f} : \mathcal{X} \to \{0, 1\}$ trained on $\{(X_i, Y_i) : i \in [n]\}$ as

$$\begin{aligned} ER(\hat{f}) &:= ER(\hat{f}; P_{XY}) \\ &= P_{XY}(Y \neq \hat{f}(X)) - \min_{f \in \mathcal{F}} P_{XY}(Y \neq f(X)) \end{aligned}$$

We will show that for $n \geq d$

$$\inf_{\hat{f} \in \mathcal{F}} \sup_{P_{XY}} \mathbb{E}_{P_{XY}}[ER(\hat{f})] = \Omega\left(\sqrt{\frac{d}{n}}\right)$$

$$\inf_{\hat{f} \in \mathcal{F}} \sup_{\{P_{XY} : \exists f \in \mathcal{F} \text{ s.t. } Y = f(X)\ a.s.\}} \mathbb{E}_{P_{XY}}[ER(\hat{f})] = \Omega\left(\frac{d}{n}\right)$$

where the first result is in an agnostic setting and the second result is in a realizable setting. Clearly, dividing by n is reduces the excess risk more than dividing by $\sqrt{n}$ so that we can do better in the realizable setting as expected.

Proof.

[Agnostic Setting]:

Take any $x_1, ..., x_d \in \mathcal{X}$ and $\{f_u\}_{u \in \{-1,1\}^d} \subseteq \mathcal{F}$ as in Definition 124. Under $u \in \{-1, 1\}^d$, construct $P_u := P_{XY,u}$ as, for some $\delta > 0$,

$$\begin{aligned} &X \sim \text{Unif}(\{x_1, ..., x_d\}) \\ &Y \mid X = x_i = \begin{cases} u_i & \text{w.p. } \frac{1}{2} + \delta \\ -u_i & \text{w.p. } \frac{1}{2} - \delta \end{cases} \end{aligned}$$

We note that $\min_{f \in \mathcal{F}} P_u(f(X) \neq Y) = \frac{1}{2} - \delta\ \forall u$ since conditional on x_i we always wish to predict u_i and will be wrong with probability $\frac{1}{2} - \delta$. We then have that $\forall u, u'$ and the corresponding f_u

$$\begin{aligned} ER(f_u, P_u) + ER(f_u, P_{u'}) &= \underbrace{P_u(Y \neq f_u(X))}_{=\frac{1}{2}-\delta} + P_{u'}(Y \neq f_u(X)) - 2\left(\frac{1}{2} - \delta\right) \\ &= \sum_{i=1}^{d} \frac{1}{d}\left(\mathbb{1}_{\{u_i \neq u_i'\}}\left(\frac{1}{2} + \delta\right) + \mathbb{1}_{\{u_i = u_i'\}}\left(\frac{1}{2} - \delta\right)\right) - \left(\frac{1}{2} - \delta\right) \\ &= \sum_{i=1}^{d} \frac{1}{d}\left(\mathbb{1}_{\{u_i \neq u_i'\}}\left(\frac{1}{2} + \delta\right) + (1 - \mathbb{1}_{\{u_i \neq u_i'\}})\left(\frac{1}{2} - \delta\right)\right) - \left(\frac{1}{2} - \delta\right) \\ &= \frac{(1/2 + 2\delta)}{d} \sum_{i=1}^{d} \mathbb{1}_{\{u_i \neq u_i'\}} \\ &\geq \frac{2\delta}{d} \sum_{i=1}^{d} \mathbb{1}_{\{u_i \neq u_i'\}} \\ &\implies \text{define } \Delta := \frac{2\delta}{d} \end{aligned}$$

for Assouad's. For any neighboring u, u', we have that

$$\begin{aligned} D_{KL}\left(P_u^{\otimes n}||P_{u'}^{\otimes n}\right) &= nD_{KL}(P_u||P_{u'}), \text{ by Property 19} \\ &= n\left(\frac{1}{d}D_{KL}\left(\text{Bern}\left(\frac{1}{2}+\delta\right)\right)||\text{Bern}\left(\frac{1}{2}-\delta\right)\right) \\ &= \frac{n}{d}\left(2\delta\log\left(\frac{1/2+\delta}{1/2-\delta}\right)\right) \\ &= \frac{2\delta n}{d}(4\delta + O(\delta^2)) \\ &= O\left(\frac{n\delta^2}{d}\right) \end{aligned}$$

Thus, applying Corollary 120 (ie., Classical Assouad), leveraging Theorem 94 (ie., Minimax Theorem), and using the fact from Section 3.8 that $TV(P_u^{\otimes n}, P_{u'}^{\otimes n}) \leq \frac{1}{2}\sqrt{D_{KL}(P_u^{\otimes n}||P_{u'}^{\otimes n})}$, with $u \sim \pi := \text{Unif}(\{-1,1\}^d)$, we get that

$$\begin{aligned} r^* &\geq r_\pi^* \\ &= \frac{d\Delta}{4}\left(1 - \max_{\{u,u':u,u' \text{ are neighbors}\}} TV(P_u^{\otimes n}, P_{u'}^{\otimes n})\right) \\ &= \frac{d}{4}\left(\frac{2\delta}{d}\right)\left(1 - O\left(\sqrt{\frac{n\delta^2}{d}}\right)\right) \\ &= \Omega\left(\sqrt{\frac{d}{n}}\right), \text{ for } \delta \asymp \sqrt{\frac{d}{n}} \end{aligned}$$

As a remark, we use the fact that $n \geq d$ to validate our expansion of $D_{KL}(P_u^{\otimes n}||P_{u'}^{\otimes n})$ when we pick $\delta \asymp \sqrt{\frac{d}{n}}$.

[Realizable Setting]:

Take any $x_1, ..., x_d \in \mathcal{X}$ and $\{f_u\}_{u \in \{1\}\times\{-1,1\}^{d-1}} \subseteq \mathcal{F}$ as in Definition 124. Now consider $u \in \{1\} \times \{-1,1\}^{d-1}$ and define $P_u := P_{XY,u}$ as

$$\begin{aligned} &X = \begin{cases} x_1 & \text{w.p. } 1 - \frac{d-1}{n}, \text{ okay since } n \geq d \text{ by assumption} \\ x_i & \text{w.p. } \frac{1}{n} \text{ for } 2 \leq i \leq d \end{cases} \\ &Y \mid X = x_i = u_i \text{ w.p. } 1 \end{aligned}$$

Clearly, $\min_{f\in\mathcal{F}} P_u(Y \neq f(X)) = 0 \; \forall u$ since conditional on x_i we're always correct if we predict u_i. Then, we have that $\forall u, u'$ and the corresponding f_u

$$\begin{aligned} ER(f_u, P_u) + ER(f_u, P_{u'}) &= \overset{0}{\cancel{P_u(Y \neq f_u(X))}} + P_{u'}(Y \neq f_u(X)) \\ &= \sum_{i=1}^{d} P_{u'}(X = x_i)\mathbb{1}_{\{u_i \neq u_i'\}} \\ &= \frac{1}{n}\sum_{i=2}^{d} \mathbb{1}_{\{u_i \neq u_i'\}} \\ &= \frac{1}{n}\sum_{i=1}^{d} \mathbb{1}_{\{u_i \neq u_i'\}} \\ &\implies \text{define } \Delta := \frac{1}{n} \end{aligned}$$

for Assouad's. Define $u' := u_{-i}$ for u to be u with the ith entry flipped for $2 \le i \le d$. Note that P_u and $P_{u'}$ are exactly the same except when $X = x_i$ for $i \in \{2, ..., d\}$. Define $B := \{X_1, ..., X_n \in \{x_1, ..., x_d\}^n : x_j \in \{X_1, ..., X_n\}\}$ we then have that, using the characterization of TV distance in Theorem 38,

$$\begin{aligned}
TV(P_u^{\otimes n}, P_{u'}^{\otimes n}) &= \max_A |P_u^{\otimes n}(A) - P_{u'}^{\otimes n}(A)| \\
&= \max_A |P_u^{\otimes n}(A \cap B) - P_{u'}^{\otimes n}(A \cap B) + \underbrace{[P_u^{\otimes n}(A \cap B^c) - P_{u'}^{\otimes n}(A \cap B^c)]}_{=0}| \\
&\le \max_A \max(P_u^{\otimes n}(A \cap B), P_{u'}^{\otimes n}(A \cap B)) \\
&\le \max_A \max(P_u^{\otimes n}(B), P_{u'}^{\otimes n}(B)) \\
&= P_u^{\otimes n}(B), \text{ since } P_u^{\otimes n} \text{ and } P_{u'}^{\otimes n} \text{ match on } X\text{'s marginal distribution} \\
&= 1 - \left(1 - \frac{1}{n}\right)^n \\
&= 1 - \Omega(1)
\end{aligned}$$

Therefore, we can apply Corollary 120 and Theorem 94 with $u \sim \pi :=$ Unif$(\{1\} \times \{-1, 1\}^{d-1})$ to get that

$$\begin{aligned}
r^* &\ge r_\pi \\
&= \frac{d\Delta}{4}\left(1 - \max_{\{u,u' : u,u' \text{ are neighbors}\}} TV(P_u^{\otimes n}, P_{u'}^{\otimes n})\right) \\
&= \frac{d}{4n}\Omega(1 - (1 - \Omega(1)) \\
&= \Omega\left(\frac{d}{n}\right)
\end{aligned}$$

□

8 Advanced Le Cam's Method

In a general binary hypothesis testing setting, we have

- $H_0 : \theta \in \Theta_0$
- $H_1 : \theta \in \Theta_1$

In Section 7, we have considered the simple case so far where Θ_0 and Θ_1 are singletons but we can consider also a more general setting where they're arbitrary sets in what we call a composite binary hypothesis test. In that case, for a test $T \in \{0, 1\}$, we write that

- Type I error: $\sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T = 1)$
- Type II error: $\sup_{\theta_1 \in \Theta_1} P_{\theta_1}(T = 0)$

Theorem 126 (Characterization of the Composite Binary Hypothesis Test)**.** Given the setting above, we have that

$$\inf_T \left(\sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T = 1) + \sup_{\theta_1 \in \Theta_1} P_{\theta_1}(T = 0)\right) = 1 - \inf_{\pi_0, \pi_1 \in P(\Theta_0) \times P(\Theta_1)} TV\left(\mathbb{E}_{\theta_0 \sim \pi_0}[P_{\theta_0}], \mathbb{E}_{\theta_1 \sim \pi_1}[P_{\theta_1}]\right)$$

Proof.

We define

$$r(T) := \sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T=1) + \sup_{\theta_1 \in \Theta_1} P_{\theta_1}(T=0)$$

Following a sequence similar to that in the proof of Theorem 94 (ie., Minimax Theorem), we have that

$$\begin{aligned}
\inf_T r(T) &= \inf_T \left(\sup_{\theta_0 \in \Theta_0} P_{\theta_0}(T=1) + \sup_{\theta_1 \in \Theta_1} P_{\theta_1}(T=0) \right) \\
&= \inf_T \sup_{\pi_0, \pi_1} \left(\mathbb{E}_{\theta_0 \sim \pi_0}[P_{\theta_0}(T=1)] + \mathbb{E}_{\theta_1 \sim \pi_1}[P_{\theta_1}(T=0)] \right) \\
&= \sup_{\pi_0, \pi_1} \inf_T \left(\mathbb{E}_{\theta_0 \sim \pi_0}[P_{\theta_0}(T=1)] + \mathbb{E}_{\theta_1 \sim \pi_1}[P_{\theta_1}(T=0)] \right), \text{ under regularity conditions} \\
&= \sup_{\pi_0, \pi_1} \inf_T P_{\pi_0}(T=1) + P_{\pi_1}(T=0)
\end{aligned}$$

where we define the probability distributions $P_{\pi_0} := \mathbb{E}_{\theta_0 \sim \pi_0}[P_{\theta_0}]$ and $P_{\pi_1} := \mathbb{E}_{\theta_1 \sim \pi_1}[P_{\theta_1}]$ for any fixed priors $\pi_0 \in P(\Theta_0)$ and $\pi_1 \in P(\Theta_1)$. Then, we have that

$$\begin{aligned}
\sup_{\pi_0, \pi_1} \inf_T P_{\pi_0}(T(X)=1) + P_{\pi_1}(T(X)=0) &= \sup_{\pi_0, \pi_1} \inf_T [1 - (P_{\pi_0}(T(X)=0) - P_{\pi_1}(T(X)=0))] \\
&= 1 - \inf_{\pi_0, \pi_1} \sup_T (P_{\pi_0}(T(X)=0) - P_{\pi_1}(T(X)=0)) \\
&= 1 - \inf_{\pi_0, \pi_1} \sup_A (P_{\pi_0}(X \in A) - P_{\pi_1}(X \in A)) \\
&= 1 - \inf_{\pi_0, \pi_1} TV(P_{\pi_0}, P_{\pi_1})
\end{aligned}$$

where the last line uses the characterization of TV distance from Theorem 38. □

8.1 Point versus Mixture

In this section, we focus on using the Le Cam method with a point versus mixture approach in a binary hypothesis setting to present asymptotic lower bounds on risk.

Theorem 127 (Advanced Le Cam: Point vs. Mixture)**.** Consider a statistical model $\{P_\theta : \theta \in \Theta\}$. Let $\theta_0 \in \Theta$ and $\Theta_1 \subseteq \Theta \setminus \{\theta_0\}$. We wish to take an action a based on θ from an observation $X \sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta, a)$.[a] In many cases the action a is a test $T : \mathcal{X} \to \{0,1\}$ and we wish to decide if $\theta = \theta_0$ or $\theta \in \Theta_1$.[b] Suppose that

$$\inf_{\theta_1 \in \Theta_1} \min_a L(\theta_0, a) + L(\theta_1, a) \geq \Delta$$

Then, for any probability distribution π on Θ_1,

$$\begin{aligned}
r^* &= \inf_a \sup_{\theta \in \{\theta_0\} \cup \Theta_1} \mathbb{E}_\theta[L(\theta, a(X))] \\
&\geq \frac{\Delta}{2} (1 - TV(P_{\theta_0}, \mathbb{E}_{\theta_1 \sim \pi}[P_{\theta_1}]))
\end{aligned}$$

Proof.

Take any probability distribution $\pi \in P(\Theta_1)$. We then have that

$$\begin{aligned} r^* &= \inf_a \sup_{\theta \in \{\theta_0\} \cup \Theta_1} \mathbb{E}_\theta[L(\theta, a(X))] \\ &\geq \inf_a \frac{1}{2}\mathbb{E}_{\theta_0}[L(\theta_0, a(X))] + \frac{1}{2}\mathbb{E}_\pi[\mathbb{E}_\theta[L(\theta, a(X))]] \\ &\geq \frac{\Delta}{2}\int_{x \in \mathcal{X}} \min(p_{\theta_0}(x), p_\pi(x))dx, \text{ by the separation condition where } p_\pi := \mathbb{E}_\pi[p_\theta] \\ &= \frac{\Delta}{2}\int_{x \in \mathcal{X}} \frac{1}{2}[p_{\theta_0}(x) + p_\pi(x) - |p_{\theta_0}(x) - p_\pi(x)|]dx \\ &= \frac{\Delta}{2}\left(1 - TV(P_{\theta_0}, P_\pi)\right), \text{ by the definition of } TV \text{ distance} \end{aligned}$$

□

[a] Assume that the models admit a density with respect to a measure on $\mathcal{X}$ though the measure-theoretic details are skirted here.

[b] Note that the separation condition here trivially holds. $L(\theta, T(x)) = \mathbb{1}_{\{T(x) \neq \mathbb{1}_{\{\theta \neq \theta_0\}}\}}$. We have that for any (deterministic) test T and any sample x, $L(\theta_0, T(x)) + L(\theta_1, T(x)) = 1$. Note that we can also let T be randomized and arrive at the same conclusion.

How do we usually think about upper-bounding $TV(P_{\theta_0}, \mathbb{E}_\pi[P_{\theta_1}])$ when applying Theorem 127?

- By Property 36, we have that $TV(P_{\theta_0}, \mathbb{E}_\pi[P_{\theta_1}]) \leq \mathbb{E}_{\theta_1 \sim \pi}[TV(P_{\theta_0}, P_{\theta_1})]$.
- We can also upper bound the χ^2 divergence $\chi^2(\mathbb{E}_\pi[P_{\theta_1}]||P_{\theta_0})$ and then leverage some inequalities between f-divergences in Section 3.8. A useful characterization of the χ^2 divergence is covered in Theorem 128.

Theorem 128 (χ^2 Method)**.** Consider a statistical model $\{P_\theta : \theta \in \Theta\}$. Let $\theta_0 \in \Theta$ and $\Theta_1 \subseteq \Theta \setminus \{\theta_0\}$. Suppose that P_θ has support $\mathcal{X}$. Let $\pi \in P(\Theta_1)$ be any distribution and suppose that $\theta_1, \theta_1' \overset{iid}{\sim} \pi$. Then, we have that[a]

$$\chi^2(\mathbb{E}_\pi[P_{\theta_1}]||P_{\theta_0}) = \mathbb{E}_{\theta_1, \theta_1' \overset{iid}{\sim} \pi}\left[\int_{x \in \mathcal{X}} \frac{P_{\theta_1}(x)P_{\theta_1'}(x)}{P_{\theta_0}(x)}dx\right] - 1$$

Proof.

We have that

$$\begin{aligned} \chi^2(\mathbb{E}_\pi[P_{\theta_1}]||P_{\theta_0}) + 1 &= \int_{x \in \mathcal{X}} \frac{(\mathbb{E}_\pi[P_{\theta_1}](x) - P_{\theta_0}(x))^2}{P_{\theta_0}(x)}dx + 1 \\ &= \int_{x \in \mathcal{X}} \frac{(\mathbb{E}_\pi[P_{\theta_1}](x))^2}{P_{\theta_0}(x)} + P_{\theta_0}(x) - 2\mathbb{E}_\pi[P_{\theta_1}](x)dx + 1 \\ &= \int_{x \in \mathcal{X}} \frac{(\mathbb{E}_\pi[P_{\theta_1}](x))^2}{P_{\theta_0}(x)}dx \\ &= \int_{x \in \mathcal{X}} \frac{\mathbb{E}_{\theta_1, \theta_1' \overset{iid}{\sim} \pi}[P_{\theta_1}(x)P_{\theta_1'}(x)]}{P_{\theta_0}(x)}dx \\ &= \mathbb{E}_{\theta_1, \theta_1' \overset{iid}{\sim} \pi}\left[\int_{x \in \mathcal{X}} \frac{P_{\theta_1}(x)P_{\theta_1'}(x)}{P_{\theta_0}(x)}dx\right], \text{ by Fubini} \end{aligned}$$

□

[a] As before, again here, we skirt measure-theoretic details.

Corollary 129 (χ^2 Method for iid Samples)**.** Take the context of Theorem 128. We have that

$$\chi^2\left(\mathbb{E}_\pi[P_{\theta_1}^{\otimes n}]||P_{\theta_0}^{\otimes n}\right)+1=\mathbb{E}_{\theta_1,\theta_1'\overset{iid}{\sim}\pi}\left[\left(\int_{x\in\mathcal{X}}\frac{P_{\theta_1}(x)P_{\theta_1'}(x)}{P_{\theta_0}(x)}dx\right)^n\right]$$

Proof.

To see the result, apply Theorem 128 with the additional observation that

$$\int_{x\in\mathcal{X}}\frac{P_{\theta_1}^{\otimes n}(x)P_{\theta_1'}^{\otimes n}(x)}{P_{\theta_0}^{\otimes n}(x)}dx=\left(\int_{x\in\mathcal{X}}\frac{P_{\theta_1}(x)P_{\theta_1'}(x)}{P_{\theta_0}(x)}dx\right)^n$$

□

Example 130 (Planted Clique)**.** Given an undirected graph G on n vertices, we aim to test between

$$H_0: G\sim g\left(n,\frac{1}{2}\right)$$
$$H_1: G\sim g\left(n,\frac{1}{2},k\right)$$

where H_0 means that $\forall i<j\in[n]$, we have that $\Pr((i,j)\in E)=\frac{1}{2}$. H_1 means that there is an unknown $S\subseteq[n]$ with $|S|=k$ (for known k) such that $\Pr((i,j)\in E)=\mathbb{1}_{\{(i,j)\in E\}}+\frac{1}{2}\mathbb{1}_{\{(i,j)\notin E\}}$. We will show that there exists a constant C such that if $k<2\log_2(n)-2\log_2(\log_2(n))+C$, then no test can asymptotically reliably distinguish between H_0 and H_1 meaning that there exists $\epsilon>0$ such that $\inf_a P_0(T=1)+P_1(T=0)\geq\epsilon$.

Proof.

Let P_0 be the distribution of $G\sim g\left(n,\frac{1}{2}\right)$ and P_S be the distribution of G with a clique planted at S where S is a uniformly selected subset of $[n]$ of size k. We note the distribution that uniformly selects a subset of $[n]$ of size k by π. Finally, let $\mathcal{G}$ be the support of all graphs on $[n]$. Then, we have that

$$\begin{aligned}\int_{x\in\mathcal{G}}\frac{P_S(x)P_{S'}(x)}{P_0(x)}dx&=\sum_{x\in\mathcal{G}}\frac{P_S(x)P_{S'}(x)}{P_0(x)}\\&=\sum_{(X_{ij}:(i,j)\in[n])\in\{0,1\}^{\binom{n}{2}}}\frac{\left(\frac{1}{2}\right)^{\binom{n}{2}-\binom{k}{2}}\Pi_{(i,j)\in S}\mathbb{1}_{\{X_{ij}=1\}}\left(\frac{1}{2}\right)^{\binom{n}{2}-\binom{k}{2}}\Pi_{(i,j)\in S'}\mathbb{1}_{\{X_{ij}=1\}}}{\left(\frac{1}{2}\right)^{\binom{n}{2}}}\\&=\frac{\left(\frac{1}{2}\right)^{2\binom{n}{2}-2\binom{k}{2}}\cdot 2^{\binom{n}{2}-2\binom{k}{2}+\binom{|S\cap S'|}{2}}}{\left(\frac{1}{2}\right)^{\binom{n}{2}}},\text{ using the Inclusion-Exclusion principle}\\&=2^{\binom{|S\cap S'|}{2}}\end{aligned}$$

Leveraging Theorem 128, we get that

$$\begin{aligned}\chi^2(\mathbb{E}_\pi[P_S]||P_0) &= \mathbb{E}_{S,S' \overset{iid}{\sim} \pi}\left[2^{\binom{|S\cap S'|}{2}}\right] - 1 \\ &= \sum_{r=0}^{k} 2^{\binom{r}{2}} \Pr(|S\cap S'| = r) - 1 \\ &= \sum_{r=0}^{k} 2^{\binom{r}{2}} \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} - 1, \text{ random select } S', \text{ pick } r \text{ overlapping and } k-r \text{ other elements for } S \\ &= \sum_{r=1}^{k} \left(2^{\binom{r}{2}} - 1\right) \frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}} \\ &= o(1), \text{ when } k < 2\log_n(n) - 2\log_2(\log_2(n)) + C\end{aligned}$$

using Stirling's approximation for some constant $C > 0$. As a result, applying continuity of $\log(\cdot)$ and $\sqrt{\cdot}$, we can leverage the following inequalities for f-divergences ($KL \leq \log(1+\chi^2)$ and $2TV^2 \leq \frac{1}{2}KL$) from Section 3.8, to get that $TV(\mathbb{E}_\pi[P_S], P_0) = o(1)$. In turn, we can apply Theorem 127 to say that

$$r^* \gtrsim \epsilon$$

for some $\epsilon > 0$. The separation condition for the application of the theorem trivially holds as explain in a footnote of Theorem 127. □

Example 131 (Uniformity Testing)**.** Given $X_1, ..., X_n \overset{iid}{\sim} P = (p_1, ..., p_k)$, unknown, we wish to test if

$$\begin{aligned}H_0 &: P = \text{Unif}([k]) =: P_0 \\ H_1 &: TV(P, \text{Unif}([k])) \geq \epsilon\end{aligned}$$

for some $\epsilon > 0$. We wish to show that we cannot reliably distinguish between the two hypotheses if $n = O\left(\frac{\sqrt{k}}{\epsilon^2}\right)$.

Proof.

WLOG, assume that k is even (we're making an asymptotic argument). Restrict the alternative hypothesis to take the form

$$\begin{aligned}H_1' : P_v &= \left(\frac{1-2\epsilon v_1}{k}, \frac{1+2\epsilon v_1}{k}, ..., \frac{1-2\epsilon v_{k/2}}{k}, \frac{1+2\epsilon v_{k/2}}{k}\right) \\ &\text{for } v_i \overset{iid}{\sim} \text{Unif}(\{-1,1\}) \text{ for } i \in [k/2]\end{aligned}$$

Note that $TV(P_v, \text{Unif}([k])) = \epsilon \; \forall v \in \{-1,1\}^{k/2}$. Also note that

$$\begin{aligned}\sum_{x\in[k]} \frac{P_v(x)P_{v'}(x)}{P_0(x)} dx &= \sum_{i=1}^{k/2} \left(\frac{(1-2\epsilon v_i)(1-2\epsilon v_i')}{k} + \frac{(1+2\epsilon v_i)(1+2\epsilon v_i)}{k}\right) \\ &= 1 + \frac{8\epsilon^2}{k} \sum_{i=1}^{k} v_i v_i'\end{aligned}$$

Let $\pi := \text{Unif}(\{-1,1\}^{k/2})$, then we have that

$$\begin{aligned}
\chi^2(\mathbb{E}_\pi[P_v]||P_0) &= \mathbb{E}_{v,v' \overset{iid}{\sim} \pi}\left[\left(\sum_{x\in[k]} \frac{P_v(x)P_{v'}(x)}{P_0(x)}dx\right)^n\right] - 1, \text{ by Corollary 129} \\
&= \mathbb{E}_{v,v' \overset{iid}{\sim} \pi}\left[\left(1 + \frac{8\epsilon^2}{k}\sum_{i=1}^{k/2} v_i v_i'\right)^n\right] - 1, \text{ by above} \\
&\leq \mathbb{E}_{v,v' \overset{iid}{\sim} \pi}\left[\exp\left(\frac{8n\epsilon^2}{k}\sum_{i=1}^{k/2} v_i v_i'\right)\right] - 1, \text{ since } 1 + x \geq \exp(x)\ \forall x \in \mathbb{R} \\
&\leq \exp\left(\frac{1}{2}\left(\frac{8n\epsilon^2}{k}\right)^2 \frac{k}{2}\right) - 1, \sum_{i=1}^{k/2} v_i v_i' \text{ is } k/2\text{-sub gaussian as in Definition 32} \\
&= \exp\left(\frac{16n^2\epsilon^4}{k}\right) - 1 \\
&= o(1), \text{ if } n = O\left(\frac{\sqrt{k}}{\epsilon^2}\right)
\end{aligned}$$

As a result, applying continuity of $\log(\cdot)$ and $\sqrt{\cdot}$, we can leverage the following inequalities for f-divergences ($KL \leq \log(1+\chi^2)$ and $2TV^2 \leq \frac{1}{2}KL$) from Section 3.8, to get that $TV(\mathbb{E}_\pi[P_v], P_0) = o(1)$ in this case. In turn, we can apply Theorem 127 to say that

$$r^* \gtrsim \epsilon$$

for some $\epsilon > 0$. The separation condition for the application of the theorem trivially holds as explain in a footnote of Theorem 127. As a final remark, note that $H_1' \subseteq H_1$ and if we can't reliably distinguish between H_0 and H_1' as we've shown, we also can't reliably distinguish between H_0 and H_1. □

Lemma 132 (Hoeffding's Lemma 2). Let $C = \{c_1, ..., c_N\} \subseteq \mathbb{R}$ be a fixed population and

- $X_1, ..., X_n : n$ draws from C with replacement. Call this distribution of n draws P_n.
- $X_1^*, ..., X_n^* : n$ draws from C without replacement. Call this distribution of n draws P_n^*.

Then, for any convex function $f : \mathbb{R} \to \mathbb{R}$

$$\mathbb{E}_{P_n}\left[f\left(\sum_{i=1}^n X_i\right)\right] \leq \mathbb{E}_{P_n^*}\left[f\left(\sum_{i=1}^n X_i^*\right)\right]$$

Proof.

For a fixed set of draws without replacement, define $\mathcal{A} := \{X_1^*, .., X_n^*\}$ and $P_n \mid \mathcal{A}$ to be P_n restricted to the elements in $\mathcal{A}$. We have that

$$\begin{aligned}
\sum_{i=1}^{n} X_i^* &= \mathbb{E}_{P_n|\mathcal{A}}\left[\sum_{i=1}^{n} X_i\right] \\
\implies f\left(\sum_{i=1}^{n} X_i^*\right) &= f\left(\mathbb{E}_{P_n|\mathcal{A}}\left[\sum_{i=1}^{n} X_i\right]\right) \\
&\leq \mathbb{E}_{P_n|\mathcal{A}}\left[f\left(\sum_{i=1}^{n} X_i\right)\right], \text{ by Jensen's inequality} \\
\implies \mathbb{E}_{P_n^*}\left[f\left(\sum_{i=1}^{n} X_i^*\right)\right] &\leq \mathbb{E}_{P_n^*}\left[\mathbb{E}_{P_n|\mathcal{A}}\left[f\left(\sum_{i=1}^{n} X_i\right)\right]\right] \\
&= \mathbb{E}_{P_n}\left[f\left(\sum_{i=1}^{n} X_i\right)\right]
\end{aligned}$$

□

Example 133 (Linear Functional of Sparse Parameters)**.** Suppose that $X \sim \mathcal{N}(\mu, I_d)$ with $||\mu||_0 \leq s$ (s is known) where $||\mu||_0 := \sum_{i=1}^{\dim(\mu)} \mathbb{1}_{\{\mu_i \neq 0\}}$. We wish to show that

$$\begin{aligned}
\inf_{T} \sup_{||\mu||_0 \leq s} \mathbb{E}_{X\sim\mathcal{N}(\mu,I_d)}\left[\left(T(X) - \sum_{i=1}^{d}\mu_i\right)^2\right] &\asymp s^2 \log\left(1 + \frac{d}{s^2}\right) \\
&\asymp \begin{cases} s^2\log(d) & \text{if } s << \sqrt{d} \\ d & \text{if } d \geq s \geq \sqrt{d} \end{cases}
\end{aligned}$$

If we cannot reliably estimate the sum of the parameters, then we also can't reliably estimate the parameters themselves. Mathematically, this result can be shown using Cauchy-Schwarz.

Proof.

[Lower Bound]:

Define the following two hypotheses:

$$\begin{aligned}
&H_0 : \mu = 0 \\
&H_1 : \mu = \rho 1_S,\ S \sim \text{Unif}\left(\binom{[d]}{S}\right)
\end{aligned}$$

where $1_S \in \mathbb{R}^d$ and has $(1_S)_i := \mathbb{1}_{\{i\in S\}}$ and ρ is currently unspecified. Define $\pi := \text{Unif}\left(\binom{[d]}{S}\right)$. Finally, call $P_0 := \mathcal{N}(0, I_d)$ and $P_S = \mathcal{N}(\rho 1_S, I_d)$. We have that

$$\begin{aligned}
\int_{x\in\mathbb{R}^d} \frac{P_S(x)P_{S'}(x)}{P_0(x)}dx &= \int_{x\in\mathbb{R}^d} \frac{\phi^d(x-\rho 1_S)\phi^d(x-\rho 1_{S'})}{\phi^d(x)}dx, \text{ where } \phi^d(\cdot) \text{ is the pdf of } \mathcal{N}(0, I_d) \\
&= \exp\left(\rho^2 < 1_S, 1_{S'} >\right) \\
&= \exp\left(\rho^2 |S \cap S'|\right)
\end{aligned}$$

Then, by Theorem 128, we have that

$$\begin{aligned}\chi^2(\mathbb{E}_\pi[P_S]||P_0)+1 &= \mathbb{E}_{S,S'\overset{iid}{\sim}\pi}\left[\int_{x\in\mathbb{R}^d}\frac{P_S(x)P_{S'}(x)}{P_0(x)}dx\right]\\ &= \mathbb{E}_{S,S'\overset{iid}{\sim}\pi}\left[\exp\left(\rho^2|S\cap S'|\right)\right]\\ &\le \mathbb{E}_{Z\sim\text{Bin}(s,s/d)}\left[\exp\left(\rho^2 Z\right)\right],\text{ by Lemma 132}\\ &= \left(1-\frac{s}{d}+\frac{s}{d}\exp(\rho^2)\right)^s,\text{ using the MGF of Bin}(n,p)\\ &\le e,\text{ when }\rho=\sqrt{\log\left(1+\frac{d}{s^2}\right)}\end{aligned}$$

since $x\mapsto\left(1+\frac{1}{x}\right)^x$ is increasing on $x\in[0,\infty)$ and its limit as $x\to\infty$ is e. Next, applying the inequalities between f-divergences in Section 3.8, we get that

$$\begin{aligned}TV(P_0,\mathbb{E}_\pi[P_S]) &\le \frac{1}{2}\sqrt{\log(1+\chi^2(\mathbb{E}_\pi[P_S]||P_0))}\\ &\le \frac{1}{2},\text{ by }\chi^2(\mathbb{E}_\pi[P_S]||P_0)+1\le e\end{aligned}$$

Finally, let θ_0 denote $\mu=0$ as in H_0 and $\theta_1\in\Theta_1$ specify which s of the d dimensions of μ are set to ρ. Note that

$$\begin{aligned}\inf_{\theta_1\in\Theta_1}\inf_T\left(T(X)-\sum_{i=1}^d\mu_i(\theta_0)\right)^2+\left(T(X)-\sum_{i=1}^d\mu_i(\theta_1)\right)^2 &= (T(X))^2+(T(X)-\rho s)^2\\ &= 2\left(T(X)-\frac{\rho s}{2}\right)^2+\left(\frac{\rho s}{2}\right)^2\\ &\ge \left(\frac{\rho s}{2}\right)^2\end{aligned}$$

Thus, we can pick $\Delta:=\frac{\rho^2 s^2}{4}$ to apply Theorem 127. The theorem then implies that

$$\begin{aligned}\inf_T\sup_{\theta\in\{\theta_0\}\cup\Theta_1}\mathbb{E}_{X\sim\mathcal{N}(\mu(\theta),I_d)}\left[\left(T(X)-\sum_{i=1}^d\mu_i\right)^2\right] &\ge \inf_T\sup_{\theta\in\{\theta_0\}\cup\Theta_1}\mathbb{E}_{X\sim\mathcal{N}(\mu(\theta),I_d)}\left[\left(T(X)-\sum_{i=1}^d\mu_i\right)^2\right]\\ &\ge \frac{\Delta}{2}(1-TV(P_0,\mathbb{E}_\pi[P_S]))\\ &\ge \frac{\rho^2 s^2}{8}\left(\frac{1}{2}\right)\\ &= \Omega\left(s^2\log\left(1+\frac{d}{s^2}\right)\right)\end{aligned}$$

□

8.2 *Mixture versus Mixture*

In this section, we focus on using the Le Cam method with a mixture versus mixture approach in a binary hypothesis setting to present asymptotic lower bounds on risk.

Theorem 134 (Advanced Le Cam: Mixture vs. Mixture)**.** Consider a statistical model $\{P_\theta:\theta\in\Theta\}$. Let $\Theta_0\subseteq\Theta$ and $\Theta_1=\Theta\setminus\Theta_0$. We wish to take an action a based on θ from an observation $X\sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta,a)$.[a] In many cases the action a is a test $T:\mathcal{X}\to\{0,1\}$ and we wish to decide if $\theta=\theta_0$ or

$\theta \in \Theta_1$.[b] Suppose that

$$\inf_{\theta_0, \theta_1 \in \Theta_0 \times \Theta_1} \min_a \left(L(\theta_0, a) + L(\theta_1, a) \right) \geq \Delta$$

then for any probability distributions $\pi_0, \pi_1 \in P(\Theta)$, we have that

$$\begin{aligned} r^* &= \inf_a \sup_{\theta \in \Theta_0 \cup \Theta_1} \mathbb{E}_{X \sim P_\theta} \left[L(\theta, a(X)) \right] \\ &\geq \frac{\Delta}{2} \left(1 - TV(\mathbb{E}_{\theta \sim \pi_0}[P_\theta], \mathbb{E}_{\theta \sim \pi_1}[P_\theta]) - \pi_0(\Theta_0^c) - \pi_1(\Theta_1^c) \right) \end{aligned}$$

where X^c is the complement of the set X.

Proof.

Take any probability distributions $\pi_0, \pi_1 \in P(\Theta)$ and call $\tilde{\pi}_0$ the restriction of π_0 on Θ_0 and $\tilde{\pi}_1$ the restriction of π_1 on Θ_1 – these restrictions are *not* normalized to 1. Lastly, also define $P_{\pi_0} = \mathbb{E}_{\theta \sim \pi_0}[P_\theta]$ and $P_{\pi_1} = \mathbb{E}_{\theta \sim \pi_1}[P_\theta]$. Then, we have that

$$\begin{aligned} \inf_a \sup_{\theta \in \Theta_0 \cup \Theta_1} \mathbb{E}_{X \sim P_\theta} \left[L(\theta, a(X)) \right] &\geq \inf_a \frac{1}{2} \mathbb{E}_{\theta \sim \pi_0} \left[\mathbb{E}_{X \sim P_\theta} \left[L(\theta, a(X)) \right] \right] + \frac{1}{2} \mathbb{E}_{\theta \sim \pi_1} \left[\mathbb{E}_{X \sim P_\theta} \left[L(\theta, a(X)) \right] \right] \\ &\geq \frac{\Delta}{2} \int_{x \in \mathcal{X}} \min(p_{\tilde{\pi}_0}(x), p_{\tilde{\pi}_1}(x)) dx, \text{ by the separation condition and } L(\cdot, \cdot) \geq 0 \\ &= \frac{\Delta}{2} \int_{x \in \mathcal{X}} \frac{1}{2} \left[p_{\tilde{\pi}_0}(x) + p_{\tilde{\pi}_1}(x) - |p_{\tilde{\pi}_0}(x) - p_{\tilde{\pi}_1}(x)| \right] dx \\ &= \frac{\Delta}{2} \left[\frac{1}{2} (\pi_0(\Theta_0) + \pi_1(\Theta_1)) - \frac{1}{2} \int_{x \in \mathcal{X}} |p_{\tilde{\pi}_0}(x) - p_{\tilde{\pi}_1}(x)| dx \right] \\ &= \frac{\Delta}{2} \left[1 - \frac{1}{2} \pi_0(\Theta_0^c) - \frac{1}{2} \pi_1(\Theta_1^c) - \frac{1}{2} \int_{x \in \mathcal{X}} |p_{\tilde{\pi}_0}(x) - p_{\tilde{\pi}_1}(x)| dx \right] \end{aligned}$$

Now, focusing just on $\frac{1}{2} \int_{x \in \mathcal{X}} |p_{\tilde{\pi}_0}(x) - p_{\tilde{\pi}_1}(x)| dx$, by the triangle inequality

$$\begin{aligned} \frac{1}{2} \int_{x \in \mathcal{X}} |p_{\tilde{\pi}_0}(x) - p_{\tilde{\pi}_1}(x)| dx &\leq \frac{1}{2} \int_{x \in \mathcal{X}} |p_{\tilde{\pi}_0}(x) - p_{\pi_0}(x)| dx + \frac{1}{2} \int_{x \in \mathcal{X}} |p_{\pi_0}(x) - p_{\pi_1}(x)| dx \\ &\quad + \frac{1}{2} \int_{x \in \mathcal{X}} |p_{\tilde{\pi}_1}(x) - p_{\pi_1}(x)| dx \\ &= TV(P_{\pi_0}, P_{\pi_1}) + \frac{1}{2} \pi_0(\Theta_0^c) + \frac{1}{2} \pi_1(\Theta_1^c) \end{aligned}$$

As a result, combining with the above, we get that

$$r^* \geq \frac{\Delta}{2} \left(1 - TV(\mathbb{E}_{\theta \sim \pi_0}[P_\theta], \mathbb{E}_{\theta \sim \pi_1}[P_\theta]) - \pi_0(\Theta_0^c) - \pi_1(\Theta_1^c) \right)$$

□

[a] Assume that the models admit a density with respect to a measure on $\mathcal{X}$ though the measure-theoretic details are skirted here.

[b] Note that the separation condition here trivially holds. $L(\theta, T(x)) = \mathbb{1}_{\{T(x) \neq \mathbb{1}_{\{\theta \neq \theta_0\}}\}}$. We have that for any (deterministic) test T and any sample x, $L(\theta_0, T(x)) + L(\theta_1, T(x)) = 1$. Note that we can also let T be randomized and arrive at the same conclusion.

Lemma 135 (Orthogonal Functions and Polynomials)**.** Suppose that $\{P_\theta : \theta \in [\theta_0 - \epsilon, \theta_0 + \epsilon]\}$ is a 1D statistical model with likelihood ratio expansion

$$\frac{P_{\theta_0+u}(x)}{P_{\theta_0}(x)} = \sum_{m=0}^{\infty} p_m(x;\theta_0)\frac{u^m}{m!}, \text{ for } |u| \leq \epsilon$$

If $\int_{x\in\mathcal{X}} \frac{P_{\theta_0+u}(x)P_{\theta_1+v}(x)}{P_{\theta_0}(x)}dx$ depends only on $(\theta_0, u\cdot v)$, then[a]

$$\mathbb{E}_{X\sim P_{\theta_0}}\left[p_m(X;\theta_0)p_n(X;\theta_0)\right] = 0, \; \forall m \neq n$$

Proof.

We have that

$$\begin{aligned}\int_{x\in\mathcal{X}} \frac{P_{\theta_0+u}(x)P_{\theta_1+v}(x)}{P_{\theta_0}(x)}dx &= \mathbb{E}_{X\sim P_{\theta_0}}\left[\left(\sum_{m=0}^{\infty} p_m(X;\theta_0)\frac{u^m}{m!}\right)\left(\sum_{n=0}^{\infty} p_n(X;\theta_0)\frac{v^n}{n!}\right)\right] \\ &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\mathbb{E}_{X\sim P_{\theta_0}}\left[p_m(X;\theta_0)p_n(X;\theta_0)\right]\frac{u^m v^n}{m!n!}\end{aligned}$$

Since this quantity depends on (u, v) through $u\cdot v$ by assumption, it must be that $\mathbb{E}_{X\sim P_{\theta_0}}\left[p_m(X;\theta_0)p_n(X;\theta_0)\right] = 0$ for $m \neq n$. □

[a] Technically, for this conclusion, I think we must assume that each model in $\{P_\theta : \theta \in [\theta_0 - \epsilon, \theta_0 + \epsilon]\}$ has full support on $\mathcal{X}$.

Example 136 (Gaussian Hermite Polynomials). For $P_\theta = \mathcal{N}(\theta, 1)$, then $\int_{x\in\mathbb{R}} \frac{P_u(x)P_v(x)}{P_0(x)}dx = \exp(u\cdot v)$, the corresponding $p_m(x;\theta_0 = 0)$ is called a *Hermite polynomial* $H_m(x)$ with

$$\mathbb{E}_{X\sim\mathcal{N}(0,1)}[H_m(X)H_n(X)] = n!\mathbb{1}_{\{m=n\}}$$

Example 137 (Poisson-Charlier Polynomials). For $P_\theta = \text{Poisson}(\theta)$, then for some fixed $\lambda > 0$, we have that

$$\begin{aligned}\int_{x\in\mathbb{R}} \frac{P_{\lambda+u}(x)P_{\lambda+v}(x)}{P_\lambda(x)}dx &= \sum_{k=0}^{\infty}\frac{\exp(-\lambda-u-v)}{k!}\left(\frac{(\lambda+u)(\lambda+v)}{\lambda}\right)^k \\ &= \exp\left(\frac{u\cdot v}{\lambda}\right)\end{aligned}$$

The corresponding $p_m(x;\theta_0 = \lambda)$ is called a *Poisson-Charlier polynomial* $c_m(x;\lambda)$ with

$$\mathbb{E}_{X\sim\text{Poisson}(\lambda)}[c_m(X;\lambda)c_n(X;\lambda)] = \frac{n!}{\lambda^n}\mathbb{1}_{\{m=n\}}$$

Theorem 138 (Upper-Bounding TV and χ^2 of Univariate Gaussian Mixtures). For $\mu \in \mathbb{R}$ and random variables U, V with support of $\mathbb{R}$, we have that

$$TV(\mathbb{E}_{U\sim P_U}[\mathcal{N}(\mu+U,1)], \mathbb{E}_{V\sim P_V}[\mathcal{N}(\mu+V,1)]) \leq \frac{1}{2}\left(\sum_{m=0}^{\infty}\frac{(\mathbb{E}_{U\sim P_U}[U^m] - \mathbb{E}_{V\sim P_V}[V^m])^2}{m!}\right)^{1/2}$$

If in addition $\mathbb{E}_{V\sim P_V}[V] = 0$ and $\mathbb{E}_{V\sim P_V}[V^2] \leq M^2$, then we have that

$$\chi^2\left(\mathbb{E}_{U\sim P_U}[\mathcal{N}(\mu+U,1)]||\mathbb{E}_{V\sim P_V}[\mathcal{N}(\mu+V,1)]\right) \leq \frac{\exp(M^2)}{2}\sum_{m=0}^{\infty}\frac{(\mathbb{E}_{U\sim P_U}[U^m]-\mathbb{E}_{V\sim P_V}[V^m])^2}{m!}$$

Proof.

WLOG, assume that $\mu = 0$ and let $\Delta_m := \mathbb{E}_{U\sim P_U}[U^m] - \mathbb{E}_{V\sim P_V}[V^m]$. Then, letting $\phi(\cdot)$ be the PDF of $\mathcal{N}(0,1)$ and using the definition of *Hermite polynomials* in Definition 136, we have that

$$\begin{aligned}
TV\left(\mathbb{E}_{U\sim P_U}[\mathcal{N}(U,1)], \mathbb{E}_{V\sim P_V}[\mathcal{N}(V,1)]\right) &= \frac{1}{2}\int_{\mathbb{R}} \left|\mathbb{E}_{U\sim P_U}[\phi(x-U)] - \mathbb{E}_{V\sim P_V}[\phi(x-V)]\right| dx \\
&= \frac{1}{2}\int_{\mathbb{R}} \phi(x)\Bigg|\mathbb{E}_{U\sim P_U}\left[\sum_{m=0}^{\infty} H_m(x)\frac{U^m}{m!}\right] \\
&\quad - \mathbb{E}_{V\sim P_V}\left[\sum_{m=0}^{\infty} H_m(x)\frac{V^m}{m!}\right]\Bigg| dx \\
&= \frac{1}{2}\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\left|\sum_{m=0}^{\infty} H_m(X)\frac{\Delta_m}{m!}\right|\right] \\
&\leq \frac{1}{2}\left(\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\left(\sum_{m=0}^{\infty} H_m(X)\frac{\Delta_m}{m!}\right)^2\right]\right)^{1/2}, \text{ by Cauchy-Schwarz} \\
&= \frac{1}{2}\left(\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!}\right)^{1/2}, \text{ by Example 136}
\end{aligned}$$

Next, we wish to show the χ^2 upper bound with the additional assumptions that $\mathbb{E}_{V\sim P_V}[V] = 0$ and $\mathbb{E}_{V\sim P_V}[V^2] \leq M^2$. As a first observation in this direction, note that for any $x \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}_{V\sim P_V}[\phi(x-V)] &= \phi(x)\mathbb{E}_{V\sim P_V}\left[\exp\left(Vx - \frac{V^2}{2}\right)\right] \\
&\geq \phi(x)\exp\left(\mathbb{E}_{V\sim P_V}\left[Vx - \frac{V^2}{2}\right]\right), \text{ by Jensen's Inequaltiy} \\
&\geq \phi(x)\exp\left(-\frac{M^2}{2}\right)
\end{aligned}$$

Then, we have that

$$\begin{aligned}
\chi^2\left(\mathbb{E}_{U\sim P_U}\left[\mathcal{N}(U,1)\right]||\mathbb{E}_{V\sim P_V}\left[\mathcal{N}(V,1)\right]\right) &= \int_{\mathbb{R}} \frac{\left(\mathbb{E}_{U\sim P_U}\left[\phi(x-U)\right]-\mathbb{E}_{V\sim P_V}\left[\phi(x-V)\right]\right)^2}{\mathbb{E}_{V\sim P_V}\left[\phi(x-V)\right]}dx \\
&\leq \int_{\mathbb{R}} \frac{\left(\mathbb{E}_{U\sim P_U}\left[\phi(x-U)\right]-\mathbb{E}_{V\sim P_V}\left[\phi(x-V)\right]\right)^2}{\phi(x)\exp\left(-\frac{M^2}{2}\right)}dx \\
&= \exp\left(\frac{M^2}{2}\right)\int_{x\in\mathcal{X}}\left(\mathbb{E}_{U\sim P_U}\left[\sum_{m=0}^{\infty} H_m(x)\frac{U^m}{m!}\right]\right. \\
&\quad \left. -\mathbb{E}_{V\sim P_V}\left[\sum_{m=0}^{\infty} H_m(x)\frac{V^m}{m!}\right]\right)^2\phi(x)dx \\
&= \exp\left(\frac{M^2}{2}\right)\mathbb{E}_{X\sim\mathcal{N}(0,1)}\left[\left(\sum_{m=0}^{\infty} H_m(X)\frac{\Delta_m}{m!}\right)^2\right] \\
&= \exp\left(\frac{M^2}{2}\right)\left(\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!}\right), \text{ by Example 136}
\end{aligned}$$

□

Theorem 139 (Upper-Bounding TV and χ^2 of Univariate Poisson Mixtures)**.** For $\lambda > 0$ and random variables U, V supported on $[-\lambda, \infty)$, defining $\Delta_m := \mathbb{E}_{U\sim P_U}[U^m] - \mathbb{E}_{V\sim P_V}[V^m]$, we have that

$$TV\left(\mathbb{E}_{U\sim P_U}\left[\text{Poisson}(\lambda+U)\right], \mathbb{E}_{V\sim P_V}\left[\text{Poisson}(\lambda+V)\right]\right) \leq \frac{1}{2}\left(\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!\lambda^m}\right)^{1/2}$$

If in addition $\mathbb{E}_{V\sim P_V}[V] = 0$ and $\max_{v\in\text{supp}(V)}|v| \leq M$, then

$$\chi^2\left(\mathbb{E}_{U\sim P_U}\left[\text{Poisson}(\lambda+U)\right]||\mathbb{E}_{V\sim P_V}\left[\text{Poisson}(\lambda+V)\right]\right) \leq \exp(M)\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!\lambda^m}$$

Proof.

WLOG, let $\lambda = 1$. Also, let $p_\lambda(\cdot)$ denote the PDF of Poisson(λ). Finally, for brevity, I will abbreviate "Poisson" by

“Poi”. We have that

$$\begin{aligned}
TV\left(\mathbb{E}_{U\sim P_U}\left[\text{Poi}(\lambda+U)\right],\mathbb{E}_{V\sim P_V}\left[\text{Poi}(\lambda+V)\right]\right) &= \frac{1}{2}\int_{[0,\infty)} \left|\mathbb{E}_{U\sim P_U}\left[p_{1+U}(x)\right]-\mathbb{E}_{V\sim P_V}\left[p_{1+V}(x)\right]\right| dx \\
&= \frac{1}{2}\int_{[0,\infty)} p_1(x)\Bigg|\mathbb{E}_{U\sim P_U}\left[\sum_{m=0}^{\infty} c_m(x;\lambda=1)\frac{U^m}{m!}\right] \\
&\quad - \mathbb{E}_{V\sim P_V}\left[\sum_{m=0}^{\infty} c_m(x;\lambda=1)\frac{V^m}{m!}\right]\Bigg| dx \\
&= \frac{1}{2}\mathbb{E}_{X\sim\text{Poi}(\lambda=1)}\left[\left|\sum_{m=0}^{\infty} c_m(x;\lambda=1)\frac{\Delta_m}{m!}\right|\right] \\
&\le \frac{1}{2}\left(\mathbb{E}_{X\sim\text{Poi}(\lambda=1)}\left[\left(\sum_{m=0}^{\infty} c_m(x;\lambda=1)\frac{\Delta_m}{m!}\right)^2\right]\right)^{1/2}, \\
&\quad \text{by Cauchy-Schwarz} \\
&= \frac{1}{2}\left(\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!}\right)^{1/2}, \text{ by Example 137}
\end{aligned}$$

Next, we wish to show the χ^2 upper bound with the additional assumptions that $\mathbb{E}_{V\sim P_V}\left[V\right]=0$ and $\max_{v\in\text{supp}(V)}|v| \le M$. As a first observation in this direction, note that for any $x\in\mathbb{N}\cup\{0\}$,

$$\begin{aligned}
\mathbb{E}_{V\sim P_V}\left[p_{1+V}(x)\right] &= \mathbb{E}_{V\sim P_V}\left[\frac{(1+V)^x\exp(-(1+V))}{x!}\right] \\
&\ge \frac{\exp(-1)}{x!}\mathbb{E}_{V\sim P_V}\left[(1+V)^x\exp(-V)\right] \\
&\ge \frac{\exp(-1)}{x!}\exp(-M)\mathbb{E}_{V\sim P_V}\left[(1+V)^x\right] \\
&\ge \frac{\exp(-1)}{x!}\exp(-M)\left(\mathbb{E}_{V\sim P_V}\left[1+V\right]\right)^x, \text{ by convexity of } t\mapsto t^x \text{ for } x\in\mathbb{N}\cup\{0\} \\
&= p_1(x)\exp(-M)
\end{aligned}$$

Then, we have that

$$\begin{aligned}
\chi^2\left(\mathbb{E}_{U\sim P_U}\left[\text{Poi}(\lambda+U)\right]||\mathbb{E}_{V\sim P_V}\left[\text{Poi}(\lambda+V)\right]\right) &= \frac{1}{2}\int_{[0,\infty)} \frac{\left(\mathbb{E}_{U\sim P_U}\left[p_{1+U}(x)\right]-\mathbb{E}_{V\sim P_V}\left[p_{1+V}(x)\right]\right)^2}{\mathbb{E}_{V\sim P_V}\left[p_{1+v}(x)\right]}dx \\
&\le \int_{[0,\infty)} \frac{\left(\mathbb{E}_{U\sim P_U}\left[p_{1+U}(x)\right]-\mathbb{E}_{V\sim P_V}\left[p_{1+V}(x)\right]\right)^2}{p_1(x)\exp(-M)}dx \\
&= \exp(M)\int_{[0,\infty)}\Bigg(\mathbb{E}_{U\sim P_U}\left[\sum_{m=0}^{\infty} c_m(x;\lambda=1)\frac{U^m}{m!}\right] \\
&\quad - \mathbb{E}_{V\sim P_V}\left[\sum_{m=0}^{\infty} c_m(x;\lambda=1)\frac{V^m}{m!}\right]\Bigg)^2 p_1(x)dx \\
&= \exp(M)\mathbb{E}_{X\sim\text{Poi}(\lambda=1)}\left[\left(\sum_{m=0}^{\infty} c_m(X;\lambda=1)\frac{\Delta_m}{m!}\right)^2\right] \\
&= \exp(M)\left(\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!}\right), \text{ by Example 137}
\end{aligned}$$

□

Example 140 (Generalized Uniformity Testing)**.** Given $X_1, ..., X_n \overset{iid}{\sim} P = (p_1, ..., p_k)$, unknown, we wish to test if

$$\begin{aligned} H_0 &: P = \text{Unif}(S) \text{ for some } S \subseteq [k] \\ H_1 &: \min_{S \subseteq [k]} TV(P, \text{Unif}(S)) \geq \frac{\epsilon}{2} \end{aligned}$$

for some $\epsilon > 0$. We wish to show that we cannot reliably distinguish between the two hypotheses if $n = O\left(\frac{\sqrt{k}}{\epsilon^2} + \frac{k^{2/3}}{\epsilon^{4/3}}\right)$.

Proof.

The proof of $n = O\left(\frac{\sqrt{k}}{\epsilon^2}\right)$ follows directly from Example 131 so we focus on the second term. Take any small $\epsilon > 0$. Assume Poissonization where the observations are $\otimes_{i=1}^{k} \text{Poi}(np_i)$.[a] We construct two product priors where under H_0, $p_1, ..., p_k \overset{iid}{\sim} \text{Law}(U)$ and under H_1, $p_1, ..., p_k \overset{iid}{\sim} \text{Law}(V)$ where

$$U = \begin{cases} 0 & \text{w.p. } \frac{\epsilon^2}{1+\epsilon^2} \\ \frac{1+\epsilon^2}{k} & \text{w.p. } \frac{1}{1+\epsilon^2} \end{cases}$$

$$V = \begin{cases} \frac{1-\epsilon}{k} & \text{w.p. } \frac{1}{2} \\ \frac{1+\epsilon}{k} & \text{w.p. } \frac{1}{2} \end{cases}$$

Next, we make some observations and comments:

- Under H_0, $p_i \in \{0, \frac{1+\epsilon^2}{k}\}$ so that $(p_1, ..., p_k)$ is generalized uniform.[b]

- Under H_1, $(p_1, ..., p_k)$ is $\Omega(\epsilon)$-far from a generalized uniform with high probability.[c]

- $\mathbb{E}[U] = \mathbb{E}[V] = \frac{1}{k}$ so that under both H_0 and H_1, $(p_1, ..., p_k)$ is a probability mass function (pmf) in expectation.[d]

- $\mathbb{E}[U^2] = \mathbb{E}[V^2] = \frac{1+\epsilon^2}{k^2}$.

- We have that for $m \geq 3$, by the triangle inequality,

$$\begin{aligned}\left|\mathbb{E}\left[\left(U-\frac{1}{k}\right)^m\right]-\mathbb{E}\left[\left(V-\frac{1}{k}\right)^m\right]\right| &\leq \left|\mathbb{E}\left[\left(U-\frac{1}{k}\right)^m\right]\right|+\left|\mathbb{E}\left[\left(V-\frac{1}{k}\right)^m\right]\right| \\ &\leq \left|\frac{\epsilon^2}{1+\epsilon^2}\left(-\frac{1}{k}\right)^m+\frac{1}{1+\epsilon^2}\left(\frac{1+\epsilon^2}{k}-\frac{1}{k}\right)^m\right| \\ &\quad+\left|\frac{1}{2}\left(\frac{1-\epsilon}{k}-\frac{1}{k}\right)^m+\frac{1}{2}\left(\frac{1+\epsilon}{k}-\frac{1}{k}\right)^m\right| \\ &= \left|\frac{\epsilon^2}{1+\epsilon^2}\left(-\frac{1}{k}\right)^m+\frac{1}{1+\epsilon^2}\left(\frac{\epsilon^2}{k}\right)^m\right| \\ &\quad+\left|\frac{1}{2}\left(\frac{-\epsilon}{k}\right)^m+\frac{1}{2}\left(\frac{\epsilon}{k}\right)^m\right| \\ &\leq \left|\frac{\epsilon^2}{1+\epsilon^2}\left(\frac{1}{k}\right)^m+\frac{1}{1+\epsilon^2}\left(\frac{\epsilon^2}{k}\right)^m\right| \\ &\quad+\left|\frac{1}{2}\left(\frac{\epsilon}{k}\right)^m+\frac{1}{2}\left(\frac{\epsilon}{k}\right)^m\right| \\ &\leq \frac{\epsilon^2}{(1+\epsilon^2)k^m}+\frac{\epsilon^2}{k^m} \\ &\leq \frac{2\epsilon^2}{k^m}\end{aligned}$$

Now, we can apply Theorem 139. We see that $\max_{v\in\text{supp}(V)}\left|n\left(V-\frac{1}{k}\right)\right| = \frac{n\epsilon}{k} =: M$. As in the theorem, defining $\Delta_m := \mathbb{E}\left[\left(U-\frac{1}{k}\right)^m\right]-\mathbb{E}\left[\left(V-\frac{1}{k}\right)^m\right]$, we get that

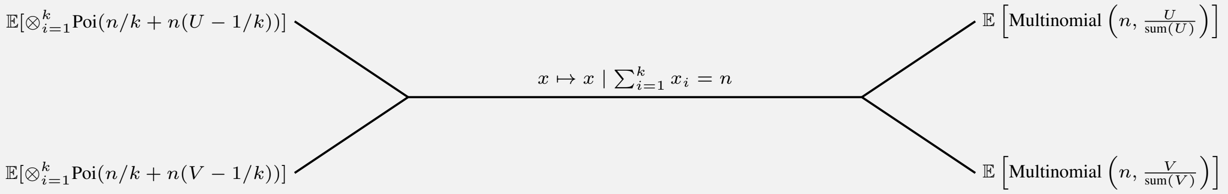
$$\begin{aligned}\chi^2\left(\mathbb{E}\left[\text{Poi}\left(n/k+n\left(U-1/k\right)\right)\right]||\mathbb{E}\left[\text{Poi}\left(n/k+n\left(V-1/k\right)\right)\right]\right) &\leq \exp(M)\sum_{m=0}^{\infty}\frac{\Delta_m^2}{m!} \\ &\leq \exp\left(\frac{n\epsilon}{k}\right)\sum_{m=3}^{\infty}\frac{4\epsilon^4 n^m}{m!k^m} \\ &\leq e\frac{n^3\epsilon^4}{k^3},\ \text{after some algebra}\end{aligned}$$

where the second line uses the above result and the fact that $\mathbb{E}[U^m] = \mathbb{E}[V^m]$ for $m \in \{0,1,2\}$. The third line follows from the fact that we're assuming $O\left(\frac{k^{2/3}}{\epsilon^{4/3}}\right)$ is tight so that $\frac{k^{2/3}}{\epsilon^{4/3}} \geq \frac{\sqrt{k}}{\epsilon^2} \iff k \geq \frac{1}{\epsilon^4} \iff k^{1/3} \geq \frac{1}{\epsilon^{4/3}} \iff k \geq \frac{k^{2/3}}{\epsilon^{4/3}}$ so $n \leq \frac{k^{2/3}}{\epsilon^{4/3}} \implies n \leq k \implies \frac{n}{k} \leq 1$. As a result, the first term (ie., $m = 3$ term) in the sum dominates and we can bound the sum as a multiple of that term using some algebra.

Next, applying the tensorization of χ^2-divergence in Section 3.3, we get that

$$\begin{aligned}&\chi^2(\mathbb{E}[\otimes_{i=1}^k\text{Poi}(n/k+n(U-1/k))]||\mathbb{E}[\otimes_{i=1}^k\text{Poi}(n/k+n(V-1/k))])+1 \\ &= \Pi_{i=1}^k\left[\chi^2(\mathbb{E}[\text{Poi}(n/k+n(U-1/k))]||\mathbb{E}[\text{Poi}(n/k+n(V-1/k))])+1\right] \\ &\leq \left(1+e\frac{n^3\epsilon^4}{k^3}\right)^k \\ &\leq \exp\left(e\frac{n^3\epsilon^4}{k^2}\right),\ \text{since } 1+x\leq\exp(x) \text{ for } x\in\mathbb{R} \\ &= \exp(-1),\ \text{if } n=\frac{1}{e^{2/3}}\frac{k^{2/3}}{\epsilon^{4/3}}\end{aligned}$$

Finally, we wish to move in the direction of applying Theorem 134 to establish that we cannot reliably distinguish between (a subset of) H_0 and (a subset of) H_1. The separation condition of the theorem holds trivially with $\Delta := 1$ as explained in a footnote of the theorem. Next, we have considered a Poissonized version of the distribution when in fact we wish to consider the multinomial version. To that end, consider the following channel:[e]


$\mathbb{E}[\otimes_{i=1}^k \text{Poi}(n/k + n(U-1/k))]$
$\mathbb{E}[\otimes_{i=1}^k \text{Poi}(n/k + n(V-1/k))]$
$x \mapsto x \mid \sum_{i=1}^k x_i = n$
$\mathbb{E}\left[\text{Multinomial}\left(n, \frac{U}{\text{sum}(U)}\right)\right]$
$\mathbb{E}\left[\text{Multinomial}\left(n, \frac{V}{\text{sum}(V)}\right)\right]$

Applying Theorem 37 (ie., Data Processing Inequality for f-Divergences), we get that

$$\begin{aligned}
&\chi^2(\mathbb{E}[\otimes_{i=1}^k \text{Poi}(n/k + n(U-1/k))] || \mathbb{E}[\otimes_{i=1}^k \text{Poi}(n/k + n(V-1/k))]) \\
&\geq \chi^2\left(\mathbb{E}\left[\text{Multinomial}\left(n, \frac{U}{\text{sum}(U)}\right)\right] \middle\| \mathbb{E}\left[\text{Multinomial}\left(n, \frac{V}{\text{sum}(V)}\right)\right]\right) \\
&\implies \chi^2\left(\mathbb{E}\left[\text{Multinomial}\left(n, \frac{U}{\text{sum}(U)}\right)\right] \middle\| \mathbb{E}\left[\text{Multinomial}\left(n, \frac{V}{\text{sum}(V)}\right)\right]\right) = \exp(-1), \text{ if } n = \frac{1}{e^{2/3}} \frac{k^{2/3}}{\epsilon^{4/3}}
\end{aligned}$$

Finally, from the inequalities between f-divergences in Section 3.8, we have that

$$\begin{aligned}
&TV \leq \frac{1}{2}\sqrt{\log(1+\chi^2)} \\
&\implies TV\left(\mathbb{E}\left[\text{Multinomial}\left(n, \frac{U}{\text{sum}(U)}\right)\right], \mathbb{E}\left[\text{Multinomial}\left(n, \frac{V}{\text{sum}(V)}\right)\right]\right) \leq 0.3
\end{aligned}$$

So that applying Theorem 134, we have that[f]

$$\begin{aligned}
r^* &\geq \frac{\Delta}{2}\left(1 - TV\left(\mathbb{E}\left[\text{Multinomial}\left(n, \frac{U}{\text{sum}(U)}\right)\right], \mathbb{E}\left[\text{Multinomial}\left(n, \frac{V}{\text{sum}(V)}\right)\right]\right) - o(1)\right) \\
&\geq \frac{7}{20}
\end{aligned}$$

□

[a] By Poissonization, we mean that we construct a new distribution so that the previously multinomial counts for $i \in \{1, ..., k\}$ are independently determined as Poisson random variables with the same expected mean.

[b] A generalized uniform random variable is one where the elements with non-zero probability have equal support.

[c] By $\Omega(\epsilon)$-far away, we mean using the TV distance metric. With high-probability, all k chosen probabilities will not agree.

[d] Additional arguments are needed to ensure that it suffices to consider "approximate pmfs" but those are skirted here

[e] We use the well-known fact that independent Poisson random variables conditioned on their sum have a multinomial distribution.

[f] The $o(1)$ accounts for the probability that H_1 produces a V that is uniform, which is vanishingly small as $k \to \infty$. H_0 never produces a U that's not generalized uniform.

Lemma 141 (Non-Negative Random Variables Distribution Law Equality)**.** Let U be a non-negative random variable supported on $\{0, x_1, ..., x_{k-1}\}$ for some $k \in \mathbb{N}$. Let V be another random variable supported on $[0, \infty)$. Suppose that

$$\mathbb{E}_{X \sim P_U}[X^m] = \mathbb{E}_{X \sim P_V}[X^m] \; \forall m \in \{0, ..., 2k-1\}$$

Then, we have that $U =_d V$.

Proof.

We have that

$$\begin{aligned} 0 &= \mathbb{E}_{X\sim P_U}\left[X(X-x_1)^2 \cdot ... \cdot (X-x_{k-1})^2\right], \text{ since one of terms must equal } 0 \\ &= \mathbb{E}_{X\sim P_V}\left[X(X-x_1)^2 \cdot ... \cdot (X-x_{k-1})^2\right], \text{ since } U, V \text{ agree on first } 2k-1 \text{ moments} \end{aligned}$$

That implies that

$$\begin{aligned} &\text{supp}(V) \subseteq \{0, x_1, ..., x_{k-1}\} \\ &\implies U =_d V \end{aligned}$$

since we can form an invertible Vandermonde matrix (ie., the columns are $[x_i^0, x_i^1, ..., x_i^{2k-1}]'$) of the various $2k$ moments. We can then pick any size k subset of moments (that must include the $m = 0$ moment) to uniquely determines the pmf of V, matching that of U. □

9 Advanced Fano's Method

In this section, we will cover some extensions of Theorem 117 and Theorem 118.

9.1 *Covering and Packing*

We first discuss some mathematical prerequisites for this section. For this section, let (X, d) be a metric space and $A \subseteq X$ be a compact set.

Definition 142 (ϵ-Covering)**.** The set $\{x_1, ..., x_n\} \subseteq X$ is an ϵ-covering (or ϵ-net) of A if

$$\begin{aligned} &A \subseteq \cup_{i=1}^n B(x_i; \epsilon) \\ &\text{with } B(x_i; \epsilon) = \{y \in X : d(x, y) \le \epsilon\} \end{aligned}$$

Definition 143 (ϵ-Packing)**.** The set $\{a_1, ..., a_n\} \subseteq A$ is an ϵ-packing of A if

$$\min_{i \neq j} d(a_i, a_j) > \epsilon$$

Definition 144 (Covering and Packing Numbers)**.** We define the ϵ-covering and ϵ-packing numbers as, respectively,

$$\begin{aligned} N(A, d, \epsilon) &:= \min\{n \in \mathbb{N} : \exists \epsilon\text{-cover of } A \text{ of size } n\} \\ M(A, d, \epsilon) &:= \max\{m \in \mathbb{N} : \exists \epsilon\text{-packing of } A \text{ of size } m\} \end{aligned}$$

Lemma 145 (Relationship between Covering and Packing Numbers)**.** We have that

$$M(A, d, 2\epsilon) \overset{(1)}{\le} N(A, d, \epsilon) \overset{(2)}{\le} M(A, d, \epsilon)$$

Proof.

[(1)]:

Suppose for the sake of contradiction that $M(A, d, 2\epsilon) \ge N(A, d, \epsilon) + 1$. Then by the pigeonhole principle, there exist two points x, x' in a 2ϵ-packing of A belonging to the same ball $B(y; \epsilon)$ for some y in an ϵ-covering of A. That implies that

$$\begin{aligned} d(x,x') &\leq d(x,y) + d(x',y), \text{ by the triangle inequality} \\ &\leq 2\epsilon \end{aligned}$$

which contradicts the 2ϵ-packing. Thus, we have that $M(A,d,2\epsilon) < N(A,d,\epsilon)+1 \implies M(A,d,2\epsilon) \leq N(A,d,\epsilon)$ since the packing and covering numbers take on positive integer values.

[(2)]:

Suppose for the sake of contradiction that $a_1,...,a_m$ is a maximal ϵ-packing of A but it's not an ϵ-covering. Then, $\exists a \in A$ such that $d(a,a_i) > \epsilon\ \forall i \in [m]$. That implies that $\{a_1,...,a_m\} \cup \{a\}$ is a larger ϵ-packing of A, which is a contradiction. This, we have that a maximal ϵ-packing is in fact an ϵ-covering. □

Lemma 146 (Bounding the Covering and Packing Numbers)**.** Let $||\cdot||$ be any norm on $\mathbb{R}^d$ and $B := \{x : ||x|| \leq 1\}$ be the unit ball. Then, we have that for any compact set A and any (small) $\epsilon > 0$,

$$\left(\frac{1}{\epsilon}\right)^d \frac{\text{Vol}(A)}{\text{Vol}(B)} \overset{(1)}{\leq} N(A,||\cdot||,\epsilon) \leq M(A,||\cdot||,\epsilon) \overset{(2)}{\leq} \left(\frac{2}{\epsilon}\right)^d \frac{\text{Vol}(A+\frac{\epsilon}{2}B)}{\text{Vol}(B)}$$

where the center inequality holds by Lemma 145, taking the metric induced by the norm (ie., $d(x,y) := ||x-y||$).

Proof.

[(1)]:

Since $A \subseteq \cup_{i=1}^n B(x_i;\epsilon)$ for an ϵ-covering $\{x_1,...,x_n\}$, we have that

$$\begin{aligned} \text{Vol}(A) &\leq \sum_{i=1}^n \text{Vol}(B(x_i;\epsilon)) \\ &= n\epsilon^d\,\text{Vol}(B) \\ \implies n &\geq \left(\frac{1}{\epsilon}\right)^d \frac{\text{Vol}(A)}{\text{Vol}(B)} \end{aligned}$$

[(2)]

Since $\cup_{i=1}^m B\left(a_i;\frac{\epsilon}{2}\right) \subseteq A + \frac{\epsilon}{2}B$ and the sets are disjoint under an ϵ-packing $\{a_1,...,a_m\}$, we have that

$$\begin{aligned} \text{Vol}\left(A+\frac{\epsilon}{2}B\right) &\geq \sum_{i=1}^m \text{Vol}\left(B\left(a_i;\frac{\epsilon}{2}\right)\right) \\ &= m\left(\frac{\epsilon}{2}\right)^d \text{Vol}(B) \\ \implies m &\leq \left(\frac{2}{\epsilon}\right)^d \frac{\text{Vol}\left(A+\frac{\epsilon}{2}B\right)}{\text{Vol}(B)} \end{aligned}$$

□

Application 147 (Unit Ball Packing and Covering)**.** Suppose that $A := \{x : ||x|| \leq 1\}$ is the unit ball in $\mathbb{R}^d$ with the metric induced by $||\cdot||$, then, we can apply Lemma 146 to say that for any $0<\epsilon\leq 1$, we have that

$$\begin{aligned}\left(\frac{1}{\epsilon}\right)^d &\overset{(1)}{\leq} N(A,||\cdot||,\epsilon)\\ &\leq M(A,||\cdot||,\epsilon)\\ &\overset{(2)}{\leq} \left(\frac{2}{\epsilon}\right)^d \frac{\operatorname{Vol}\left(A+\frac{\epsilon}{2}B\right)}{\operatorname{Vol}(B)}\\ &\leq \left(\frac{2}{\epsilon}\right)^d \left(1+\frac{\epsilon}{2}\right)^d \frac{\cancel{\operatorname{Vol}(B)}}{\cancel{\operatorname{Vol}(B)}}\\ &\leq \left(1+\frac{2}{\epsilon}\right)^d\\ &\leq \left(\frac{3}{\epsilon}\right)^d, \text{ since } \epsilon\in(0,1]\end{aligned}$$

Example 148 (Gilbert-Varshamov Bound)**.** Let $A:=\{0,1\}^d$ and $d_H(x,x'):=\sum_{i=1}^d \mathbb{1}_{\{x_i\neq x_i'\}}$ be the Hamming distance in the metric space (A,d_H). Then, for $1\leq r\leq d-1$, we have that

$$\begin{aligned}\frac{2^d}{\sum_{i=0}^r\binom{d}{i}} &\overset{(1)}{\leq} M(A,d_H,r)\\ &\overset{(2)}{\leq} \frac{2^d}{\sum_{i=0}^{\lfloor r/2\rfloor}\binom{d}{i}}\end{aligned}$$

If in addition, $r:=\rho d$ with $d\to\infty$ and $\rho\in(0,1)$, then by Stirling's approximation, we have that

$$\begin{aligned}2^{d(1-h(\rho)+o(1))} &\leq M\left(\{0,1\}^d,d_H,\rho d\right)\\ &\leq 2^{d(1-h(\rho/2)+o(1))}\end{aligned}$$

where $h(p):=p\log_2(1/p)+(1-p)\log_2(1/(1-p))$ as a reminder.

Proof.

The asymptotic behavior is left to Lemma 122 where something very similar to the asymptotic lower bound is shown.

[(1)]:

First, observe that $|B(x;r)|=\sum_{i=1}^r\binom{d}{i}$ since a point is within r units (according to the Hamming distance) of x if at most r bits are flipped and that sum gives the number of ways to flip r bits. Then, we note that for any r-covering of A given by $\{x_1,...,x_n\}$, $\{0,1\}^d\subseteq\cup_{i=1}^n B(x_i,r)\implies 2^d\leq n\sum_{i=1}^r\binom{d}{i}$. In turn since this holds for all r-coverings, by Lemma 145, we get that

$$\begin{aligned}\frac{2^d}{\sum_{i=0}^r\binom{d}{i}} &\leq N(A,d_H,r)\\ &\leq M(A,d_H,r)\end{aligned}$$

[(2)]:

Take any r-packing of A given by $\{a_1,...,a_m\}$. We note that since $A=\{0,1\}^d$ is the entire metric space, we have

that

$$\cup_{i=1}^{m} B(a_i; \lfloor r/2 \rfloor) \subseteq \{0,1\}^d$$

and since each of the balls are disjoint, we have that

$$\begin{aligned} m \sum_{i=1}^{\lfloor r/2 \rfloor} \binom{d}{i} &\leq 2^d, \text{ for all } r\text{-packings} \\ \implies M(A, d_H, r) &\leq \frac{2^d}{\sum_{i=0}^{\lfloor r/2 \rfloor} \binom{d}{i}} \end{aligned}$$

□

Lemma 149 (Stein's Lemma)**.** Let $X \sim \mathcal{N}(\mu, \Sigma)$ where $\mathrm{supp}(X) = \mathbb{R}^d$ and the density is given by p_X. Also let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with at most polynomial growth and assume that all relevant expectations exist. Then, we have that

$$\mathbb{E}[(X - \mu) f(X)] = \Sigma \mathbb{E}[\nabla f(X)]$$

Proof.

As a first remark, we note that

$$\begin{aligned} \nabla p_X(x) &= -\Sigma^{-1}(x - \mu) p_X(x) \\ \implies \Sigma \nabla p_X(x) &= -(x - \mu) p_X(x) \end{aligned}$$

We have that

$$\begin{aligned} \mathbb{E}[(X - \mu) f(X)] &= \int_{x \in \mathbb{R}^d} (x - \mu) f(x) p_X(x) dx \\ &= -\int_{x \in \mathbb{R}^d} f(x) \Sigma \nabla p_X(x) dx, \text{ by above} \\ &= [-f(x) \Sigma p_X(x)]_{-\infty^d}^{\infty^d} + \Sigma \int_{x \in \mathbb{R}^d} \nabla f(x) p_X(x) dx \\ &= 0 + \Sigma \mathbb{E}[\nabla f(X)], \text{ by assumption that } f \text{ has at most polynomial growth} \\ &= \Sigma \mathbb{E}[\nabla f(X)] \end{aligned}$$

□

Lemma 150 (Slepian's Lemma)**.** Let X, Y be independent centered Gaussian random variables in $\mathbb{R}^d$ with

$$\mathbb{E}\left[(Y_i - Y_j)^2\right] \leq \mathbb{E}\left[(X_i - X_j)^2\right] \ \forall i \neq j \in [d]$$

Define their covariances of X and Y by Σ_X and Σ_Y, respectively. Also denote $\Sigma^{(i,)}$ as the ith row of Σ and $\Sigma^{(,j)}$ as the jth column of Σ. Then, we have that

$$\mathbb{E}\left[\max_{i \in [d]} Y_i\right] \leq \mathbb{E}\left[\max_{i \in [d]} X_i\right]$$

Proof.

Define $Z(t) := X\sqrt{t}+Y\sqrt{1-t}$ for $t \in [0,1]$. Also define $F_\beta(Z) := \frac{1}{\beta}\log\left(\sum_{i=1}^d \exp(Z_i\beta)\right)$ so that $\lim_{\beta\to\infty} F_\beta(Z)$ $= \max_{i\in[d]} Z_i$. We wish to show here that for any $\beta \geq 1$, $\frac{\partial \mathbb{E}[F_\beta(Z(t))]}{\partial t} \geq 0$. We have that

$$
\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}[F_\beta(Z(t))] = \mathbb{E}\left[\frac{\partial}{\partial t}F_\beta(Z(t))\right]&, \text{ under regularity conditions}\\
&= \sum_{i=1}^d \mathbb{E}\left[\underbrace{\frac{\partial F_\beta(Z(t))}{\partial z_i}}_{=:p_i^\beta(Z(t))}\frac{\partial z_i(t)}{t}\right]\\
&= \sum_{i=1}^d \mathbb{E}\left[\frac{\partial F_\beta(Z(t))}{\partial z_i}\left(\frac{X_i}{2\sqrt{t}}-\frac{Y_i}{2\sqrt{1-t}}\right)\right]\\
&= \sum_{i=1}^d \frac{1}{2\sqrt{t}}\mathbb{E}\left[\frac{\partial F_\beta(Z(t))}{\partial z_i}X_i\right]-\frac{1}{2\sqrt{1-t}}\mathbb{E}\left[\frac{\partial F_\beta(Z(t))}{\partial z_i}Y_i\right]\\
&= \sum_{i=1}^d \frac{1}{2\sqrt{t}}\mathbb{E}\left[\mathbb{E}\left[\frac{\partial F_\beta(Z(t))}{\partial z_i}X_i \mid Y_i\right]\right]-\frac{1}{2\sqrt{1-t}}\mathbb{E}\left[\mathbb{E}\left[\frac{\partial F_\beta(Z(t))}{\partial z_i}Y_i \mid X_i\right]\right]\\
&= \sum_{i=1}^d \frac{1}{2t}\mathbb{E}\left[\mathbb{E}\left[\frac{\partial F_\beta(Z(t))}{\partial z_i}(Z_i(t)-Y_i\sqrt{1-t}) \mid Y_i\right]\right]\\
&\quad -\frac{1}{2(1-t)}\mathbb{E}\left[\mathbb{E}\left[\frac{\partial F_\beta(Z(t))}{\partial z_i}(Z_i(t)-\sqrt{t}X_i) \mid X_i\right]\right]\\
&= \sum_{i=1}^d \frac{1}{2t}\mathbb{E}\left[t\Sigma_X^{(i,)'}\partial_{z,z_i}F_\beta(Z(t))\right]-\frac{1}{2(1-t)}\mathbb{E}\left[(1-t)\Sigma_Y^{(i,)'}\partial_{z,z_i}F_\beta(Z(t))\right], \text{ by Lemma 149}
\end{aligned}
$$

Continuing, we have that

$$\begin{aligned}
\frac{\partial}{\partial t}\mathbb{E}[F_\beta(Z(t))] &= \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\Sigma_X^{(i,j)} - \Sigma_Y^{(i,j)}\right)\mathbb{E}\left[\partial_{z_j,z_i}F_\beta(Z(t))\right] \\
&= \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\left(\Sigma_X^{(i,j)} - \Sigma_Y^{(i,j)}\right)\mathbb{E}\left[\mathbb{1}_{\{i=j\}}p_i^\beta(Z(t)) - p_i^\beta(Z(t))p_j^\beta(Z(t))\right] \\
&= \frac{1}{2}\sum_{i=1}^{d}\left(\Sigma_X^{(i,i)} - \Sigma_Y^{(i,i)}\right)\mathbb{E}[p_i^\beta(Z(t))(1-p_i^\beta(Z(t)))] \\
&\quad - \frac{1}{2}\sum_{i=1}^{d}\sum_{j\neq i}^{d}\left(\Sigma_X^{(i,j)} - \Sigma_Y^{(i,j)}\right)\mathbb{E}[p_i^\beta(Z(t))p_j^\beta(Z(t))] \\
&= \frac{1}{2}\sum_{i=1}^{d}\sum_{j\neq i}^{d}\left(\Sigma_X^{(i,i)} - \Sigma_Y^{(i,i)}\right)\mathbb{E}[p_i^\beta(Z(t))p_j^\beta(Z(t))] \\
&\quad - \frac{1}{2}\sum_{i=1}^{d}\sum_{j\neq i}^{d}\left(\Sigma_X^{(i,j)} - \Sigma_Y^{(i,j)}\right)\mathbb{E}[p_i^\beta(Z(t))p_j^\beta(Z(t))] \\
&= \frac{1}{2}\sum_{i=1}^{d}\sum_{j\neq i}^{d}\left(\Sigma_X^{(i,i)} - \Sigma_Y^{(i,i)} - \Sigma_X^{(i,j)} + \Sigma_Y^{(i,j)}\right)\mathbb{E}[p_i^\beta(Z(t))p_j^\beta(Z(t))] \\
&= \frac{1}{4}\sum_{i=1}^{d}\sum_{j\neq i}^{d}\left(\mathbb{E}[(X_i - X_j)^2] - \mathbb{E}[(Y_i - Y_j)^2]\right)\mathbb{E}[p_i^\beta(Z(t))p_j^\beta(Z(t))] \\
&\geq 0
\end{aligned}$$

since $p_i^\beta(Z(t)) > 0$ for any $Z(t)$, $i \in [d]$, and $\beta \geq 1$ and the remaining difference is non-negative by assumption for each $i \neq j$. As a result, we have that

$$\begin{aligned}
&\mathbb{E}[F_\beta(Z(1))] \geq \mathbb{E}[F_\beta(Z(0))] \\
\implies &\mathbb{E}[F_\beta(X)] \geq \mathbb{E}[F_\beta(Y)] \\
\implies &\lim_{\beta\to\infty}\mathbb{E}[F_\beta(X)] \geq \lim_{\beta\to\infty}\mathbb{E}[F_\beta(X)] \\
\implies &\mathbb{E}\left[\lim_{\beta\to\infty}F_\beta(X)\right] \geq \mathbb{E}\left[\lim_{\beta\to\infty}F_\beta(Y)\right], \text{ under dominance} \\
\implies &\mathbb{E}\left[\max_{i\in[d]}X_i\right] \geq \mathbb{E}\left[\max_{i\in[d]}Y_i\right]
\end{aligned}$$

A remark[a]. □

[a] The dominance condition is relatively easy to show. We need that $|F_\beta(X)| < M$ for some M for each β along the asymptotic sequence and $\mathbb{E}[|M|] < \infty$ (the same must hold for Y).

Lemma 151 (Asymptotics of Maximum of n Standard Gaussians)**.** Suppose that $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Then, we have that

$$\mathbb{E}\left[\max_{i\in[n]}X_i\right] = (1 + o(1))\sigma\sqrt{2\log(n)}$$

Proof.

[Upper Bound]:

We have that for any $s > 0$

$$\begin{aligned}
\exp\left(s\mathbb{E}\left[\max_{i\in[n]} X_i\right]\right) &\leq \mathbb{E}\left[\exp\left(s\max_{i\in[n]} X_i\right)\right] \\
&= \mathbb{E}\left[\max_{i\in[n]} \exp(sX_i)\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} \exp(sX_i)\right] \\
&= n\exp\left(\frac{\sigma^2 s^2}{2}\right), \text{ by MGF of } \mathcal{N}(0,\sigma^2) \\
\implies \mathbb{E}\left[\max_{i\in[n]} X_i\right] &\leq \frac{\log(n)}{s} + \frac{\sigma^2 s}{2}
\end{aligned}$$

Taking the derivative of the RHS with respect to s and setting it equal to 0 to find the minimum (the second order condition can be easily verified) gives that

$$\begin{aligned}
0 &= -\frac{\log(n)}{s^2} + \frac{\sigma^2}{2} \\
\implies s &= \frac{\sqrt{2\log(n)}}{\sigma} \\
\implies \mathbb{E}\left[\max_{i\in[n]} X_i\right] &\leq \sqrt{2\log(n)}
\end{aligned}$$

[Lower Bound]:

Omitted. Requires thinking about some tedious asymptotics. To see a proof, see here.

Theorem 152 (Sudakov Minoration)**.** Let $w(A) := \mathbb{E}\left[\sup_{a\in A}\langle a, Z\rangle\right]$ where $Z \sim \mathcal{N}(0, I_d)$ be the Gaussian width of the compact $A \subseteq \mathbb{R}^d$. Then, we have that for some $C > 0$

$$w(A) \geq C \sup_{\epsilon>0} \epsilon\sqrt{\log\left(M(A), ||\cdot||_2, \epsilon\right)}$$

Proof.

Take any $\epsilon > 0$ and let $\{a_1, ..., a_m\}$ be an optimal ϵ-packing of A.[a] Define $X_i := \langle a_i, Z\rangle$ and let $Y_i \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\epsilon^2}{2}\right)$

for $i \in [m]$. Then, we have that for any $i \neq j \in [m]$

$$\begin{aligned}
\mathbb{E}[(Y_i - Y_j)^2] &= \epsilon^2 \\
&\leq ||a_i - a_j||^2 \\
&= \mathbb{E}[(X_i - X_j)^2] \\
\implies w(A) &\geq \mathbb{E}\left[\max_{i\in[m]} X_i\right] \\
&\geq \mathbb{E}\left[\max_{i\in[m]} Y_i\right], \text{ by Lemma 150} \\
&= (1+o(1))\frac{\epsilon}{\sqrt{2}}\sqrt{2\log(m)}, \text{ by Lemma 151} \\
&\geq C\epsilon\sqrt{\log\left(M(A, ||\cdot||_2, \epsilon)\right)}
\end{aligned}$$

for some universal constant $C > 0$ that doesn't depend on ϵ (since we can rewrite $X_i =_d \frac{\epsilon}{\sqrt{2}}\tilde{X}_i$ for $\tilde{X}_i \overset{iid}{\sim} \mathcal{N}(0,1)$) and is relevant whenever $M(A, ||\cdot||_2, \epsilon) \geq 2$.[b] In light of the arbitrariness of ϵ, we have the result. □

[a] The attainment of the optimal packing is guaranteed by the compactness of A. One can argue that any 2ϵ-covering of A has a finite subcover since A is compact so that with Lemma 145, one has the ϵ-packing number is finite. Finally, since the number of elements in a packing is an integer, the ϵ-packing number is attained.

[b] When $M(A, ||\cdot||, \epsilon) = 1$ we have the result trivially since $w(A) \geq 0$ since $0 = \sup_{a\in A}\mathbb{E}[\langle a, Z\rangle] \leq \mathbb{E}\left[\sup_{a\in A}\langle a, Z\rangle\right]$ by the convexity of sup and Jensen's Inequality. $M(A, ||\cdot||_2, \epsilon) = 0$ is not possible.

Application 153 (Application of Sudakov Minoration)**.** Let $B_1 = \{x \in \mathbb{R}^d : ||x||_1 \leq 1\}$. Then, we have that

$$\begin{aligned}
w(B_1) &= \mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\left[\sup_{||x||\leq 1}\langle x, Z\rangle\right], \text{ by definition} \\
&= \mathbb{E}[||Z||_\infty] \\
&\leq \sqrt{w\log(d)}, \text{ by the upper bound of Lemma 151}
\end{aligned}$$

We can then apply Theorem 152 to say that for any $\epsilon > 0$

$$\log\left(M(B_1, ||\cdot||_2, \epsilon)\right) = O\left(\frac{\log(d)}{\epsilon^2}\right)$$

In fact, it holds that

$$\log\left(M(B_1, ||\cdot||_2, \epsilon)\right) \asymp \begin{cases} d\left(1+\log\left(\frac{1}{\epsilon^2 d}\right)\right) & \text{if } \epsilon \leq \frac{1}{\sqrt{d}} \\ \frac{1+\log(\epsilon^2 d)}{\epsilon^2} & \text{if } \frac{1}{\sqrt{d}} < \epsilon < 1 \end{cases}$$

Theorem 154 (Maurey's Empirical Method)**.** Let $(H, \langle\cdot,\cdot\rangle)$ be an inner product space and $T \subseteq H$ be a finite set. Then, we have that

$$N\left(\text{conv}(T), ||\cdot||, \epsilon\right) \leq \binom{|T| + \lceil\frac{r^2}{\epsilon^2}\rceil - 2}{\lceil\frac{r^2}{\epsilon^2}\rceil - 1}, \text{ if } 0 < \epsilon \leq r^2$$

with r, the radius of T, defined as

$$r := \inf_{y\in H}\sup_{x\in T}||x - y||$$

Proof.

We will use a probabilistic argument. Let $T := \{t_1, ..., t_m\}$ and $c \in H$ satisfy $r = \max_{i\in[m]} ||t_i - c||$.[a] Next, take any $x \in \operatorname{conv}(T)$ so that $x = \sum_{i=1}^m y_i t_i$ for $y \in \Delta^{m-1}$. Also take any $\epsilon > 0$. We let Z be a H-valued random variable with $\Pr(Z = t_i) = y_i$ so that $\mathbb{E}[Z] = x$. Lastly, let $Z_1, ..., Z_n$ be n iid copies of Z and $\bar{Z} := \frac{1}{n+1}\left(c + \sum_{i=1}^n Z_i\right)$. Then, we have that

$$\begin{aligned}
\mathbb{E}\left[||\bar{Z} - x||^2\right] &= \mathbb{E}\left[\left|\left|\frac{1}{n+1}\left(c + \sum_{i=1}^n Z_i\right) - x\right|\right|^2\right] \\
&= \mathbb{E}\left[\left|\left|\frac{1}{n+1}\left((c-x) + \sum_{i=1}^n (Z_i - x)\right)\right|\right|^2\right] \\
&= \frac{1}{(n+1)^2}\left(||c-x||^2 + n\mathbb{E}\left[||Z-x||^2\right] + \sum_{i\neq j}^n \overset{0}{\cancel{\mathbb{E}[Z_i - x]}}\overset{0}{\cancel{\mathbb{E}[Z_i - x]}} + 2\sum_{i=1}^n \overset{0}{\cancel{\mathbb{E}[Z_i - x]}}(c-x)\right) \\
&\leq \frac{r^2}{n+1}
\end{aligned}$$

since $||c - x||^2 \leq r^2$ and $\mathbb{E}\left[||Z - x||^2\right] \leq \mathbb{E}\left[||Z - c||^2\right]$ as $\mathbb{E}[Z] = x$, and $\mathbb{E}\left[||Z - c||^2\right] \leq r^2$ by construction of c. Consequently, if $n := \lceil \frac{r^2}{\epsilon^2} \rceil - 1$, there exists a realization of $\bar{Z}$ such that $||\bar{Z} - x|| \leq \epsilon$. Meanwhile,

$$\operatorname{supp}(\bar{Z}) = \{\frac{1}{n+1}\left(c + \sum_{i=1}^m t_i n_i\right) : n_i \geq 0, \sum_{i=1}^n n_i = n\}$$

which has cardinality $\binom{n+m-1}{n}$ by stars and bars. In light of the arbitrariness of x, we have that $\operatorname{supp}(\bar{Z})$ is an ϵ-covering for $\operatorname{conv}(T)$, which implies that

$$N\left(\operatorname{conv}(T), ||\cdot||, \epsilon\right) \leq \binom{|T| + \lceil \frac{r^2}{\epsilon^2} \rceil - 2}{\lceil \frac{r^2}{\epsilon^2} \rceil - 1}, \text{ if } 0 < \epsilon \leq r^2$$

□

[a] Note that since $|T| < \infty$, $\sup_{i\in[m]} ||t_i - y|| = \max_{i\in[m]} ||t_i - y||$ and that $f(y) := \max_{i\in[m]} ||t_i - y||$ is continuous. Also note that $f(y) \to \infty$ as $||y|| \to \infty$ so that f is coercive implying that it's bounded from below. Thus, if H is finite dimensional, we can restrict our search of the optimal $c \in H$ that attains the radius to a compact set (eg., $\inf_{y\in H} f(y) = \inf_{||y||\leq M} f(y)$ for some $M > 0$) and the existence of c is then given by the fact that a continuous function on a compact set attains its extremal points. If H is infinite dimensional, I think we may need to additionally assume something stronger about H (eg., it's a Hilbert space).

Application 155 (Application of Maurey's Empirical Method)**.** Let $B_1 := \{\pm e_1, ..., \pm e_d\} \subseteq \mathbb{R}^d$ endowed with the natural inner product. We have that the radius of B_1 is 1. By Theorem 154, we have that

$$\begin{aligned}
\log\left(N(\operatorname{conv}(B_1), ||\cdot||_2, \epsilon)\right) &\leq \log\left(\binom{2d + \lceil \frac{1}{\epsilon^2} \rceil - 2}{\lceil \frac{1}{\epsilon^2} \rceil - 1}\right) \\
&= O\left(\frac{1 + \log(\epsilon^2 d)}{\epsilon^2}\right), \text{ by Stirling's approximation if } \frac{1}{\sqrt{d}} < \epsilon < 1
\end{aligned}$$

To see the approximation, let $n := 2d + \lceil \frac{1}{\epsilon^2} \rceil - 2$ and $k := \lceil \frac{1}{\epsilon^2} \rceil - 1$. We have that

$$\begin{aligned}
\binom{n}{k} &= \frac{n \cdot (n-1) \cdot (n-k+1)}{k!} \\
&\leq \left(\frac{ne}{k}\right)^k, \text{ since } k! \geq \left(\frac{k}{e}\right)^k \text{ by a finite version of Stirling} \\
\implies \log\left(\binom{n}{k}\right) &\leq k \log\left(\frac{ne}{k}\right) \\
&= O\left(\left(\frac{1}{\epsilon^2} - 1\right)\left(1 + \log\left(\frac{2d + \frac{1}{\epsilon^2} - 2}{\frac{1}{\epsilon^2} - 1}\right)\right)\right) \\
&= O\left(\left(\frac{1}{\epsilon^2} - 1\right)\left(1 + \log\left(\frac{2d\epsilon^2 + 1 - 2\epsilon^2}{1 - \epsilon^2}\right)\right)\right) \\
&= O\left(\frac{1 + \log\left(\epsilon^2 d\right)}{\epsilon^2}\right), \text{ since } \epsilon = o(1) \text{ and } \frac{1}{\sqrt{d}} < \epsilon < 1
\end{aligned}$$

9.2 Onto the Global Fano's Method

Recall the settings and steps of applying Theorem 117 (ie., Fano's Inequality). We consider a statistical model $\{P_\theta : \theta \in \Theta\}$ and we wish to construct an estimator $\hat{\theta}(X)$ for θ from an observation $X \sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta, a)$.

(1) We find a pairwise separated set $\{\theta_1, ..., \theta_m\} \subseteq \Theta$ such that

$$\min_{i \neq j} \inf_a \left(L(\theta_i, a) + L(\theta_j, a)\right) \geq \Delta$$

(2) Try to upper bound $I(\theta; X)$ with $\theta \sim \text{Unif}(\{\theta_1, ..., \theta_m\})$ and $X \mid \theta \sim P_\theta$.

(3) If $I(\theta; X) < \frac{1}{2}\log(m)$ then the minimax risk satisfies $r^* = \Omega(\Delta)$.

A packing of $\Theta_0 \subseteq \Theta$ as defined in Definition 143 gives an alternative strategy to complete Step 1. If there's a metric $d(\theta, \theta')$ on $\Theta \times \Theta$ so that

$$\min_a L(\theta, a) + L(\theta', a) \geq h(d(\theta, \theta'))$$

for a strictly increasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$, then a δ-packing of $\Theta_0 = \{\theta_1, ..., \theta_m\}$ of Θ under d satisfies the separation condition with $\Delta = h(\delta)$.

Definition 156 (Smallest f-Divergence Covering)**.** For a family $\mathcal{P}$ of distributions and $\epsilon > 0$, for some f-divergence D, let $N_D(\mathcal{P}, \epsilon)$ be the smallest integer n such that there exist distributions $Q_1, ..., Q_n$ (not necessarily in $\mathcal{P}$) satisfying

$$\sup_{P \in \mathcal{P}} \min_{i \in [n]} D(P || Q_i) \leq \epsilon^2$$

As a first remark, sometimes we will say that the LHS is $\leq \epsilon$ as opposed to ϵ^2, this will be interpreted within the context. As a second remark, recall that D might be a distance metric and also might be a divergence (eg., KL-divergence). Also note that the distributions $\{Q_i : i \in [n]\}$ are in the second argument of D, which is relevant when we're dealing with a divergence.

Theorem 157 (Entropic Upper Bound of $I(\theta; X^n)$)**.** Consider a statistical model $\{P_\theta : \theta \in \Theta\}$ and let $\theta \sim \pi$ where $\text{supp}(\pi) = \Theta_0$ for $\Theta_0 \subseteq \Theta$ and $X^n \mid \theta \sim P_\theta^{\otimes n}$. Suppose that for any $\epsilon > 0$, there exists a finite ϵ-KL covering of $\{P_\theta : \theta \in \Theta_0\}$ so that $N_{KL}(\{P_\theta : \theta \in \Theta_0\}, \epsilon) < \infty$.[a] Then, we have that

$$I(\theta; X^n) \leq \inf_{\epsilon > 0} \left[n\epsilon^2 + \log\left(N_{KL}(\{P_\theta : \theta \in \Theta_0\}, \epsilon)\right)\right]$$

Proof.

Recall from Lemma 116, we have that

$$I(\theta; X^n) = \min_{Q_{X^n}} \mathbb{E}_{\theta\sim\pi}\left[D_{KL}(P_\theta^{\otimes n} || Q_{X^n})\right]$$

Take any $\epsilon > 0$. For an ϵ-KL covering of $\{P_\theta : \theta \in \Theta\}$ given by $\{Q_1, ..., Q_N\}$ with $N := N_{KL}(\{P_\theta : \theta \in \Theta_0\}, \epsilon)$, we define $Q_{X^n} := \frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}$. Then, for $\theta \sim \pi$, we have that

$$\begin{aligned}
D_{KL}\left(P_\theta^{\otimes n} || Q_{X^n}\right) &= \mathbb{E}_{X\sim P_\theta^{\otimes n}}\left[\log\left(\frac{P_\theta^{\otimes n}(X)}{\frac{1}{N}\sum_{i=1}^N Q_i^{\otimes n}(X)}\right)\right] \\
&\leq \mathbb{E}_{X\sim P_\theta{}^{\otimes n}}\left[\min_{i\in[N]} \log\left(\frac{P_\theta^{\otimes n}(X)}{Q_i^{\otimes n}(X)}\right)\right] + \log(N), \text{ since } Q_i^{\otimes n}(X) \geq 0 \\
&\leq \min_{i\in[N]} \mathbb{E}_{X\sim P_\theta{}^{\otimes n}}\left[\log\left(\frac{P_\theta^{\otimes n}(X)}{Q_i^{\otimes n}(X)}\right)\right] + \log(N), \text{ by Jensen's Inequality with concave } \min(\cdot) \\
&= \min_{i\in[N]} nD_{KL}\left(P_\theta || Q_i\right) + \log(N), \text{ by Property 19} \\
&\leq n\epsilon^2 + \log(N), \text{ by } \epsilon\text{-KL Covering}
\end{aligned}$$

□

[a] As a remark, we're assuming the existence of a finite n such that $\{P_\theta : \theta \in \Theta_0\}$ is "covered" by n distributions in the sense of KL-divergence. Obviously, this holds trivially if $\{P_\theta : \theta \in \Theta_0\}$ is finite since we can take $\{P_\theta : \theta \in \Theta_0\}$ itself as a candidate covering. That said, this assumption must be argued when applying the theorem.

We now discuss the Global Fano's Method. Again, we consider a statistical model $\{P_\theta : \theta \in \Theta\}$ and we wish to construct an estimator $\hat{\theta}(X)$ for θ from an observation $X \sim P_\theta$ with support $\mathcal{X}$ penalizing losses with $L(\theta, a)$. For hyperparameters $\Theta_0 \subseteq \Theta$ and $\epsilon, \delta > 0$,

(1) We find a metric $d(\cdot, \cdot)$ satisfying $\min_{\theta,\theta'\in\Theta_0} \min_a L(\theta, a) + L(\theta', a) \geq h(d(\theta, \theta'))$ for a strictly increasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$. Next, we find a δ-packing of Θ_0 under d.

(2) We find an ϵ-KL covering of $\{P_\theta : \theta \in \Theta_0\}$, arguing that it exists in the first place.

(3) We apply Fano's method in Theorem 117 and Theorem 157 to conclude that

$$r^* \geq \frac{h(\delta)}{2}\left(1 - \frac{\log(N_{KL}(\{P_\theta : \theta \in \Theta_0\}, \epsilon)) + n\epsilon^2 + \log(2)}{\log\left(M(\Theta_0, d, \delta)\right)}\right)$$

(4) We optimize over $(\Theta_0, \delta, \epsilon)$ to make the lower bound as large as possible.

Example 158 (Gaussian Location Model with Global Fano's Method)**.** Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0, I_d)$ with unknown

$\theta \in \mathbb{R}^d$ that we want to estimate under $||\cdot||_p$. We wish to show that

$$r^* = \inf_{\hat{\theta}} \sup_{\theta} \mathbb{E}_{X^n \sim P_\theta^{\otimes n}} \left[||\hat{\theta}(X) - \theta||_p \right]$$
$$\gtrsim \begin{cases} \frac{d^{1/p}}{\sqrt{n}} & \text{if } 2 < p < \infty \\ \sqrt{\frac{\log(d)}{n}} & \text{if } p = \infty \end{cases}$$

Proof.

Choose Θ_0 to be a maximally sized δ-packing of $\{\theta \in \mathbb{R}^d : ||\theta||_2 \leq r\}$ under $||\cdot||_p$ for $p > 2$ and r tbd so that the separation condition holds with $||\cdot||_p$ and $h(x) = x$ by the triangle inequality. Then, for any $\epsilon, \delta > 0$, Global Fano's Method gives that

$$r^* \gtrsim \delta \left(1 - \frac{\log\left(N_{KL}(\{\mathcal{N}(\theta, I_d) : \theta \in \Theta_0\})\right) + n\epsilon^2 + \log(2)}{\log\left(M(\Theta_0, ||\cdot||_p, \delta)\right)} \right)$$
$$= \delta \left(1 - \frac{\log\left(N(\Theta_0, ||\cdot||_2, \epsilon\sqrt{2})\right) + n\epsilon^2 + \log(2)}{\log\left(M(\Theta_0, ||\cdot||_p, \delta)\right)} \right)$$

where the second line follows from the fact that $D_{KL}(\mathcal{N}(\theta, I_d)||\mathcal{N}(\theta', I_d)) = \frac{1}{2}||\theta - \theta'||_2^2$ where this equality also guarantees the existence of $N_{KL}(\{\mathcal{N}(\theta, I_d) : \theta \in \Theta_0\})$ since Θ_0 is compact so that any covering (eg., Θ_0 itself) has a finite subcovering for any $\epsilon > 0$.

Now, we think about picking the hyperparameters. We choose $\epsilon = \frac{r}{\sqrt{2}}$ so that $\log(N(\Theta_0, ||\cdot||_2, \sqrt{2}\epsilon)) = \log(1) = 0$, since we get an r-covering by picking the origin. For $p \in (2, \infty)$, we pick $\frac{\delta}{r} = \frac{1}{K} d^{\frac{1}{p} - \frac{1}{2}}$ for some $K > 0$ tbd so that $\log(M(\Theta_0, ||\cdot||_p, \delta)) \gtrsim d$.[a] To see this, take the inequality from Lemma 146 to say that ($B_p := \{x \in \mathbb{R}^d : ||x||_p \leq 1\}$)

$$\begin{aligned} M(\Theta_0, ||\cdot||_p, \delta) &\geq \left(\frac{1}{\delta}\right)^d \frac{\mathrm{Vol}(\Theta_0)}{\mathrm{Vol}(B_p)} \\ &= \left(\frac{r}{\delta}\right)^d \frac{\mathrm{Vol}(B_2)}{\mathrm{Vol}(B_p)} \\ &= \left(\frac{r}{\delta}\right)^d \frac{\left(\frac{(2\Gamma(1/2+1))^d}{\Gamma(d/2+1)}\right)}{\left(\frac{(2\Gamma(1/p+1))^d}{\Gamma(d/p+1)}\right)}, \text{ using link} \\ &\asymp \left(\frac{r}{\delta}\right)^d \left(d^{\frac{1}{p}-\frac{1}{2}}\right)^d C^d, \text{ using } \Gamma(x+1) \asymp \sqrt{2\pi x}\,(x/e)^x \text{ and some } C > 0 \\ &= (KC)^d, \text{ using our selection of } \frac{\delta}{r} = \frac{1}{K} d^{\frac{1}{p}-\frac{1}{2}} \\ \implies \log\left(M(\Theta_0, ||\cdot||_p, \delta)\right) &\gtrsim d, \text{ with } K \text{ picked so that } KC = e \end{aligned}$$

For $p = \infty$, we choose $\frac{\delta}{r} = 1$ so that $\log(M(\Theta_0, ||\cdot||_p, \delta)) = \log(d)$. Finally, we move to pick r. As of now, we have that

$$r^* \gtrsim \begin{cases} r d^{\frac{1}{p}-\frac{1}{2}} \left(1 - \frac{\frac{1}{2}r^2 n + \log(2)}{d}\right) & \text{if } p \in (2, \infty) \\ r\left(1 - \frac{\frac{1}{2}r^2 n + \log(2)}{\log(d)}\right) & \text{if } p = \infty \end{cases}$$

so that choosing

$$r = \begin{cases} \sqrt{\frac{d}{n}} & \text{for } p \in (2,\infty) \\ \sqrt{\frac{\log(d)}{n}} & \text{for } p = \infty \end{cases}$$

gives that

$$r^* \gtrsim \begin{cases} \frac{d^{1/p}}{\sqrt{n}} & \text{if } p \in (2,\infty) \\ \sqrt{\frac{\log(d)}{n}} & \text{if } p = \infty \end{cases}$$

□

[a] We use the classical result that for $x \in \mathbb{R}^d$, $||x||_p \leq ||x||_q \leq d^{\frac{1}{q}-\frac{1}{p}}||x||_p$ for $1 \leq q \leq p < \infty$.

Lemma 159 (Asymptotic Log Covering Number for Bounded s-Derivative Functions)**.** Let $H^s := \{f \in C^s([0,1]) : ||f^{(s)}||_\infty \leq 1\}$ for some $s \in \mathbb{N}$. We have that

$$\log\left(N(H^s, ||\cdot||_p, \epsilon)\right) \asymp_p \epsilon^{-\frac{1}{s}}$$

for any $p \in [1,\infty)$. By $\asymp_p$, we mean that that asymptotic rate is achieved up to a constant that only depends on p.

Proof.

Omitted.

Lemma 160 (Asymptotic Log Covering Number for Monotonic Functions)**.** Let $\mathcal{F}_M := \{f : [0,1] \to [0,1],$ f is non-decreasing$\}$. We have that

$$\log\left(N(\mathcal{F}_M, ||\cdot||_p, \epsilon)\right) \asymp_p \frac{1}{\epsilon}$$

for any $p \in [1,\infty)$. By $\asymp_p$, we mean that that asymptotic rate is achieved up to a constant that only depends on p.

Proof.

Omitted.

Example 161 (Nonparametric Density Estimation with Global Fano's Method)**.** Suppose that $X_1, ..., X_n \overset{iid}{\sim} f$ where $\text{supp}(X_i) = [0,1]$ for $i \in [n]$ and $||f^{(s)}||_\infty \leq 1$ for some $s \in \mathbb{N}$. We define $H^s := \{f : ||f^{(s)}||_\infty \leq 1\}$ We wish to show that

$$\begin{aligned} r^* &= \inf_{\hat{f}} \sup_{f \in H^s} \mathbb{E}_{X^n \sim f^{\otimes n}}\left[||\hat{f}(X^n) - f(X^n)||_p\right] \\ &\gtrsim n^{-\frac{-s}{2s+1}}, \text{ for } p \in [1,\infty) \end{aligned}$$

Proof.

Consider $H_0^s \subseteq H^s$ given by $H_0^s := \{f \in H^s : f \geq \frac{1}{2} \text{ on } [0,1]\}$. Then, for $f, g \in H_0^s$, we have that

$$\begin{aligned}
D_{KL}(f||g) &\leq \log\left(1 + \chi^2(f||g)\right), \text{ by Section 3.8} \\
&\leq \chi^2(f||g), \text{ since } \log(1+x) \leq x \; \forall x \in \mathbb{R}_+ \\
&= \int_{x\in[0,1]} \frac{|f(x)-g(x)|^2}{f(x)} dx \\
&\leq 2||f-g||_2^2 \\
\implies N_{KL}(H_0^s, \epsilon) &\leq N(H_0^s, ||\cdot||_2, \epsilon/\sqrt{2})
\end{aligned}$$

for any $\epsilon > 0$. By results from analysis, we know that H_0^s is compact in $(C^s([0,1]), ||\cdot||_2)$ so that $N(H_0^s, ||\cdot||_2, \epsilon/\sqrt{2}) < \infty$ giving us that $N_{KL}(H_0^s, \epsilon) < \infty$. The separation condition holds trivially by the triangle inequality with $h(x) = x$ for a δ-packing of H_0^s under $||\cdot||_p$. Then, by the Global Fano's Method, we have that for any $\epsilon, \delta > 0$,

$$\begin{aligned}
r^* &\gtrsim \delta \left(1 - \frac{\log\left(N_{KL}(H_0^s, \epsilon) + n\epsilon^2 + \log(2)\right)}{\log\left(M(H_0^s, ||\cdot||_p, \epsilon/\sqrt{2})\right)}\right) \\
&\gtrsim \delta \left(1 - \frac{c_1\epsilon^{-1/s} + n\epsilon^2 + \log(2)}{c_2\delta^{-1/s}}\right), \text{ Lemma 159 with } c_1, c_2 > 0
\end{aligned}$$

Picking $\epsilon \asymp \delta \asymp n^{-\frac{s}{2s+1}}$, we get that

$$\begin{aligned}
r^* &\gtrsim n^{-\frac{s}{2s+1}} \left(1 - \frac{c_1 n^{\frac{1}{2s+1}} + n^{\frac{1}{2s+1}} + \log(2)}{c_2 n^{\frac{1}{2s+1}}}\right) \\
&= \Omega\left(n^{-\frac{-s}{2s+1}}\right)
\end{aligned}$$

□

Example 162 (Isotonic Regression with Global Fano's Method)**.** Suppose that $X_1, ..., X_n \overset{iid}{\sim} P_X$ where the (known or unknown) P_X has a bounded density on $[0,1]$. We also assume that $Y_i \mid X_i \overset{iid}{\sim} \mathcal{N}(f(X_i), 1) =: P_f$ with $f \in \mathcal{F}_M = \{f : [0,1] \to [0,1], f \text{ is increasing}\}$. We wish to show that

$$\begin{aligned}
r^* &= \inf_{\hat{f}} \sup_{f\in\mathcal{F}_M} \mathbb{E}_{P_f^{\otimes n}}\left[||\hat{f} - f||_p\right] \\
&= \Omega\left(n^{-1/3}\right), \; \forall p \in [1, \infty)
\end{aligned}$$

Proof.

Since P_X has a bounded density, we have that

$$\begin{aligned}
D_{KL}(P_f||P_{f'}) &= \frac{1}{2}||f - f'||^2_{L^2(P_X)}, \text{ (ie., integral is over } P_X) \\
&\leq M||f - f'||_2^2, \text{ for some universal constant } M > 0 \\
\implies N_{KL}(\{P_f : f \in \mathcal{F}_M\}, \epsilon) &\leq N\left(\mathcal{F}_M, ||\cdot||_2, \frac{\epsilon}{M}\right)
\end{aligned}$$

for any $\epsilon > 0$. By results from analysis, we know that $\mathcal{F}_M$ is compact in $(L^2([0,1],[0,1]), ||\cdot||_2)$ so that $N\left(\mathcal{F}_M, ||\cdot||_p, \frac{\epsilon}{M}\right) < \infty$ giving us that $N_{KL}(\{P_f : f \in \mathcal{F}_M\}, \epsilon) < \infty$. The separation condition holds trivially by the triangle inequality for a δ-packing of $\mathcal{F}_M$ with $h(x) = x$ under $||\cdot||_p$. Then, as a result, we can apply the Global Fano Method to say that for any $\epsilon, \delta > 0$, we have that

$$\begin{aligned} r^* &\gtrsim \delta \left(1 - \frac{\log(N(\mathcal{F}_M, ||\cdot||_2, \frac{\epsilon}{M})) + n\epsilon^2 + \log(2)}{\log(M(\mathcal{F}_M, ||\cdot||_p, \delta))}\right) \\ &\gtrsim \delta \left(1 - \frac{\frac{c_1}{\epsilon} + n\epsilon^2 + \log(2)}{c_2/\delta}\right), \text{ by Lemma 160 with } c_1, c_2 > 0 \end{aligned}$$

Choosing $\epsilon \asymp \delta \asymp n^{-1/3}$, we get that

$$\begin{aligned} r^* &\gtrsim n^{-1/3} \left(1 - \frac{c_1 n^3 + n^{1/3} + \log(2)}{c_2 n^3}\right) \\ &= \Omega\left(n^{-1/3}\right) \end{aligned}$$

□

Example 163 (Sparse Linear Regression with Global Fano's Method)**.** Suppose that $Y \mid X \overset{iid}{\sim} \mathcal{N}(X\theta, I_n) =: P_\theta$ with fixed design matrix $X \in \mathbb{R}^{n\times d}$ where all singular values of X are $O(\sqrt{n})$. We also assume that the unknown parameter $\theta \in \mathbb{R}^d$ is sparse, which we mathematically convey by $||\theta||_q \leq R$ for some $q \in (0,1)$ and some $R > 0$. We wish to show that

$$\begin{aligned} r^* &= \inf_{\hat{\theta}} \sup_{||\theta||_q \leq R} \mathbb{E}_{Y|X\sim P_\theta^{\otimes n}} \left[||\hat{\theta} - \theta||_p\right] \\ &= \Omega\left(R^{q/p}\left(\frac{\log(d)}{n}\right)^{\frac{p-q}{2p}}\right) \end{aligned}$$

for some small enough $R < f(n,d)$ and $p \in [1, \infty)$.

Proof.

For notation, define $B_p := \{x \in \mathbb{R}^d : ||x||_p \leq 1\}$ for $p > 0$. Applying Lemma 146, we have that for any $\delta > 0$ (recall that $p \geq 1 > q > 0$)

$$\begin{aligned} M(R\cdot B_q, ||\cdot||_p, \delta) &\geq \left(\frac{1}{\delta}\right)^d \frac{\text{Vol}(R\cdot B_q)}{\text{Vol}(B_p)} \\ &\geq \left(\frac{R}{\delta}\right)^d \frac{\text{Vol}(B_q)}{\text{Vol}(B_p)} \\ &= \left(\frac{R}{\delta}\right)^d \frac{\frac{(2\Gamma(1/q+1))^d}{\Gamma(d/q+1)}}{\frac{(\Gamma(1/p+q))^d}{\Gamma(d/p+1)}}, \text{ using link} \\ &\asymp \left(\frac{R}{\delta}\right)^d \left(d^{\frac{1}{p}-\frac{1}{q}}\right)^d C^d, \text{ using } \Gamma(x+1) \asymp \sqrt{2\pi x}(x/e)^x \text{ and some } C > 0 \\ \implies \log(M(R\cdot B_q, ||\cdot||_p, \delta)) &\gtrsim d\log\left(\frac{R}{\delta} d^{\frac{q-p}{pq}}\right) + d\log(C) \\ &\gtrsim \left(\frac{R}{\delta}\right)^{\frac{pq}{p-q}} \log(d) \quad (*) \end{aligned}$$

if $\delta << Rd^{\frac{1}{p}-\frac{1}{q}}$. That's because

$$\begin{aligned} &\delta << Rd^{\frac{1}{p}-\frac{1}{q}} \\ \implies &d^{\frac{p-q}{pq}} << \frac{R}{\delta} \\ \implies &d << \left(\frac{R}{\delta}\right)^{\frac{pq}{p-q}} \end{aligned}$$

I CANT SEEM TO SHOW THAT WHICH ENDS IN $(*)$

Next, we wish to produce a KL-covering of $\mathcal{P} = \{\mathcal{N}(X\theta, I_n) : ||\theta||_q \leq R\}$. To that end, note that

$$\begin{aligned} D_{KL}(\mathcal{N}(X\theta, I_n)||\mathcal{N}(X\theta', I_n)) &= \frac{1}{2}||X(\theta-\theta')||_2^2 \\ &\leq \sigma_{max}^2(X)||\theta-\theta'||_2^2, \text{ where } \sigma_{max}^2(X) \text{ is the largest singular value of } X \\ &= O(n) \cdot ||\theta-\theta'||_2^2, \text{ by assumption} \\ \implies \log(N_{KL}(\mathcal{P}, \epsilon)) &\leq \log\left(N\left(R \cdot B_q, ||\cdot||_2, \frac{\epsilon}{O(\sqrt{n})}\right)\right) \\ &\asymp \left(\frac{R\sqrt{n}}{\epsilon}\right)^{\frac{2q}{2-q}} \log(d), \text{ if } \epsilon >> R\sqrt{n}\,(d)^{\frac{1}{2}-\frac{1}{q}} \end{aligned}$$

for any $\epsilon > 0$ using an approach like that which bounds $\log(M(R \cdot B_q, ||\cdot||_p, \delta))$ and Lemma 145.[a] Notice that the separation condition holds trivially for a packing of $R \cdot B_q$ by the triangle inequality with $h(x) = x$ under $||\cdot||_p$. Picking $\epsilon \asymp n^{-\frac{q}{4}} R^{\frac{q}{2}} (\log(d))^{\frac{2-q}{4}}$ and $\delta \asymp R^{\frac{q}{p}} \left(\frac{\log(n)}{n}\right)^{\frac{p-q}{2p}}$, we have that

$$\begin{aligned} \log(M(R \cdot B_q, ||\cdot||_p, \delta)) &\gtrsim R^q n^{q/2} (\log(d))^{1-q/2} \\ \log(N_{KL}(\mathcal{P}, \epsilon)) &\lesssim R^q n^{q/2} (\log(d))^{1-q/2} \end{aligned}$$

Then, applying the Global Fano's Method

$$\begin{aligned} r^* &\gtrsim \delta\left(1 - \frac{\log(N_{KL}(\mathcal{P}, \epsilon)) + n\epsilon^2 + \log(2)}{\log\left(M(R \cdot B_q, ||\cdot||_p, \delta)\right)}\right) \\ &= \Omega\left(R^{q/p}\left(\frac{\log(n)}{n}\right)^{\frac{p-q}{2p}}\right) \end{aligned}$$

□

[a] The singular value steps follows from the SVD of $X = U\Lambda V'$ for $U \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{d \times d}$ with orthonormal columns and Λ a diagonal matrix of singular values. We note that for a matrix U with orthonormal columns $||Ux||_2^2 = ||x||_2^2$.

10 Entropic Upper Bounds of Density Estimation

In Section 9, we considered techniques that use packing and covering to prove asymptotic lower bounds for risk via Fano's Inequality. In this section, we will use the techniques to prove upper bounds for density estimation.

In density estimation, we consider the following problem. For $X_1, .., X_n \overset{iid}{\sim} P$ where $P \in \mathcal{P}$ is an unknown distribution target and $\mathcal{P}$ is such that we can form an a finite ϵ-covering using the appropriate metric. For $D \in \{D_{KL}, TV, H^2\}$, we wish to construct an estimator $\hat{P} = \hat{P}(X^n)$ such that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{X^n \sim P^{\otimes n}}\left[D(P, \hat{P})\right]$$

is small. Recall that $r^* := \inf_{\hat{P}} \sup_{P \in \mathcal{P}} \mathbb{E}_{X^n \sim P^{\otimes n}}\left[D(P, \hat{P})\right]$.

10.1 Yang-Barron with KL

In this section, we will show an upper bound on density estimation that depends on KL-divergence thanks to Yang and Barron.

Lemma 164 (Online Guarantee)**.** Let $P_1, ..., P_N$ be an ϵ-covering of $\mathcal{P}$. In other words, we have that

$$\sup_{P \in \mathcal{P}} \min_{i \in [N]} D_{KL}(P||P_i) \le \epsilon^2$$

Also, let $Q_{X^{n+1}} = \frac{1}{N} \sum_{i=1}^N P_i^{\otimes(n+1)}$. We have that

$$\sup_{P \in \mathcal{P}} D_{KL}(P^{\otimes(n+1)}||Q_{X^{n+1}}) \le (n+1)\epsilon^2 + \log(N)$$

This is called an "online" guarantee since it concerns the density estimation performance for joint distributions of $X^{n+1} \sim P^{\otimes(n+1)}$.

Proof.

For any $P \in \mathcal{P}$, we have that

$$\begin{aligned} D_{KL}(P^{\otimes(n+1)}||Q_{X^{n+1}}) &= \mathbb{E}_{X^{n+1} \sim P^{\otimes(n+1)}} \left[\log \left(\frac{P^{\otimes(n+1)}(X^{n+1})}{\frac{1}{N} \sum_{i=1}^N P_i^{\otimes(n+1)}(X^{n+1})} \right) \right] \\ &\le \mathbb{E}_{X^{n+1} \sim P^{\otimes(n+1)}} \left[\min_{i \in [N]} \log \left(\frac{P^{\otimes(n+1)}(X^{n+1})}{P_i^{\otimes(n+1)}(X^{n+1})} \right) \right] + \log(N), \\ &\quad \text{since } P_i^{\otimes(n+1)}(X^{n+1}) \ge 0 \\ &\le \min_{i \in [N]} \mathbb{E}_{X^{n+1} \sim P^{\otimes(n+1)}} \left[\log \left(\frac{P^{\otimes(n+1)}(X^{n+1})}{P_i^{\otimes(n+1)}(X^{n+1})} \right) \right] + \log(N), \text{ by Jensen's} \\ &= \min_{i \in [N]} D_{KL} \left(P^{\otimes(n+1)}||P_i^{\otimes(n+1)} \right) + \log(N) \\ &\le (n+1)\epsilon^2 + \log(N), \text{ by Property 19 and the } \epsilon\text{-covering assumption} \end{aligned}$$

□

Lemma 165 (Progressive Mixing)**.** Take the context of Lemma 164. Given $Q_{X^{n+1}}$, we can define

$$\hat{P}(X) = \frac{1}{n+1} \sum_{t=0}^{n} Q_{X_{t+1}=x|X^t}$$

Note that $\hat{P}$ is a well-defined estimator and depends on X^n. Expanding out the definition of $Q_{X^{n+1}}$ gives that

$$\begin{aligned} \hat{P}(X) &= \frac{1}{n+1} \sum_{t=0}^{n} \frac{\frac{1}{N} \sum_{i=1}^N \left(\Pi_{s \le t} P_i(X_s) \right) P_i(X)}{\frac{1}{N} \sum_{i=1}^N \Pi_{s \le t} P_i(X_s)} \\ &\in \operatorname{conv}(\mathcal{P}) \end{aligned}$$

The idea in the construction of $\hat{P}$ is that we're weighting the distribution that assigns more weight to the observed Xs by more. We use the convention that $P_i(X_0) := 1 \, \forall i \in [N]$ since X_0 is not actually a datapoint and its presence in the equation is just for brevity. We have that for any $P \in \mathcal{P}$

$$\mathbb{E}_P \left[D_{KL}(P||\hat{P}) \right] \le \frac{1}{n+1} D_{KL}(P^{\otimes(n+1)}||Q_{X^{n+1}})$$

Proof.

For any $P \in \mathcal{P}$,

$$\begin{aligned}\mathbb{E}_P\left[D_{KL}(P||\hat{P})\right] &= \mathbb{E}_P\left[D_{KL}\left(P\middle\|\frac{1}{n+1}\sum_{t=0}^{n}Q_{X_{t+1}|X^t}\right)\right] \\ &\leq \frac{1}{n+1}\sum_{t=0}^{n}\mathbb{E}_P\left[D_{KL}\left(P||Q_{X_{t+1}|X^t}\right)\right], \text{ by Property 18} \\ &= \frac{1}{n+1}D_{KL}\left(P^{\otimes(n+1)}||Q_{X^{n+1}}\right), \text{ by Property 19}\end{aligned}$$

□

Theorem 166 (Yang-Barron (KL))**.** Take the context from the start of Section 10. We have that there exists an estimator $\hat{P} \in \mathrm{conv}(\mathcal{P})$ for P such that

$$\sup_{P\in\mathcal{P}}\mathbb{E}_P[D_{KL}(P||\hat{P})] \precsim \inf_{\epsilon>0}\epsilon^2 + \frac{1}{n}\log(N_{KL}(\mathcal{P},\epsilon))$$

We note that this "online-to-batch" strategy provides a general paradigm for converting a redundancy bound to prediction risk bound.[a] We also note that the estimated $\hat{P}$ as defined in Lemma 165 is *improper* in the sense that $P \in \mathrm{conv}(\mathcal{P})$ but potentially $\hat{P} \notin \mathcal{P}$. Further $\hat{P}$ might be computationally difficult to obtain.

Proof.

Take any $\epsilon > 0$. By assumption, we can construct a minimal ϵ-KL covering of $\mathcal{P}$. Define $Q_{X^{n+1}}$ as in Lemma 164 and $\hat{P}$ as in Lemma 165. Applying Lemma 165 and Lemma 164 in turn, we have that

$$\begin{aligned}\mathbb{E}_P\left[D_{KL}(P||\hat{P})\right] &\leq \frac{1}{n+1}D_{KL}(P^{\otimes(n+1)}||Q_{X^{n+1}}) \\ &\leq \frac{1}{n+1}\left((n+1)\epsilon^2 + \log(N_{KL}(\mathcal{P},\epsilon))\right) \\ &= \epsilon^2 + \frac{1}{n+1}\log(N_{KL}(\mathcal{P},\epsilon))\end{aligned}$$

In light of the arbitrariness of ϵ and the fact that $\frac{1}{n} \asymp \frac{1}{n+1}$, we get the result. □

[a] As we will see in Definition 182, we define the minimax redundancy of a distribution class $\mathcal{P}$ on X^n as $\mathrm{Red}(\mathcal{P}) = \inf_{Q_{X^n}}\sup_{P_{X^n}\in\mathcal{P}}D_{KL}(P_{X^n}||Q_{X^n})$. We bound this quantity in Lemma 164.

10.2 Yatrocos with TV

In this section, we will show an upper bound on density estimation that depends on TV-distance thanks to Yatrocos.

Lemma 167 (Hoeffding's Lemma 3)**.** Let $Z^n \sim P^{\otimes n}$ with $\mathrm{supp}(Z_i) \in [a,b]$ for $i \in [n]$ with $a < b \in \mathbb{R}$. Then, we have that $\forall t \geq 0$

$$\Pr\left(\left|\mathbb{E}[Z_i] - \frac{1}{n}\sum_{i=1}^{n}Z_i\right| \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Theorem 168 (Minimum Distance Estimator)**.** Take the context in the start of Section 10. Let $Q_1, ..., Q_N$ be arbitrary candidate distributions. Then, there exists an estimator $\hat{P}$ such that

$$TV(P,\hat{P}) \leq 3 \min_{i\in[N]} TV(P,Q_i) + \epsilon_n, \text{ with } \mathbb{E}[\epsilon_n^2] = O\left(\frac{\log(N)}{n}\right)$$

Proof.

We will prove the theorem using a minimum distance estimator

$$\hat{P} = \underset{Q\in\{Q_1,\ldots,Q_N\}}{\arg\min} \widetilde{TV}(P_n,Q)$$

where $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution and $\widetilde{TV}$ is a pseudo-distance. We cannot use $\widetilde{TV} = TV$ since if $Q_1,...,Q_N$ are all continuous distributions, since P_n is discrete, $TV(P_n,Q_i) = 1\ \forall i \in [N]$ so it's not useful. Let's defer the choice of $\widetilde{TV}$ for now and proceed to the analysis.

Let $Q^* := \arg\min_{Q\in\{Q_1,\ldots,Q_N\}} TV(P,Q)$. Then, we have that

$$\begin{aligned}
TV(\hat{P},P) &\leq TV(\hat{P},Q^*) + TV(Q^*,P), \text{ by the Triangle inequality}\\
&\overset{\text{hope}}{=} \widetilde{TV}(\hat{P},Q^*) + TV(Q^*,P), \text{ if (2) below holds}\\
&\leq \widetilde{TV}(\hat{P},P_n) + \widetilde{TV}(P_n,Q^*) + TV(Q^*,P), \text{ by the Triangle inequality}\\
&\leq 2\widetilde{TV}(P_n,Q^*) + TV(Q^*,P), \text{ by the definition of } \hat{P} \text{ and } Q^*\\
&\leq 2\widetilde{TV}(P_n,P) + 2\widetilde{TV}(P,Q^*) + TV(P,Q^*), \text{ by the Triangle inequality}\\
&\overset{\text{hope}}{\leq} 2\widetilde{TV}(P_n,P) + 3TV(P,Q^*), \text{ if (1) below holds}
\end{aligned}$$

To make this above analysis go through (ie., (1) and (2)) and have the ability to reach the desired result (ie., (3)), we need that

(1) $\widetilde{TV}(P,Q) \leq TV(P,Q)\ \forall P,Q$

(2) $\widetilde{TV}(Q_i,Q_j) = TV(Q_i,Q_j)\ \forall i,j \in [N]$

(3) $\mathbb{E}\left[\left(\widetilde{TV}(P_n,P)\right)^2\right] = O\left(\frac{\log(N)}{n}\right)$

Motivated by (1) and (2), we define

$$\begin{aligned}
&\widetilde{TV}(P,Q) := \sup_{A\in\mathcal{A}} |P(A) - Q(A)|\\
&\text{with } \mathcal{A} := \{A_{ij} : i,j \in [N]\}, A_{ij} := \{x : Q_i(x) \geq Q_j(x)\}
\end{aligned}$$

(1) is immediate since $TV(P,Q) = \sup_A |P(A) - Q(A)|$ so that TV takes the supremum over a larger set of events. (2) is true since $TV(Q_i,Q_j) = \max\{Q_i(A_{ij}) - Q_j(A_{ij}), Q_j(A_{ji}) - Q_i(A_{ji})\} \leq \widetilde{TV}(Q_i,Q_j)$, which we can combine with the conclusion from (1) to get the desired result. As for (3), we note that $|\mathcal{A}| \leq \binom{N}{2}$ and for fixed A and any $\epsilon > 0$,

$$\begin{aligned}
&\Pr\left(|P(A) - P_n(A)| > \epsilon\right) \leq \exp\left(-2n\epsilon^2\right), \text{ by Lemma 167}\\
\implies &\Pr\left(\widetilde{TV}(P,P_n) > \epsilon\right) \leq 2N^2\exp\left(-2n\epsilon^2\right), \text{ by a Union bound and that } \binom{N}{2} \leq N^2\\
\implies &\Pr\left(\left(\widetilde{TV}(P,P_n)\right)^2 > \epsilon\right) < 2N^2\exp\left(-2n\epsilon^2\right), \text{ since } \widetilde{TV}(P,P_n) \geq 0
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
\mathbb{E}\left[\left(\widetilde{TV}(P,P_n)\right)^2\right] &= \int_0^\infty \Pr\left(\left(\widetilde{TV}\left(P,P_n\right)\right)^2 \geq r\right) dr, \text{ since } \left(\widetilde{TV}\left(P,P_n\right)\right)^2 \geq 0 \\
&\leq \int_0^\infty \min\left(1, 2N^2\exp\left(-2nr\right)\right) dr, \text{ by above} \\
&= \frac{\log\left(2N^2\right)}{2n} + \int_{\frac{\log\left(2N^2\right)}{2n}}^\infty 2N^2\exp\left(-2nr\right) dr \\
&= \frac{\log\left(2N^2\right)}{2n} + \left[-\frac{2N^2}{2n}\exp\left(-2nr\right)\right]_{\frac{\log\left(2N^2\right)}{2n}}^\infty \\
&= O\left(\frac{\log(N)}{n}\right)
\end{aligned}$$

Lemma 169 (Cauchy-Schwarz Implication)**.** For $a,b\in\mathbb{R}$, we have that $(a+b)^2 \leq 2a^2+2b^2$.

Proof.

Notice that $2ab \leq a^2+b^2$ implies the result so that it suffices to show this. To that end, we have that $(a-b)^2 \geq 0 \implies a^2+b^2 \geq 2ab$. □

Corollary 170 (Yatrocos: Minimum Distance Estimator)**.** Take the context from the start of Section 10. We have that there exists an estimator $\hat{P}\in\mathcal{P}$ such that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\left(TV(P,\hat{P})\right)^2\right] \precsim \inf_{\epsilon>0} \epsilon^2 + \frac{1}{n}\log\left(N_{KL}(\mathcal{P},\epsilon)\right)$$

Proof.

Take any $\epsilon > 0$ and any $P \in \mathcal{P}$. By assumption, we can construct a minimal ϵ-TV covering of $\mathcal{P}$ given by $\{Q_1,...,Q_N\}$ with $N := N_{TV}(\mathcal{P},\epsilon)$. From Theorem 168 and it's $\hat{P}\in\mathcal{P}$, we have that

$$TV(P,\hat{P}) \leq \underbrace{3\min_{i\in[N]} TV(P,Q_i)}_{=:TV(P,Q^*)} + \epsilon_n, \text{ with } \mathbb{E}\left[\epsilon_n^2\right] = O\left(\frac{\log\left(N_{TV}(\mathcal{P},\epsilon)\right)}{n}\right)$$

$$\begin{aligned}
&\implies (TV(P,\hat{P}))^2 \leq 6\left(TV(P,Q^*)\right)^2 + 2\epsilon_n^2, \text{ by Lemma 169} \\
&\implies \left(TV(P,\hat{P})\right)^2 \leq 6\epsilon^2 + 2\epsilon_n^2, \text{ by } \epsilon\text{-TV covering} \\
&\implies \mathbb{E}\left[\left(TV(P,\hat{P})\right)^2\right] \precsim \epsilon^2 + \frac{\log(N_{TV}(\mathcal{P},\epsilon))}{n}
\end{aligned}$$

In light of the arbitrariness of ϵ and $P\in\mathcal{P}$, we have the result. □

10.3 Le Cam-Birge with H^2

In this section, we will show an upper bound on density estimation that depends on H^2-divergence thanks to Le Cam and Birge.

Consider a composite hypothesis testing situation. We test between

$$\begin{aligned}
H_0 &: X^n \sim P \text{ for } P\in\mathcal{P} \\
H_1 &: X^n \sim Q \text{ for } Q\in\mathcal{Q}
\end{aligned}$$

using a test $T := T(X^n) \in \{0, 1\}$. We define the errors by

- Type I error: $\sup_{P\in\mathcal{P}} P^{\otimes n}(T = 1)$
- Type II error: $\sup_{Q\in\mathcal{Q}} Q^{\otimes n}(T = 0)$

Definition 171 (Squared-Hellinger Distance on Family of Distributions)**.** For a family of distributions $\mathcal{P}$ and $\mathcal{Q}$ on the same ground space $\mathcal{X}$, we define

$$H^2\left(\mathrm{conv}(\mathcal{P}), \mathrm{conv}(\mathcal{Q})\right) := \inf_{P\in\mathrm{conv}(\mathcal{P}), Q\in\mathrm{conv}(\mathcal{Q})} H^2(P, Q)$$

where we have abused notation to let H^2 on the RHS to be the Squared-Hellinger distance between two distributions defined in Section 3.1.

Lemma 172 (H^2 and Convolution of Families of Distributions)**.** Consider families of distributions $\mathcal{P}_1, ..., \mathcal{P}_n$ and $\mathcal{Q}_1, ..., \mathcal{Q}_n$ on the ground space $\mathcal{X}$. We have that

$$1 - \frac{1}{2}H^2\left(\mathrm{conv}(\otimes_{i=1}^n \mathcal{P}_i), \mathrm{conv}(\otimes_{i=1}^n \mathcal{Q}_i)\right) \leq \Pi_{i=1}^n \left(1 - \frac{1}{2}H^2\left(\mathrm{conv}(\mathcal{P}_i), \mathrm{conv}(\mathcal{Q}_i)\right)\right)$$

Proof.

By a simple inductive argument, it suffices to prove the case of $n = 2$. As a first remark, observe that

$$\begin{aligned} 1 - \frac{1}{2}H^2(P, Q) &= 1 - \frac{1}{2}\int_{x\in\mathcal{X}} \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 dx \\ &= \int_{x\in\mathcal{X}} \sqrt{P(x)Q(x)}dx \end{aligned}$$

As a second remark, observe that any $P_{XY} \in \mathrm{conv}(\mathcal{P}_1 \otimes \mathcal{P}_2)$ can be written as $P_{XY} = \mathbb{E}_Z\left[P_{X|Z}P_{Y|Z}\right]$ with $P_{X|Z} \in \mathcal{P}_1$ and $P_{Y|Z} \in \mathcal{P}_2$. Then, taking $P_{XY}, Q_{XY} \in \mathrm{conv}(\mathcal{P}_1 \otimes \mathcal{P}_2) \times \mathrm{conv}(\mathcal{Q}_1 \otimes \mathcal{Q}_2)$ we have that

$$\begin{aligned} 1 - \frac{1}{2}H^2(P_{XY}, Q_{XY}) &= \int_{(x,y)\in\mathcal{X}^2} \sqrt{P_{XY}(x,y)Q_{XY}(x,y)}dxdy \\ &= \int_{x\in\mathcal{X}} \sqrt{P_X(x)Q_X(x)} \int_{y\in\mathcal{X}} \sqrt{\underbrace{P_{Y|X}(y \mid x)}_{\in\mathrm{conv}(\mathcal{P}_2)}\underbrace{Q_{Y|X}(y \mid x)}_{\in\mathrm{conv}(\mathcal{Q}_2)}}dydx \\ &\leq \int_{x\in\mathcal{X}} \sqrt{\underbrace{P_X(x)}_{\in\mathrm{conv}(\mathcal{P}_1)}\underbrace{Q_X(x)}_{\in\mathrm{conv}(\mathcal{Q}_1)}}dx \left(1 - \frac{1}{2}H^2(\mathrm{conv}(\mathcal{P}_2), \mathrm{conv}(\mathcal{Q}_2)\right), \\ &\quad \text{by Defintion 171} \\ &\leq \left(1 - \frac{1}{2}H^2(\mathrm{conv}(\mathcal{P}_1), \mathrm{conv}(\mathcal{Q}_1)\right)\left(1 - \frac{1}{2}H^2(\mathrm{conv}(\mathcal{P}_2), \mathrm{conv}(\mathcal{Q}_2)\right), \\ &\quad \text{by Definition 171} \end{aligned}$$

□

Lemma 173 (Bound on Sum of Errors for Composite Test)**.** Take the context at the start of Section 10.3. We have

that

$$\inf_T \sup_{P\in\mathcal{P}} P^{\otimes n}(T=1) + \sup_{Q\in\mathcal{Q}} Q^{\otimes n}(T=0) \le \exp\left(-\frac{n}{2}H^2(\operatorname{conv}(\mathcal{P}),\operatorname{conv}(\mathcal{Q}))\right)$$

Proof.

Analogous to Definition 171, for families of distributions $\mathcal{P}$, $\mathcal{Q}$ on the same ground space $\mathcal{X}$,

$$TV(\operatorname{conv}(\mathcal{P}),\operatorname{conv}(\mathcal{Q})) = \inf_{P\in\operatorname{conv}(\mathcal{P}),Q\in\operatorname{conv}(\mathcal{Q})} TV(P,Q)$$

By Theorem 126, we have that the LHS of the lemma statement is equal to

$$\begin{aligned}
\text{LHS} &= 1 - TV\left(\operatorname{conv}(\mathcal{P}^{\otimes n}),\operatorname{conv}(\mathcal{Q}^{\otimes n})\right)\\
&\le 1-\frac{1}{2}H^2\left(\operatorname{conv}(\mathcal{P}^{\otimes n}),\operatorname{conv}(\mathcal{Q}^{\otimes n})\right),\ \text{since } TV \ge \frac{H^2}{2} \text{ by Section 3.8}\\
&\le \left(1-\frac{1}{2}H^2\left(\operatorname{conv}(\mathcal{P}),\operatorname{conv}(\mathcal{Q})\right)\right)^n\\
&= \left(1-\frac{\frac{n}{2}H^2\left(\operatorname{conv}(\mathcal{P}),\operatorname{conv}(\mathcal{Q})\right)}{n}\right)^n\\
&\le \exp\left(-\frac{n}{2}H^2\left(\operatorname{conv}(\mathcal{P}),\operatorname{conv}(\mathcal{Q})\right)\right)
\end{aligned}$$

where the last step follows since $x\mapsto\left(1+\frac{1}{x}\right)^x$ is an increasing function of x for $x\in\mathbb{R}_{++}$ and $\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x = e$. □

Corollary 174 (Bound on Sum of Errors for Hellinger Separated Composite Test)**.** Take the context from the start of Section 10.3 but with, for some reference distributions P_0, Q_0 and any $\epsilon > 0$

$$\begin{aligned}
\mathcal{P} &:= B_H(P_0,\epsilon) := \{P : H^2(P,P_0)\le\epsilon^2\}\\
\mathcal{Q} &:= B_H(Q_0,\epsilon)
\end{aligned}$$

If $H(P_0,Q_0)\ge 4\epsilon$, then

$$\inf_T \sup_{P\in B_H(P_0,\epsilon)} P^{\otimes n}(T=1) + \sup_{Q\in B_H(Q_0,\epsilon)} Q^{\otimes n}(T=0) \le \exp\left(-\frac{n}{8}H^2(P_0,Q_0)\right)$$

Proof.

Since $(P,Q)\mapsto H^2(P,Q)$ is jointly convex (ie., Property 36), we have that both balls $B_H(P_0,\epsilon)$ and $B_H(Q_0,\epsilon)$ are convex. Therefore, we have that

$$\begin{aligned}
H(B_H(P_0,\epsilon),B_H(Q_0,\epsilon)) &= \inf_{P,Q\in B_H(P_0,\epsilon)\times B_H(Q_0,\epsilon)} H(P,Q)\\
&\ge \inf_{P,Q\in B_H(P_0,\epsilon)\times B_H(Q_0,\epsilon)} H(P_0,Q_0)-H(P,P_0)-H(Q,Q_0),\\
&\quad \text{using Triangle inequality on Hellinger distance}\\
&= H(P_0,Q_0)-2\epsilon\\
&\ge \frac{1}{2}H(P_0,Q_0)
\end{aligned}$$

Applying Lemma 173, we have that

$$\begin{aligned}\inf_T \sup_{P\in B_H(P_0,\epsilon)} P^{\otimes n}(T=1) + \sup_{Q\in B_H(Q_0,\epsilon)} Q^{\otimes n}(T=0) &\leq \exp\left(-\frac{n}{2}H^2(B_H(Q_0,\epsilon), B_H(Q_0,\epsilon))\right)\\ &\leq \exp\left(-\frac{n}{8}H^2\left(P_0,Q_0\right)\right)\end{aligned}$$

□

Theorem 175 (Pairwise Comparison Estimator)**.** Take the context from the start of Section 10. For any sequence $\epsilon_n > 0$ that satisfies

$$n\epsilon_n^2 > \max\{\log(N_H(\mathcal{P},\epsilon_n)), 1\}$$

there exists an estimator $\hat{P} \in \mathcal{P}$ that satisfies

$$\sup_{P\in\mathcal{P}} \Pr_P\left(H(P,\hat{P}) > 4t\epsilon_n\right) \leq C\cdot\exp\left(-t^2\right)$$

$\forall t \geq 1$ and some universal $C > 0$. Consequently, $\sup_{P\in\mathcal{P}} \mathbb{E}\left[H^2(P,\hat{P})\right] = O\left(\epsilon_n^2\right)$.

Proof.

Take any $\epsilon > 0$ (along the sequence $(\epsilon_n)_{n\in\mathbb{N}}$, which satisfies that condition). Let $P_1, ..., P_N$ be a maximal ϵ-packing of $\mathcal{P}$ under H (ie., $H(P_i,P_j) \geq \epsilon\ \forall i \neq j \in [N]$). Since a maximal ϵ-packing is also an ϵ-covering[a], which means that $\sup_{P\in\mathcal{P}} \min_{i\in[N]} H(P,P_i) \leq \epsilon$. For $\delta := 4\epsilon$ and any $i,j \in [N]$ such that $H(P_i,P_j) \geq \delta$, construct a test T_{ij} for

$$\begin{aligned}H_0 &: P \in B_H(P_i,\epsilon)\\ H_1 &: P \in B_H(P_j,\epsilon)\end{aligned}$$

By Corollary 174, $\exists T_{ij}$ such that for any $i \neq j \in [N]$

$$\sup_{P\in B_H(P_i,\epsilon)} \Pr_P(T_{ij}=1) \leq \exp\left(-\frac{n}{8}\left(H(P_i,P_j)\right)^2\right)$$

Restrict so that $T_{jj} := 1 - T_{ij}$ for any $i \neq j \in [N]$, which still satisfies the condition. Now, define the following estimator. For $i \in [N]$, let

$$\begin{aligned}&\psi_i := \max\left(\{H(P_i,P_j) : T_{ij}(X^n) = 1, H(P_i,P_j) \geq \delta\} \cup \{0\}\right)\\ &\hat{P} := P_{i^*},\ \text{with } i^* := \arg\min_{i\in[N]} \psi_i\end{aligned}$$

so that $\hat{P} \in \mathcal{P}$. Now, consider any $P \in \mathcal{P}$. Since $\{P_1, ..., P_N\}$ is an ϵ-covering, WLOG (relabeling if necessary), assume that $H(P,P_1) \leq \epsilon$. Thus, we have that for any $t \geq 1$,

$$\begin{aligned}\mathbb{1}_{\{H(\hat{P},P_1)\geq t\delta\}} &= \mathbb{1}_{\{H(P_{i^*},P_1)\geq t\delta\}}\\ &\leq \mathbb{1}_{\{\max\{\psi_{i^*},\psi_1\}\geq t\delta\}},\\ &\quad \text{since } T_{1i^*}(X^n) = 1 \text{ or } T_{i^*1}(X^n) = 1 \text{ and } H(\hat{P},P_1) \geq t\delta \implies \max\{\psi_{i^*},\psi_1\} \geq t\delta\\ &= \mathbb{1}_{\{\psi_1\geq t\delta\}},\ \text{since } \psi_{i^*} \leq \psi_1\\ &\leq \mathbb{1}_{\{\vee_{j\in[N]\setminus\{i\}} H(P_1,P_j)\geq t\delta\ \wedge\ T_{1j}=1\}}\end{aligned}$$

Thus, by a union bound,

$$\begin{aligned}\Pr_P(H(\hat{P}, P_1) \geq t\delta) &\leq \Pr_P(\vee_{j\in[N]\setminus\{i\}} H(P_1, P_j) \geq t\delta \ \wedge\ T_{1j} = 1)\\ &\leq \sum_{j\in[N]\setminus\{i\}} \Pr_P(H(P_1, P_j) \geq t\delta \ \wedge\ T_{1j} = 1)\\ &\leq N \exp\left(-\frac{n}{8}(H(P_1, P_j))^2\right), \text{ by above with } P \in B_H(P_1, \epsilon)\\ &= N_H(\mathcal{P}, \epsilon) \exp\left(-2nt^2\epsilon^2\right)\end{aligned}$$

where we replace N by $N_H(\mathcal{P}, \epsilon)$ leveraging Lemma 145. Since $n\epsilon^2 > \max\{1, \log(N_H(\mathcal{P}, \epsilon))\}$, we have that

$$\Pr_P(H(\hat{P}, P_1) \geq t\delta) \leq \exp(-2)\exp(-t^2)$$

Finally, note that $\Pr_P(H(\hat{P}, P) \geq t\delta) \leq \Pr_P(H(\hat{P}, P_1) \geq t\delta - \epsilon) \leq \Pr_P(H(\hat{P}, P_1) \geq \frac{3}{4}t\delta)$ since $H(\hat{P}, P) \leq H(\hat{P}, P_1) + \underbrace{H(P_1, P)}_{\leq\epsilon}$ and $\epsilon < \delta, t \geq 1$. As a result, with a different universal constant $C' > 0$ that's chosen to account for the behavior with $t \in [1, 4/3)$, we can say that $\Pr_P(H(\hat{P}, P) \geq t\delta) = C' \exp(-t^2)\ \forall t \geq 1$.

To see the second result, recall that for a non-negative random variable $Z \sim P$ we have that $\mathbb{E}_P[Z^2] = \int_{z\in\operatorname{supp} Z} 2zP(Z > z)dz$. Thus, we have that for any $P \in \mathcal{P}$ and any ϵ_n in the sequence $(\epsilon_n)_{n\in\mathbb{N}}$

$$\begin{aligned}\mathbb{E}\left[\left(H(P, \hat{P})\right)^2\right] &= \int_{\mathbb{R}_+} 2u \Pr_P\left(H(P, \hat{P}) > u\right) du\\ &\leq \int_0^{4\epsilon_n} 1 \cdot 2udu + \int_{4\epsilon_n}^{\infty} 2uC \exp\left(-\frac{u^2}{16\epsilon_n^2}\right) du, \text{ for some } C > 0\\ &= 16\epsilon_n^2 + 32\epsilon_n^2 C \underbrace{\int_1^{\infty} t\exp(-t^2)dt}_{<\infty}\\ &= O(\epsilon_n^2)\end{aligned}$$

In light of the arbitrariness of $P \in \mathcal{P}$, we have both results. □

[a] Suppose that $\{P_1, ..., P_N\}$ is a maximal ϵ-packing of $\mathcal{P}$ but not an ϵ-covering. Then, there exists a distribution P^* which is not in $\cup_{i=1}^N B(P_i, \epsilon)$. But then, $\{P_1, ..., P_N\} \cup \{P^*\}$ is an ϵ-packing of strictly larger size, which contradicts the original maximal packing.

Corollary 176 (Le Cam-Birge: Pairwise Comparison Estimator)**.** Take the context from the start of Section 10. We have that there exists an estimator $\hat{P} \in \mathcal{P}$ such that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\left[\left(H(P, \hat{P})\right)^2\right] \precsim \inf_{\epsilon>0} \epsilon^2 + \frac{1}{n}\log\left(N_H(\mathcal{P}, \epsilon)\right)$$

Proof.

By Theorem 175, we have that any for any $(\epsilon_n)_{n\in\mathbb{N}}$ with $\epsilon_n > 0\ \forall n \in \mathbb{N}$ satisfying

$$n\epsilon_n^2 > \max\{\log(N_H(\mathcal{P}, \epsilon_n)), 1\}$$

there exists an estimator $\hat{P} \in \mathcal{P}$ such that

$$\sup_{P\in\mathcal{P}} \mathbb{E}\left[H^2(P, \hat{P})\right] = O(\epsilon_n^2)$$

Now, consider any fixed $\epsilon > 0$. Trivially, note that $\epsilon^2 \precsim \epsilon^2 + \frac{1}{n}\log(N_H(\mathcal{P}, \epsilon))$. For n big enough, we will have that $n\epsilon^2 > \max\{\log(N_H(\mathcal{P}, \epsilon)), 1\}$. Thus, by transitivity we have the result. □

10.4 Refinement via Local Entropy

It turns out that the *global* entropy $\log(N_H(\mathcal{P}), \epsilon)$ can be improved to a *local* entropy $\log(N_{loc}(\mathcal{P}, \epsilon))$ in the setting of Corollary 176.

Definition 177 (Local Entropy of Family of Distributions)**.** Consider a Family of distributions $\mathcal{P}$ and take any $\epsilon > 0$. We define the *local* entropy of the family of distributions: $\log(N_{loc}(\mathcal{P}, \epsilon))$, with

$$N_{loc}(\mathcal{P}, \epsilon) := \sup_{P \in \mathcal{P}} \sup_{\eta \geq \epsilon} N_H\left(B_H(P, \eta) \cap \mathcal{P}, \frac{\eta}{2}\right)$$

In other words, we search for the distribution in the family so that it's hard to cover a ball around the distribution with balls of half the radius.

Example 178 (Improvement from Local Entropy)**.** For many d-dimensional families $\mathcal{P}$, we typically have that for any $\epsilon > 0$

$$\log(N_H(\mathcal{P}, \epsilon)) \asymp d \log(1/\epsilon)$$

whereas

$$\log(N_{loc}(\mathcal{P}, \epsilon)) \asymp d$$

so that if we can replace the global entropy by local entropy in Corollary 176, we will have an asymptotic improvement on the guarantee.

Lemma 179 (Local Entropy Lemma)**.** Take any family of distributions $\mathcal{P}$, any $\epsilon > 0$, any $\eta \geq \epsilon$, and any fixed $P \in \mathcal{P}$. We have that

$$N_H(B_H(P, 2^k\eta) \cap \mathcal{P}, \eta/2) \leq N_{loc}(\mathcal{P}, \epsilon)^{k+1}$$

for any $k \in \{0\} \cup \mathbb{N}$.

Proof.

We proceed by induction on $k \in \{0\} \cup \mathbb{N}$. The result holds trivially for $k = 0$ by the definition of $N_{loc}(\mathcal{P}, \epsilon)$ in Definition 177. For the inductive step, suppose the statement holds for some $k - 1 \in \{0\} \cup \mathbb{N}$, we wish to show it holds for k.

To that end, consider $B_H(P, 2^k\eta)$ and cover it using balls of radius $2^{k-1}\eta$. Then, cover each of those balls with small balls of size $\frac{\eta}{2}$. We then have that

$$\begin{aligned}
N_H\left(B_H(P, 2^k\eta) \cap \mathcal{P}, \frac{\eta}{2}\right) &\leq N_H\left(B_H(P, \underbrace{2^k\eta}_{=:\delta}) \cap \mathcal{P}, 2^{k-1}\eta\right) \cdot N_H\left(B_H(P, 2^{k-1}\eta), \frac{\eta}{2}\right) \\
&\leq N_H\left(B_H(P, \delta) \cap \mathcal{P}, \frac{\delta}{2}\right) \cdot N_H\left(B_H(P, 2^{k-1}\eta) \cap \mathcal{P}, \frac{\eta}{2}\right) \\
&\leq N_{loc}(\mathcal{P}, \epsilon) \cdot (N_{loc}(\mathcal{P}, \epsilon))^{k-1}, \text{ by the inductive hypothesis and since } \delta \geq \epsilon \\
&= (N_{loc}(\mathcal{P}, \epsilon))^k
\end{aligned}$$

Theorem 180 (Local Entropy: Pairwise Comparison Estimator)**.** Take the context from the start of Section 10. For any sequence $\epsilon_n > 0$ that satisfies

$$n\epsilon_n^2 > \max\{\log(N_H(\mathcal{P}, \epsilon_n)), 1\}$$

there exists an estimator $\hat{P} \in \mathcal{P}$ that satisfies

$$\sup_{P \in \mathcal{P}} \Pr_P \left(H(P, \hat{P}) > 4t\epsilon_n \right) \leq C \cdot \exp\left(-t^2\right)$$

$\forall t \geq 1$ and some universal $C > 0$. Consequently, $\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[H^2(P, \hat{P}) \right] = O(\epsilon_n^2)$. There also exists a corollary to this result analogous to Corollary 176 that has an identical proof (and so is omitted). Within the context of this theorem, the statement of the corollary is that $\exists \hat{P} \in \mathcal{P}$ such that $\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[H^2(P, \hat{P}) \right] \lesssim \epsilon^2 + \frac{1}{n} \log(N_H(\mathcal{P}, \epsilon))$.

Proof.

Take[a] any $\epsilon > 0$ (along the sequence $(\epsilon_n)_{n \in \mathbb{N}}$, which satisfies that condition). Let $P_1, ..., P_N$ be a maximal ϵ-packing of $\mathcal{P}$ under H (ie., $H(P_i, P_j) \geq \epsilon \; \forall i \neq j \in [N]$). Since a maximal ϵ-packing is also an ϵ-covering[b], which means that $\sup_{P \in \mathcal{P}} \min_{i \in [N]} H(P, P_i) \leq \epsilon$. For $\delta := 4\epsilon$ and any $i, j \in [N]$ such that $H(P_i, P_j) \geq \delta$, construct a test T_{ij} for

$$\begin{aligned} H_0 &: P \in B_H(P_i, \epsilon) \\ H_1 &: P \in B_H(P_j, \epsilon) \end{aligned}$$

By Corollary 174, $\exists T_{ij}$ such that for any $i \neq j \in [N]$

$$\sup_{P \in B_H(P_i, \epsilon)} \Pr_P(T_{ij} = 1) \leq \exp\left(-\frac{n}{8} \left(H(P_i, P_j) \right)^2 \right)$$

Restrict so that $T_{jj} := 1 - T_{ij}$ for any $i \neq j \in [N]$, which still satisfies the condition. Now, define the following estimator. For $i \in [N]$, let

$$\begin{aligned} &\psi_i := \max\left(\{ H(P_i, P_j) : T_{ij}(X^n) = 1, H(P_i, P_j) \geq \delta \} \cup \{0\} \right) \\ &\hat{P} := P_{i^*}, \text{ with } i^* := \arg\min_{i \in [N]} \psi_i \end{aligned}$$

so that $\hat{P} \in \mathcal{P}$. Now, consider any $P \in \mathcal{P}$. Since $\{P_1, ..., P_N\}$ is an ϵ-covering, WLOG (relabeling if necessary), assume that $H(P, P_1) \leq \epsilon$. Also consider $t \geq 1$. Let $l \in \{0\} \cup \mathbb{N}$ be such that $2^l \leq t < 2^{l+1}$. Define $A_k := \{ j \in [N] : 2^k \leq H(P_1, P_j) < 2^{k+1}\delta \}$ so that

$$\{ j \in [N] : H(P_1, P_j) \geq t\delta \} \subseteq \cup_{k \geq l} A_k$$

We have that

$$\begin{aligned} \Pr_P(H(\hat{P}, P_1) \geq t\delta) &\leq \Pr_P(\psi_1 \geq t\delta) \\ &\leq \sum_{k \geq l} \Pr_P \left(2^k\delta \leq \psi_1 < 2^{k+1}\delta \wedge T_{1j} = 1 \right) \\ &\leq \sum_{k \geq l} |A_k| \exp\left(-\frac{n}{8} \left(2^k \delta \right)^2 \right), \text{ by above with } P \in B_H(P_1, \epsilon) \end{aligned}$$

To upper bound $|A_k|$, since $\{P_1, ..., P_N\}$ is an ϵ-packing of $\mathcal{P}$, defining the auxiliary set $O := \{ \tilde{P} \in \mathcal{P} : 2^k\delta \leq$

$H(P_1, \tilde{P}) < 2^{k+1}\delta\}$, we have that

$$\begin{aligned}
|A_k| &\leq M\left(O, H(\cdot,\cdot), \epsilon\right) \\
&\leq M\left(B_H(P_1, 2^{k+1}\delta) \cap \mathcal{P}, H(\cdot,\cdot), \epsilon\right), \text{ since } O \subseteq B_H(P_1, 2^{k+1}\delta) \cap \mathcal{P} \\
&\leq N(B_H(P_1, 2^{k+1}\delta) \cap \mathcal{P}, H(\cdot,\cdot), \epsilon/2), \text{ by Lemma 145} \\
&= N(B_H(P_1, 2^{k+3}\epsilon) \cap \mathcal{P}, H(\cdot,\cdot), \epsilon/2), \text{ since } \delta = 4\epsilon \\
&\leq \left(N_{loc}(\mathcal{P}, \epsilon)\right)^{k+4}, \text{ by Lemma 179}
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
\Pr_P(H(\hat{P}, P_1) \geq t\delta) &\leq \sum_{k \geq l} \exp\left[(k+4)\log(N_{loc}(\mathcal{P}, \epsilon)) - 2n\epsilon^2 4^k\right] \\
&\leq \sum_{k \geq l} \exp\left[\left((k+4) - 2 \cdot 4^k\right)\left(\log(N_{loc}(\mathcal{P}, \epsilon))\right)\right], \\
&\qquad \text{assuming } n\epsilon^2 > \max\{1, \log(N_{loc}(\mathcal{P}, \epsilon))\} \\
&\leq \exp\left(-\Omega_l(4^l)\right) \\
&= \exp(-\Omega_t(t^2)), \text{ since } t \in [2^l, 2^{l+1}) \\
&= O_t\left(\exp(-t^2)\right) \\
\implies \Pr_P(H(\hat{P}, P_1) \geq t\delta) &\leq C\exp(-t^2)
\end{aligned}$$

for some $C > 0$ and all $t \geq 1$. We defer the remaining parts of the proof to the proof of Theorem 175 as the arguments are totally identical.[c] □

[a] A lot of this proof is similar to the proof of Theorem 175.

[b] Suppose that $\{P_1, ..., P_N\}$ is a maximal ϵ-packing of $\mathcal{P}$ but not an ϵ-covering. Then, there exists a distribution P^* which is not in $\cup_{i=1}^{N} B(P_i, \epsilon)$. But then, $\{P_1, ..., P_N\} \cup \{P^*\}$ is an ϵ-packing of strictly larger size, which contradicts the original maximal packing.

[c] Start from "Finally, note that $\Pr_P(H(\hat{P}, P) \geq t\delta) \leq$ ...".

11 Universal Compression and Redundancy

Recall Theorem 6. If we have that $X^n \overset{iid}{\sim} \mathcal{P}$ with support in $\mathcal{X}$, there exists a uniquely decodable code $f : \mathcal{X} \to \{0,1\}^*$ such that $\mathbb{E}_{P^{\otimes n}}\left[l(f(X^n))\right] < H(P^{\otimes n}) + 1$. Examples of such codes include the Shannon code and Huffman code. The unideal assumption in this result is that we require perfect knowledge of P.

11.1 Universal Code with Minimal Overhead

In this section, we seek to find a universal code that is (a) uniquely decodable, (b) doesn't require the knowledge of P, and achieves an expected code-length close to $H(P)$ for every $P \in \mathcal{P}$.

Example 181 (Universal Code for iid Bernoulli)**.** Suppose that $X_1, ..., X_n \overset{iid}{\sim}$ Bern(p) with unknown $p \in [0,1]$. We have that $H(\text{Bern}(p)^{\otimes n}) = n\left(p\log\left(\frac{1}{p}\right) + (1-p)\log\left(\frac{1}{1-p}\right)\right)$, which would give us the expected code length limit (up to an additive factor of 1) if we knew p. Consider the following universal code that doesn't assume knowledge of p:

- Compute $n_1 := \sum_{i=1}^{n} \mathbb{1}_{\{X_i = 1\}}$ and express it using $\log(n+1)$ bits as $n_1 \in \{0, 1, ..., n\}$.

- Conditioning on n_1, there are $\binom{n}{n_1}$ possibilities for $(X_1, ..., X_n)$ so we can encode the final sequence using $\log\left(\binom{n}{n_1}\right)$ bits.

Clearly, this code is uniquely decodable; the decoder first decodes n_1 from the first $\log(n+1)$ bits and then decodes

$(X_1, ..., X_n)$ from n_1 and the last $\log\left(\binom{n}{n_1}\right)$ bits. The expected codelength is

$$\begin{aligned}
\mathbb{E}_{\text{Bern}(p)^{\otimes n}}\left[l(f(X^n)\right] &= \log(n+1) + \mathbb{E}_{\text{Bern}(p)^{\otimes n}}\left[\log\left(\binom{n}{n_1}\right)\right] \\
&\leq \log(n+1) + n\mathbb{E}_{\text{Bern}(p)}\left[H\left(\text{Bern}\left(\frac{n_1}{n}\right)\right)\right], \text{ see below} \\
&\leq \log(n+1) + nH\left(\mathbb{E}_{\text{Bern}(p)}\left[\text{Bern}\left(\frac{n_1}{n}\right)\right]\right), \\
&\quad \text{since } q \mapsto q\log\left(\frac{1}{q}\right) + (1-q)\log\left(\frac{1}{1-q}\right) \text{ is concave and apply Jensen's} \\
&= \log(n+1) + nH(\text{Bern}(p))
\end{aligned}$$

To see the first inequality, note that for a binomial random variable $S \sim \text{Bin}(n, q)$,

$$\begin{aligned}
1 &\geq \Pr(S = k) \\
&= \binom{n}{k} q^k (1-q)^{n-k} \\
\implies \log\left(\binom{n}{k}\right) &\leq k\log\left(\frac{1}{q}\right) + (n-k)\log\left(\frac{1}{1-q}\right) \\
&\leq nH(k/n), \text{ since selecting } q := \frac{k}{n} \text{ gives the armax of the RHS}
\end{aligned}$$

In other words, compared with the limit with the knowledge of p, there is an additional overhead of (only) $O(\log(n))$ bits.

In the general scenario, we have $X^n \sim P$ where $P \in \mathcal{P}$ is an unknown source distribution. The encoding limit with knowledge of P is $H(P)$. By Theorem 3, any uniquely decodable f can be equivalently represented by a probability distribution Q via $Q(X^n) = 2^{-l(f(X^n))}$. Therefore, the expected code-length of the code (represented by Q) under $X^n \sim P$ is given by

$$\mathbb{E}_P\left[l(f(X^n))\right] = \mathbb{E}_P\left[\log\left(\frac{1}{Q(X^n)}\right)\right]$$

and the "overhead" of a universal code is

$$\begin{aligned}
\mathbb{E}_P\left[\log\left(\frac{1}{Q(X^n)}\right)\right] - H(P) &= \mathbb{E}_P\left[\log\left(\frac{P(X^n)}{Q(X^n)}\right)\right] \\
&= D_{KL}(P||Q)
\end{aligned}$$

Definition 182 (Minimax Redundancy)**.** The minimax redundancy of a family of distributions $\mathcal{P}$ on X^n is defined as

$$\begin{aligned}
\text{Red}\left(\mathcal{P}\right) &:= \inf_{Q_{X^n}} \text{Red}\left(Q_{X^n}; \mathcal{P}\right) \\
&:= \inf_{Q_{X^n}} \sup_{P_{X^n} \in \mathcal{P}} D_{KL}\left(P_{X^n}||Q_{X^n}\right)
\end{aligned}$$

Q_{X^n} corresponds to a universal code that can achieve an overhead of at most $\text{Red}(\mathcal{P})$ bits with respect to every $P_{X^n} \in \mathcal{P}$. In most cases, $\text{Red}(\mathcal{P}) = o(n)$ and even $\text{Red}(\mathcal{P}) = O(\log(n))$.

Example 183 (Laplace Estimator for Universal Code for *iid* Bernoulli)**.** Take the context of Example 181. How can we find a "good" Q_{X^n} as measured by minimax redundancy when $\mathcal{P} = \{\text{Bern}(p)^{\otimes n} : p \in [0, 1]\}$? Let's try the following conditional distributions:

- Set $Q_{X_1} = (\{0,1\})$. We're setting a uniform prior on p (ie., $p \sim$ Beta$(1,1)$).

- For $t \geq 1$, let $n_1(X^t)$ and $n_0(X^t)$ be the number of 1s and 0s in X^t, respectively. Then, we set

$$Q_{X_{t+1}|X^t}(x_{t+1}=1) = \frac{n_1(X^t)+1}{t+2}$$
$$Q_{X_{t+1}|X^t}(x_{t+1}=0) = \frac{n_0(X^t)+1}{t+2}$$

to form the so-called *Add-1 estimator* or the *Laplace estimator*. We then have that

$$\begin{aligned} Q_{X^n}(x^n) &= \Pi_{t=0}^{n-1} Q_{X_{t+1}|X^t}(x_{t+1}) \\ &= \left(\frac{(1\cdot 2\cdot \ldots \cdot n_1(x^n))\cdot(1\cdot 2\cdot \ldots \cdot n_0(x^n))}{2\cdot 3\cdot \ldots \cdot (n+1)}\right) \\ &= \frac{n_1(x^n)!n_0(x^n)!}{(n+1)!} \end{aligned}$$

We also have that for any $P \in \mathcal{P}$

$$\begin{aligned} P_{X^n}(x^n) &= p^{n_1(x^n)}\cdot(1-p)^{n_0(x^n)} \\ &\leq \left(\frac{n_1(x^n)}{n}\right)^{n_1(x^n)}\cdot\left(\frac{n_0(x^n)}{n}\right)^{n_0(x^n)}, \text{ by using a MLE argument} \\ \implies \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} &= (n+1)\cdot\frac{\frac{n_1^{n_1}}{n_1!}\cdot\frac{n_0^{n_0}}{n_0!}}{\frac{n^n}{n!}}, \text{ defining } n_1 := n_1(x^n), n_0 := n_0(x^n) \\ &\asymp (n+1)\cdot\frac{\frac{\exp(n_1)}{\sqrt{n_1}}\cdot\frac{\exp(n_0)}{\sqrt{n_0}}}{\frac{\exp(n)}{\sqrt{n}}}, \text{ by Stirling} \\ &= (n+1)\frac{\sqrt{n}}{\sqrt{n_1 n_0}}, \text{ since } n_0+n_1=n \\ &\lesssim n, \text{ since the RHS is maximzed with } n_1 = n-1 \text{ (or } n_0 = n-1\text{).} \end{aligned}$$

As a remark, one can separately consider asymptotic sequences where $n_1 = 0$ WLOG and can come to the same asymptotic conclusion before applying Stirling. Anyhoo, that then means that

$$\begin{aligned} \text{Red}(Q_{X^n};\mathcal{P}) &= \sup_{P_{X^n}\in\mathcal{P}} D_{KL}(P_{X^n}||Q_{X^n}) \\ &= \sup_{P_{X^n}} \mathbb{E}_{P_{X^n}}\left[\log\left(\frac{P_{X^n}}{Q_{X^n}}\right)\right] \\ &\leq \log(n) + O(1) \end{aligned}$$

Example 184 (Krichecsky-Trofimov Estimator for Universal Code for *iid* Bernoulli)**.** Take the context of Example 181. Now we try the following conditional distributions:

- Set $Q_{X_1} = (\{0,1\})$. We're setting the following prior on p: $p \in$ Beta$(1/2,1/2)$.

- For $t \geq 1$, we define $n_1(X^t)$ and $n_0(X^t)$ as above. We set

$$Q_{X_{t+1}|X^t}(x_{t+1}=1) = \frac{n_1(X^t)+\frac{1}{2}}{t+1}$$
$$Q_{X_{t+1}|X^t}(x_{t+1}=0) = \frac{n_0(X^t)+\frac{1}{2}}{t+1}$$

to form the so-called *Add-$\frac{1}{2}$ estimator* or the *Krichevsky-Trofimov estimator*. We then have that

$$\begin{aligned} Q_{X^n}(x^n) &= \Pi_{t=0}^{n-1} Q_{X_{t+1}|X^t}(x_{t+1}) \\ &= \frac{1}{n!}\left(\frac{1}{2}\cdot\frac{3}{2}\cdot\ldots\cdot\left(n_1(x^n)-\frac{1}{2}\right)\right)\left(\frac{1}{2}\cdot\frac{3}{2}\cdot\ldots\cdot\left(n_0(x^n)-\frac{1}{2}\right)\right) \\ &= \frac{(2n_1(x^n)-1)!!\,(2n_0(x^n)-1)!!}{2^n n!} \end{aligned}$$

where $x!!$ is defined recursively by $x!! := x\cdot(x-2)!!$ with the base cases $-1!! := 0!! := 1$. We can do the same math to upper bound $P_{X^n}(x^n)$ as in Example 183 for any $P\in\mathcal{P}$. We then have that, defining $n_1 := n_1(x^n), n_0 := n_0(x^n)$,

$$\begin{aligned} \frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} &\le \frac{2^n n!\cdot\frac{n_1^{n_1}\cdot n_0^{n_0}}{n^n}}{(2n_1-1)!!(2n_0-1)!!} \\ &\asymp \frac{2^n n!\cdot\frac{n_1^{n_1}\cdot n_0^{n_0}}{n^n}}{\frac{(2n_1)!}{n_1!2^{n_1}}\cdot\frac{(2n_0)!}{n_0!2^{n_0}}} \\ &= \frac{n_1^{n_1}\cdot n_0^{n_0}\cdot 2^{2n}\cdot n!\cdot n_1!\cdot n_0!}{n^n\cdot(2n_1)!\cdot(2n_0)!} \\ &\asymp \frac{\cancel{n_1^{n_1}}\cdot\cancel{n_0^{n_0}}\cdot\cancel{2^{2n}}\cdot\cancel{n^n}\sqrt{n}\cdot\cancel{n_1^{n_1}}\cancel{\sqrt{n_1}}\cdot\cancel{n_0^{n_0}}\cancel{\sqrt{n_0}}\cdot\cancel{\exp(2n_1+2n_0)}}{\cancel{n^n}\cdot\cancel{2^{2n_1}\cdot 2^{2n_0}}\cdot\cancel{n_1^{2n_1}}\cdot\cancel{n_0^{2n_0}}\cdot\cancel{\exp(n+n_1+n_0)}\cdot\cancel{\sqrt{n_1}}\cdot\cancel{\sqrt{n_0}}},\ \text{by Stirling} \\ &\precsim \sqrt{n} \end{aligned}$$

As a remark, one can separately consider asymptotic sequences where $n_1 = 0$ WLOG and come to the same asymptotic conclusion before applying Stirling. Anyhoo, that then means that

$$\begin{aligned} \mathrm{Red}(Q_{X^n};\mathcal{P}) &= \sup_{P_{X^n}\in\mathcal{P}} D_{KL}(P_{X^n}||Q_{X^n}) \\ &= \sup_{P_{X^n}} \mathbb{E}_{P_{X^n}}\left[\log\left(\frac{P_{X^n}}{Q_{X^n}}\right)\right] \\ &\le \frac{1}{2}\log(n) + O(1) \end{aligned}$$

As it turns out, this constant of $\frac{1}{2}$ turns out to be tight so that in fact $\mathrm{Red}(\mathcal{P}) = \frac{1}{2}\log(n)+\Theta(1)$.

Definition 185 (Worst-Case Redundancy)**.** We define the worst-case redundancy (ie., pointwise redundancy) of a family of distributions $\mathcal{P}$ on X^n as

$$R^*(\mathcal{P}) := \inf_{Q_{X^n}}\sup_{P_{X^n}\in\mathcal{P}}\sup_{x^n}\log\left(\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)}\right)$$

We make some remarks on the worst-case redundancy:

- It's clear that $R^*(\mathcal{P}) \ge \mathrm{Red}(\mathcal{P})$. We use this fact in the derivations of Examples 183 and 184.
- $R^*(\mathcal{P})$ treats x^n as an individual sequence instead of drawing from a probability distribution.
- Unlike $\mathrm{Red}(\mathcal{P})$, which can be hard to characterize, $R^*(\mathcal{P})$ has a combinatorial characterization given in Theorem 186.

Theorem 186 (Combinatorial Characterization of Worst-Case Redundancy)**.** Consider a family of distributions $\mathcal{P}$

where all distributions in the family have finite support. We have that

$$R^*(\mathcal{P}) = \log\left(\sum_{x^n} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)\right)$$

The quantity $\sum_{x^n} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)$ is called the *Shtarkov sum*.

Proof.

[Upper bound]:

Let $\mathcal{Z} := \sum_{x^n} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)$ and $Q^*_{X^n}(x^n) := \frac{1}{\mathcal{Z}} \sup_{P_{X^n}\in\mathcal{P}} P_{X^n}(x^n)$, which is the normalized maximum likelihood distribution. Then, we have that

$$\begin{aligned} R^*(\mathcal{P}) &= \inf_{Q_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log\left(\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)}\right) \\ &\le \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log\left(\frac{P_{X^n}(x^n)}{Q^*_{X^n}(x^n)}\right) \\ &= \log(\mathcal{Z}) + \underbrace{\sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log\left(\frac{P_{X^n}(x^n)}{\sup_{\tilde{P}_{X^n}(x^n)\in\mathcal{P}} \tilde{P}_{X^n}(x^n)}\right)}_{=0} \\ &= \log(\mathcal{Z}) \end{aligned}$$

[Lower bound]:

Continue using the definitions in the proof of the upper bound. Consider any Q_{X^n}. We have that

$$\begin{aligned} \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log\left(\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)}\right) &= \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \left(\log\left(\frac{P_{X^n}(x^n)}{Q^*_{X^n}(x^n)}\right) + \log\left(\frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)}\right)\right) \\ &= \log(\mathcal{Z}) + \sup_{x^n} \log\left(\frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)}\right) \\ &\ge \log(\mathcal{Z}) + \sum_{x^n} Q^*_{X^n}(x^n) \log\left(\frac{Q^*_{X^n}(x^n)}{Q_{X^n}(x^n)}\right), \text{ max is greater than (weighted) average} \\ &= \log(\mathcal{Z}) + D_{KL}\left(Q^*_{X^n} || Q_{X^n}\right) \\ &\ge \log(\mathcal{Z}), \text{ since } D_{KL}\left(Q^*_{X^n} || Q_{X^n}\right) \ge 0 \text{ by Property 16} \end{aligned}$$

In light of the arbitrariness of Q^n, we have that $R^*(\mathcal{P}) \ge \log(\mathcal{Z})$. □

The combinatorial nature of $R^*(\mathcal{P})$ makes it easy to upper bound $\mathrm{Red}(\mathcal{P})$ for non-iid families $\mathcal{P}$.

Example 187 (Upper Bounding Overhead of Universal Code for Markov Chain)**.** Let

$$\mathcal{P} := \{P_{X^n} = p(x_1)\Pi_{t=1}^{n-1} M(x_{t+1} \mid x_t)\}$$

be the family of all time-homogeneous (first-order) Markov chains on the state space $[k]$. We claim that

$$\mathrm{Red}(\mathcal{P}) \precsim \frac{k(k-1)}{2}\log(n) + O_k(1)$$

Proof.

Following the approach of Example 184, we will apply the *Add-$\frac{1}{2}$ estimator* to all transition probabilities. That is, we will define

- $Q_{X_1} = ([k])$
- For $t \geq 1$ and $i, j \in [k]$, we define $n_{j\to i}(x^t) := \sum_{s=1}^{t} \mathbb{1}_{\{x_s=j, x_{s+1}=i\}}$ and $n_j(x^t) := \sum_{s=1}^{t} \mathbb{1}_{\{x_s=j\}}$. We then define the adaptive transition kernel

$$Q_{X_{t+1}|X^t}(x_{t+1}=i) = \frac{n_{j\to i}(x^t)+\frac{1}{2}}{n_j(x^t)+\frac{k}{2}}, \text{ if } x_t = j$$

Then, for any $x^n \in [k]^n$ and $P \in \mathcal{P}$, we have that

$$\begin{aligned}\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)} &= \frac{p(x_1)}{1/k}\Pi_{j=1}^{k} \underbrace{\Pi_{t\in[n-1],x_t=j}\frac{M(x_{t+1}\mid j)}{Q_{X_{t+1}|X^t}(x_{t+1})}}_{=O(\sqrt{n})^{k-1}} \\ &\leq k\cdot\left(C\sqrt{n}\right)^{k(k-1)}\end{aligned}$$

for some universal $C > 0$. To see that that product is $O(\sqrt{n})^{k-1}$, we use the analysis of Example 184. In that example, we looked at an iid binary problem. In this example, conditional on being in a state we have an iid multinomial problem but we can reduce the complexity to the binary problem by conditioning $k-1$ times on which subset the element draw is in.[a] From that analysis, we know that the ratio between the true probability and the learned adaptive probability for any single one of these binary probabilities has the rate $O(\sqrt{n})$ and there are $k-1$ of them. Meanwhile, there are k states for which we need to learn these probabilities and hence the additional multiplicative factor of k in the exponent.

As a result, we have that

$$\begin{aligned}\mathrm{Red}(\mathcal{P}) &\leq R^*(\mathcal{P}) \\ &= \inf_{\tilde{Q}_{X^n}} \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log\left(\frac{P_{X^n}(x^n)}{\tilde{Q}_{X^n}(x^n)}\right) \\ &\leq \sup_{P_{X^n}\in\mathcal{P}} \sup_{x^n} \log\left(\frac{P_{X^n}(x^n)}{Q_{X^n}(x^n)}\right) \\ &\leq \frac{k(k-1)}{2}\log(n) + \log(k) + k(k-1)\log(C) \\ &= \frac{k(k-1)}{2}\log(n) + O(k^2) \\ &= \frac{k(k-1)}{2}\log(n) + O_k(1)\end{aligned}$$

□

[a] For instance, if $k = 3$, we can write $p(x = x_1) = p(x \in \{x_1, x_2\})p(x = x_1 \mid x \in \{x_1, x_2\})$, where each of those terms is characterized by a single scalar parameter in the true distribution.

Theorem 188 (Entropic Upper Bound for Redundancy in iid Family)**.** Let $\mathcal{P}$ be a family of distributions such that it's ϵ-KL coverable for any $\epsilon > 0$. We have that

$$\mathrm{Red}(\mathcal{P}^{\otimes n}) \leq \inf_{\epsilon>0} n\epsilon^2 + \log\left(N_{KL}(\mathcal{P},\epsilon)\right)$$

Proof.

The proof is a subset of the proof of Theorem 157 and so is omitted.

Example 189 (Upper Bound on Redundancy for Parametric Family)**.** Suppose $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ is a parametric family with d parameters. Then, usually we have that $\log\left(N_{KL}(\mathcal{P}, \epsilon)\right) \asymp d \log(1/\epsilon)$. Choosing $\epsilon \asymp \sqrt{\frac{d}{n}}$ then gives the upper bound, applying Theorem 188,

$$\mathrm{Red}(\mathcal{P}^{\otimes n}) \leq \frac{d}{2} \log\left(\frac{n}{d}\right) + O(d)$$

Theorem 190 (Redundancy-Capacity Theorem)**.** Suppose that $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ is a parametric family with d parameters. We have that

$$\mathrm{Red}(\mathcal{P}) = \sup_{\pi \in \Delta(\Theta)} I(\theta; X), \text{ where } \theta \sim \pi, X \mid \theta \sim P_\theta$$

Proof.

Recall from Lemma 116, we have that

$$I(\theta; X) = \inf_{Q_X} \mathbb{E}_{\theta \sim \pi}\left[D_{KL}(P_\theta || Q_X)\right]$$

for some fixed prior $\pi \in \Delta(\Theta)$. As a result, we have that

$$\begin{aligned}
\sup_{\pi} I(\theta; X) &= \sup_{\pi} \inf_{Q_X} \mathbb{E}_{\theta \sim \pi}\left[D_{KL}(P_\theta || Q_X)\right] \\
&= \inf_{Q_X} \sup_{\pi} \mathbb{E}_{\theta \sim \pi}\left[D_{KL}(P_\theta) || Q_X\right], \text{ under regularity assumptions} \\
&= \inf_{Q_X} \sup_{\theta} D_{KL}(P_\theta || Q_X) \\
&= \mathrm{Red}(\mathcal{P}), \text{ by Definition 182}
\end{aligned}$$

□

An implication of Theorem 190 is that to lower bound $\mathrm{Red}(\mathcal{P})$ when $\mathcal{P}$ is a parametric family, we can search for a prior distribution π on the parametric family such that $I(\theta; X)$ is large when $\theta \sim \pi$.

Lemma 191 (Arithmetic Mean- Geometric Mean (AM-GM) Inequality)**.** For $x_1, ..., x_n \geq 0$, we have that

$$\Pi_{i=1}^n x_i \leq \left(\frac{\sum_{i=1}^n x_i}{d}\right)^d$$

Proof.

If any $x_i = 0$, the result follows directly so focus on the other case where $x_i > 0 \, \forall i \in [n]$. We have that

$$\begin{aligned}
\frac{1}{n}\log\left(\Pi_{i=1}^n x_i\right) &= \frac{1}{n}\sum_{i=1}^n \log\left(x_i\right) \\
&\leq \log\left(\frac{1}{n}\sum_{i=1}^n x_i\right), \text{ by Jensen's inequality for concave } x \mapsto \log(x) \\
\implies \left(\Pi_{i=1}^n x_i\right)^{1/n} &\leq \frac{1}{n}\sum_{i=1}^n x_i \\
\implies \Pi_{i=1}^n x_i &\leq \left(\frac{\sum_{i=1}^n x_i}{d}\right)^d
\end{aligned}$$

□

Lemma 192 (Symmetric Matrix Determinant Bound)**.** Suppose that $A \in \mathbb{R}^{d\times d}$ is a symmetric matrix.[a] Then, we have that

$$\det(A) \leq \left(\frac{\mathrm{Tr}(A)}{d}\right)^d$$

Proof.

We can then diagonalize $A = U\Lambda U'$ where $U, \Lambda \in \mathbb{R}^{d\times d}$, U has orthonormal columns, and Λ is a diagonal matrix that has the eigenvalues of A, denoted $\{\lambda_1, ...\lambda_d\}$ on the diagonal. We then have that

$$\begin{aligned}
\det(A) &= \det(U\Lambda U') \\
&= \overset{1}{\cancel{\det(U)}}\det(\Lambda)\overset{1}{\cancel{\det(U')}} \\
&= \Pi_{i=1}^d \lambda_i, \text{ with } \lambda_i \geq 0 \text{ since } A \text{ is PSD} \\
&\leq \left(\frac{\sum_{i=1}^d \lambda_i}{d}\right)^d, \text{ by Lemma 191} \\
&= \left(\frac{\mathrm{Tr}(\Lambda)}{d}\right)^d \\
&= \left(\frac{\mathrm{Tr}(\Lambda U'U)}{d}\right)^d, U \text{ has orthonormal columns} \\
&= \left(\frac{\mathrm{Tr}(U\Lambda U')}{d}\right)^d, \ \mathrm{Tr}(\cdot) \text{ is invariant to circular permutations} \\
&= \left(\frac{\mathrm{Tr}(A)}{d}\right)^d
\end{aligned}$$

□

[a] It then is also PSD.

Theorem 193 (Rissanen's Program)**.** Suppose that $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ is a parametric family where $\Theta \subseteq \mathbb{R}^d$ and has a non-empty interior. Suppose that for a positive sequence $(\epsilon_n)_{n\in\mathbb{N}}$, there exists an estimator $\hat{\theta}(X^n)$ such that

$\sup_{\theta\in\Theta} \mathbb{E}_{P_\theta^{\otimes n}}\left[||\theta - \hat{\theta}(X^n)||^2\right] \leq \epsilon_n^2$. Then, we have that

$$\mathrm{Red}(\mathcal{P}^{\otimes n}) \geq \log(\mathrm{Vol}_d(\Theta)) - \frac{d}{2}\log\left(\frac{2\pi e\epsilon_n^2}{d}\right)$$

Proof.

Let $\theta \sim \pi := (\Theta)$ and $h(\cdot)$ denote the differential entropy on $\mathbb{R}^d$. Then, we have that

$$\begin{aligned} I(\theta; X) &= h(\theta) - h(\theta \mid X^n) \\ &= \log\left(\mathrm{Vol}_d(\Theta)\right) - h(\theta \mid X^n) \end{aligned}$$

and

$$\begin{aligned} h(\theta \mid X^n) &= h(\theta - \hat{\theta}(X^n) \mid X^n) \\ &\leq h(\theta - \hat{\theta}(X^n)), \text{ conditioning reduces entropy} \\ &\leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\left(\det\left(\mathbb{E}\left[(\theta - \hat{\theta}(X^n))(\theta - \hat{\theta}(X^n))'\right]\right)\right), \text{ by Theorem 78 and Example 69} \\ &\leq \frac{d}{2}\log(2\pi e)\frac{d}{2}\log\left(\frac{\mathbb{E}\left[||\theta - \hat{\theta}(X^n)||^2\right]}{d}\right), \text{ by Lemma 192} \\ &\leq \frac{d}{2}\log\left(\frac{2\pi e\epsilon_n^2}{d}\right), \text{ by assumption} \end{aligned}$$

As a result, we have that

$$\begin{aligned} \mathrm{Red}(\mathcal{P}^{\otimes n}) &= \log\left(\mathrm{Vol}_d(\Theta)\right) - h(\theta \mid X^n) \\ &\geq \log\left(\mathrm{Vol}_d(\Theta)\right) - \frac{d}{2}\log\left(\frac{2\pi e\epsilon_n^2}{d}\right) \end{aligned}$$

□

Application 194 (Rissanen's Program Typical Application)**.** Take the context of Theorem 193. We usually have that $\mathrm{Vol}_d(\Theta) = \Omega\left(\frac{1}{d}\right)^{d/2}$ and $\epsilon_n^2 = O(d/n)$ so that Theorem 193 gives the lower bound (after a little algebra)

$$\mathrm{Red}(\mathcal{P}^{\otimes n}) \geq \frac{d}{2}\log\left(\frac{n}{d}\right) - O(d)$$

Lemma 195 (Parametric Mutual Information Lower Bound)**.** Suppose that $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ is a parametric family with d parameters. Let $\theta \sim \pi$ for $\pi \in \Delta(\Theta)$ and $X \mid \theta \sim P_\theta$ with support $\mathcal{X}$. For any $\lambda \in (0, 1]$, we have that[a]

$$I(\theta; X) \geq \underbrace{-\mathbb{E}_{P_{\theta X}}\left[\log\left(\mathbb{E}_{\theta'\sim\pi}\left[\left(\frac{P_{\theta'}(X)}{P_\theta(X)}\right)^\lambda\right]\right)\right]}_{=:f(\lambda)}$$

where $\theta' \overset{indep}{\sim} \pi$ is an independent copy of θ (ie., $\theta' \perp\!\!\!\perp (\theta, X)$).

Proof.

Take the definition of $f(\lambda)$ made using the underbrace. We observe that $f(1) = I(\theta; X)$. To see that

$$\begin{aligned}
f(1) &= -\mathbb{E}_{P_{\theta X}}\left[\log\left(\mathbb{E}_{\theta' \sim \pi}\left[\left(\frac{P_{\theta'}(X)}{P_\theta(X)}\right)^1\right]\right)\right] \\
&= -\mathbb{E}_{P_{\theta X}}\left[\log\left(\frac{P_X(X)}{P_\theta(X)}\right)\right], \text{ where } P_X \text{ is the marginal of } X \\
&= \mathbb{E}_{P_{\theta X}}\left[\log\left(\frac{P_\theta(X)}{P_X(X)}\right)\right] \\
&= \mathbb{E}_{P_{\theta X}}\left[\log\left(\frac{P_{\theta X}(\theta, X)}{P_X(X)\pi(\theta)}\right)\right] \\
&= I(\theta; X), \text{ by definition}
\end{aligned}$$

Since cumulant generating functions (CGFs) are convex,[b] f is concave. Finally, with some calculus and algebra, we have that

$$f'(\lambda) = \mathbb{E}_{P_{\theta X}}\left[\log(P_\theta(X))\right] - \mathbb{E}_{P_{\theta X}}\left[\frac{\mathbb{E}_{\theta' \sim \pi}\left[P_{\theta'}(X)^\lambda \log(P_{\theta'}(X))\right]}{\mathbb{E}_{\theta' \sim \pi}\left[P_{\theta'}(X)^\lambda\right]}\right]$$

When $\lambda = 1$, we have that

$$\begin{aligned}
\mathbb{E}_{P_{\theta X}}\left[\frac{\mathbb{E}_{\theta' \sim \pi}\left[P_{\theta'}(X)^1 \log(P_{\theta'}(X))\right]}{\mathbb{E}_{\theta' \sim \pi}\left[P_{\theta'}(X)^1\right]}\right] &= \int_{\mathcal{X}} P_X(x) \frac{\int_\Theta \pi(\theta') P_{\theta'}(x) \log(P_{\theta'}(x)) d\theta'}{\int_\Theta \pi(\theta') P_{\theta'}(x) d\theta'} dx, \\
&\quad \text{marginalizing immediately over } \theta \\
&= \int_{\mathcal{X}} \cancel{P_X(x)} \frac{\int_\Theta \pi(\theta') P_{\theta'}(x) \log(P_{\theta'}(x)) d\theta'}{\cancel{P_X(x)}} dx \\
&= \mathbb{E}_{P_{\theta X}}\left[\log(P_\theta(X))\right]
\end{aligned}$$

Therefore, we have that $f'(1) = \mathbb{E}_{P_{\theta X}}\left[\log(P_\theta(X))\right] - \mathbb{E}_{P_{\theta X}}\left[\log(P_\theta(X))\right] = 0$, which implies that $f'(\lambda) \geq 0$ by the concavity of f for $\lambda \in [0, 1]$. In turn, that implies that

$$\begin{aligned}
f(\lambda) &\leq f(1) \\
&= I(\theta; X)
\end{aligned}$$

□

[a] Make all necessary regularity assumptions so that all distributions admit densities with respect to the Lesbegue measure.

[b] A cumulant generating for a random variable X is given by $K_X(\lambda) := \log\left(\mathbb{E}[\exp(\lambda X)]\right)$. The convexity of this function is a consequence of Hölder's Inequality.

Theorem 196 (Haussler and Opper Program). Suppose that $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ is a parametric family with d parameters. Also, suppose that $\mathcal{P}$ is such that it admits a finite maximal ϵ-packing with Hellinger distance for any $\epsilon > 0$. We have that

$$\text{Red}(\mathcal{P}) \geq \sup_{\epsilon > 0} \min\left(\frac{n\epsilon^2}{2}, \log(M_H(\mathcal{P}, \epsilon))\right) - \log(2)$$

Proof.

Take any $\epsilon > 0$ and let $P_{\theta_1}, .., , P_{\theta_M}$ be an ϵ-packing of $\mathcal{P}$ under Hellinger distance. Define $\pi := \frac{1}{M} \sum_{i=1}^M \delta_{\theta_i}$.

Then, for $\theta \sim \pi$, we have that

$$\begin{aligned}
I(\theta; X^n) &\geq -\mathbb{E}_{P_{\theta X^{\otimes n}}}\left[\log\left(\mathbb{E}_{\theta' \sim \pi}\left[\left(\frac{P_{\theta'}^{\otimes n}(X^n)}{P_{\theta}^{\otimes n}(X^n)}\right)^{1/2}\right]\right)\right], \text{ by Lemma 195} \\
&= -\frac{1}{M}\sum_{i=1}^{M}\mathbb{E}_{P_{\theta_i}^{\otimes n}}\left[\log\left(\frac{1}{M}\sum_{j=1}^{M}\left(\frac{P_{\theta_j}^{\otimes n}(X^n)}{P_{\theta_i}^{\otimes n}(X^n)}\right)^{1/2}\right)\right] \\
&\geq -\frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{1}{M}\sum_{j=1}^{M}\mathbb{E}_{P_{\theta_i}^{\otimes n}}\left[\left(\frac{P_{\theta_j}^{\otimes n}(X^n)}{P_{\theta_i}^{\otimes n}(X^n)}\right)^{1/2}\right]\right), \text{ since } x \mapsto -\log(x) \text{ is convex and Jensen's} \\
&= -\frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{1}{M}\sum_{j=1}^{M}\left(1 - \frac{1}{2}H^2\left(P_{\theta_i}^{\otimes n}, P_{\theta_j}^{\otimes n}\right)\right)\right) \\
&= -\frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{1}{M}\sum_{j=1}^{M}\left(1 - \frac{1}{2}H^2\left(P_{\theta_i}, P_{\theta_j}\right)\right)^{n}\right), \text{ by Section 3.3} \\
&\geq -\frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{1}{M} + \frac{M-1}{M}\left(1 - \frac{\epsilon^2}{2}\right)^{n}\right), \text{ since } H^2(P_{\theta_i}, P_{\theta_j}) \geq \epsilon^2 \text{ for } i \neq j \in [M] \text{ by packing} \\
&\geq -\frac{1}{M}\sum_{i=1}^{M}\log\left(\frac{1}{M} + \frac{M-1}{M}\exp\left(-\frac{n\epsilon^2}{2}\right)\right), \\
&\qquad \text{since } x \mapsto \left(1 + \frac{c}{x}\right)^{x} \text{ is increasing for } x > 0 \text{ and } \lim_{x \to \infty}\left(1 + \frac{c}{x}\right)^{x} = \exp(c) \\
&\geq -\log\left(\frac{2}{\min\left(\frac{M}{M-1}\exp\left(\frac{n\epsilon^2}{2}\right), M\right)}\right), \text{ since } \frac{1}{a} + \frac{1}{b} \leq \frac{2}{\min(a,b)} \\
&= \min\left(\frac{M}{M-1}\frac{n\epsilon^2}{2}, \log(M)\right) - \log(2), \text{ exchanging log and min since arguments are positive} \\
&\geq \min\left(\frac{n\epsilon^2}{2}, \log(M)\right) - \log(2)
\end{aligned}$$

Note that $\epsilon > 0$ was arbitrary as was the ϵ-packing. Thus, by Theorem 190, which says that

$$\mathrm{Red}(\mathcal{P}) = \sup_{\tilde{\pi} \in \Delta(\Theta)} I(\theta; X)$$

we have the result. □

Application 197 (Haussler and Opper Program Typical Application)**.** Take the context of Theorem 196. In d-dimensional parametric families, we usually have that $\log(M_H(\mathcal{P}, \epsilon)) \asymp d\log(1/\epsilon)$. Applying Theorem 196, we have that

$$\begin{aligned}
\mathrm{Red}(\mathcal{P}) &\geq \sup_{\epsilon>0} \min\left(\frac{n\epsilon^2}{2}, \log(M_H(\mathcal{P},\epsilon))\right) \\
&\asymp \sup_{\epsilon>0} \min\left(\frac{n\epsilon^2}{2}, d\log(1/\epsilon)\right) \\
&\gtrsim \min\left(\frac{n}{2}\cdot\frac{d\log(n)}{n}, \frac{d}{2}\log\left(\frac{n}{d\log(n)}\right)\right), \\
&\qquad \text{picking } \epsilon^2 \asymp \frac{d\log(n)}{n} \text{ to make both terms asymptotically equivalent and the min as large as possible} \\
&\gtrsim \frac{d}{2}\log\left(\frac{n}{d\log(n)}\right)
\end{aligned}$$

11.2 Relationship between Prediction Risk and Redundancy

In this section, we will define prediction risk and investigate its relationship with redundancy.

Definition 198 (Prediction Risk)**.** Let $\mathcal{P}$ be a family of distributions. We define the next symbol *prediction risk* as

$$\mathrm{Risk}_n(\mathcal{P}) := \inf_{Q_{X_{n+1}|X^n}} \sup_{P_{X^{n+1}}} \mathbb{E}_{P_{X^n}}\left[D_{KL}(P_{X_{n+1}|X^n}||Q_{X_{n+1}|X^n})\right]$$

Proposition 199 (Mutual Information Representation of Prediction Risk)**.** Suppose that $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ is a parametric family with d parameters. We have that

$$\mathrm{Risk}_n(\mathcal{P}) = \sup_{\pi\in\Delta(\Theta)} I(\theta; X_{n+1} \mid X^n)$$

Proof.

We have that

$$\begin{aligned}
\mathrm{Risk}_n(\mathcal{P}) &= \inf_{Q_{X_{n+1}|X^n}} \sup_{\theta\in\Theta} \mathbb{E}_{X^n\sim P_\theta}\left[D_{KL}(P_{X_{n+1}|X^n,\theta}||Q_{X_{n+1}|X^n})\right] \\
&= \inf_{Q_{X_{n+1}|X^n}} \sup_{\pi\in\Delta(\Theta)} \mathbb{E}_{\theta\sim\pi,X^n\sim P_\theta}\left[D_{KL}(P_{X_{n+1}|X^n,\theta}||Q_{X_{n+1}|X^n})\right] \\
&= \sup_{\pi\in\Delta(\Theta)} \inf_{Q_{X_{n+1}|X^n}} \mathbb{E}_{\theta\sim\pi,X^n\sim P_\theta}\left[D_{KL}(P_{X_{n+1}|X^n,\theta}||Q_{X_{n+1}|X^n})\right], \text{ under regularity conditions} \\
&= I(\theta; X_{n+1} \mid X^n), \text{ by Lemma 116}
\end{aligned}$$

□

Proposition 200 (Redundancy-Risk Inequality)**.** Suppose that $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ is a parametric family with d parameters. We have that

$$\mathrm{Red}_n(\mathcal{P}) \leq \sum_{t=0}^{n-1} \mathrm{Risk}_t(\mathcal{P})$$

We use the subscript n to be precise but the quantity is equivalent to that of Definition 182.

Proof.

By a simple consequence of Property 19, when we write $I(\theta; X^n) = \mathbb{E}_{\theta\sim\pi}\left[D_{KL}(P_{X^n|\theta}||P_{X^n})\right]$, we have that for any $\pi \in \Delta(\Theta)$

$$\begin{aligned} I(\theta; X^n) &= \sum_{t=1}^{n} I(\theta; X_t \mid X^{t-1}) \\ \implies \mathrm{Red}_n(\mathcal{P}) &= \sup_{\pi\in\Delta(\Theta)} I(\theta; X^n), \text{ by Theorem 190} \\ &= \sup_{\pi\in\Delta(\Theta)} \sum_{t=1}^{n} I(\theta; X_t \mid X^{t-1}) \\ &\le \sum_{t=1}^{n} \sup_{\pi\in\Delta(\Theta)} I(\theta; X_t \mid X^{t-1}) \\ &= \sum_{t=1}^{n} \mathrm{Risk}_t(\mathcal{P}), \text{ by Proposition 199} \end{aligned}$$

□

Theorem 201 (Online-to-Batch Conversion for Stationary Distributions)**.** Consider a family of distributions $\mathcal{P}$. Suppose that each $P_{X^{n+1}}$ is stationary in the sense that for any $K \subseteq [n+1]$, we have that $P_{X_{K_1},\ldots,X_{K_{|K|}}} = P_{X_{K_1+\Delta},\ldots,X_{K_{|K|}+\Delta}}$ for any suitable $\Delta \in \mathbb{Z}$. Then, we have that

$$\mathrm{Risk}_n(\mathcal{P}) \le \frac{1}{n}\mathrm{Red}(\mathcal{P}) + \mathrm{Mem}(\mathcal{P})$$

where

$$\mathrm{Mem}(\mathcal{P}) := \sup_{P_{X^{n+1}}} \frac{1}{n}\sum_{t=1}^{n} I(X_{n+1}; X^{n-t} \mid X_{n-t+1}^{n})$$

Proof.

Suppose that $Q_{X^{n+1}} = \Pi_{t=1}^{n+1} Q_{X_t|X^{t-1}}$ attains the minimax redundancy $\mathrm{Red}(\mathcal{P})$. Define

$$\tilde{Q}_{X_{n+1}|X^n} := \frac{1}{n}\sum_{t=1}^{n} Q_{X_{t+1}|X^t}(\cdot \mid X_{n-t+1}^{n})$$

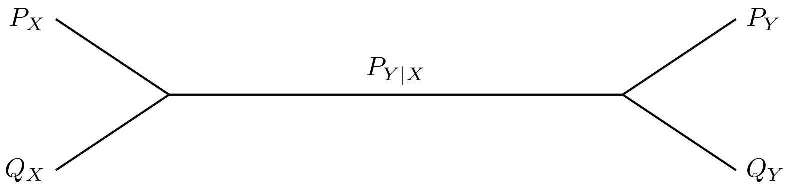
The notation $Q_{X_{t+1}|X^t}(\cdot \mid X_{n-t+1}^{n})$ might be a little confusing so we wish to clarify. The X^t in the subscript refers to the fact that the function conditions on t elements; the X_{n-t+1}^{n} are the t inputted elements. We're averaging a predictor of X_{n+1} that conditions on one element, two elements, ..., n elements. We then have that for any $P \in \mathcal{P}$

$$\begin{aligned}
\mathbb{E}_{P_{X^n}}\left[D_{KL}(P_{X_{n+1}|X^n}||\tilde{Q}_{X_{n+1}|X^n})\right] &\leq \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{P_{X^n}}\left[D_{KL}(P_{X_{n+1}|X^n}||Q_{X_{t+1}|X^t})\right], \text{ by Property 18} \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{P_{X^{n+1}}}\left[\log\left(\frac{P_{X_{n+1}|X^n}(X_{n+1} \mid X^n)}{Q_{X_{t+1}|X^t}(X_{n+1} \mid X^n_{n-t+1})}\right)\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{P_{X^{n+1}}}\left[\log\left(\frac{P_{X_{n+1}|X^n_{n-t+1}}(X_{n+1} \mid X^n_{n-t+1})}{Q_{X_{t+1}|X^t}(X_{n+1} \mid X^n_{n-t+1})}\right)\right] \\
&\quad + \mathbb{E}_{P_{X^{n+1}}}\left[\log\left(\frac{P_{X_{n+1}|X^n}(X_{n+1} \mid X^n)}{P_{X_{n+1}|X^n_{n-t+1}}(X_{n+1} \mid X^n_{n-t+1})}\right)\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{P_{X^{n+1}}}\left[\log\left(\frac{P_{X_{t+1}|X^t}(X_{t+1} \mid X^t)}{Q_{X_{t+1}|X^t}(X_{t+1} \mid X^n_{t-t+1})}\right)\right] \\
&\quad + I(X_{n+1}; X^{n-t} \mid X^n_{n-t+1}), \text{ since } (X_{n+1}, X^n_{n-t+1}) =_d (X_{t+1}, X^t) \\
&\leq \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{P_{X^t}}\left[D_{KL}(P_{X_{t+1}|X^t}||Q_{X_{t+1}|X^t})\right] + \mathrm{Mem}(\mathcal{P}), \text{ by marginalizing} \\
&\quad \text{over irrelevant } X_t \text{ and taking supremum over } P \in \mathcal{P}, \text{respectively} \\
&= \frac{1}{n}D_{KL}(P_{X^{n+1}}||Q_{X^{n+1}}) + \mathrm{Mem}(\mathcal{P}), \text{ by Property 19} \\
&\leq \frac{1}{n}\,\mathrm{Red}(\mathcal{P}) + \mathrm{Mem}(\mathcal{P}), \text{ taking the supremum over } P \in \mathcal{P}
\end{aligned}$$

By transitivity and the definition of $\mathrm{Risk}_n(\mathcal{P})$ in Definition 198 (we could appropriately time taking the supremums over $P \in \mathcal{P}$ on both sides), we have the result. □

12 Strong Data Processing Inequalities (SDPIs)

Recall the data processing inequality of Property 20. Suppose the distributions P_X, Q_X are inputted and the distributions P_Y, Q_Y are defined as follows using the channel depicted below: $P_Y(y) = \sum_{x\in\mathcal{X}} P_X(x)P_{Y|X}(Y|X = x)$; $Q_Y(y) = \sum_{x\in\mathcal{X}} Q_X(x)P_{Y|X}(Y|X = x)$.



Property 20 says that

$$D_{KL}(P_Y||Q_Y) \leq D_{KL}(P_X||Q_X)$$

Meanwhile, the strong data processing inequality (SDPI) will say that

$$D_{KL}(P_Y||Q_Y) \leq \eta(P_{Y|X}) \cdot D_{KL}(P_X||Q_X)$$

for some $\eta(P_{Y|X}) < 1$.

12.1 Input-Independent SDPI

References

Sheldon Axler. *Linear Algebra Done Right*. Springer, 4th edition, 2025. Electronic edition, available at `https://linear.axler.net/LADR4e.pdf`.

Yanjun Han. Information theory for statistics and learning. `https://yanjunhan2021.github.io/courses/info_theory/notes.html`, 2025. Accessed: 2026-01-14.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.