**High-dimensional Statistics & Learning Theory Study Group** **March 31, 2016**

*f-divergences*

**Leader: I-Hsiang Wang** **Notes: Chi-Ning Chou**

*This week we are going to study f-divergences, which is a class of measurements for the dissimilarity of two distributions.*

# Contents

# 1 Overview

The goal of f-divergences is to *quantitatively* measure the differences of two probability distributions. Most of the time, these divergences will have some direct relation to some useful applications. For instance, total variation quantifies the error in hypothesis testing, KL-divergence appears in Sanov's theorem and relates to information theory, $\chi^2-$divergence provides lower bound for statistical estimation, etc. As a result, sometimes we can use f-divergences to find the fundamental limits of certain statistical analyses.

Moreover, there are some relation among f-divergences. For example, Pinsker's inequality told us that we can use KL-divergence to upper bound the total variation. With this useful fact, if the total variation of two distributions is hard to compute, we can use KL-divergence to bound the error in hypothesis testing instead.

| **f-divergence** | **Application** |
|---|---|
| Total variation | Hypothesis testing |
| KL-divergence | Information theory, Sanov's theorem, MLE |
| $\chi^2-$divergence | Estimation lower bound (HCR, CR) |
| Hellinger distance | Bounds for total variation, Disjointness lower bound |

Table 1: Common f-divergences and their applications.

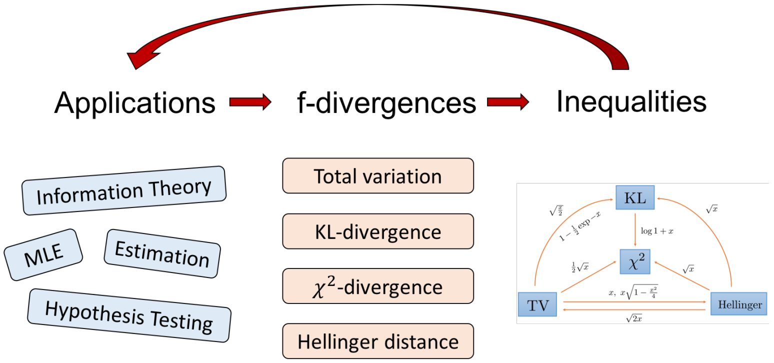

Figure 1: Overview of f-divergences: We defined f-divergences from observations in some applications. Then, we find inequalities among each f-divergence. Finally, we can use these relations to construct new bound for applications.

In the following discussion, we are going to learn the properties, inequalities, and technical details of f-divergences so that we can use this powerful tools in the future. In Section 2, we are going to see the basic definition of f-divergence and why they are defined in that way. Then, four common f-divergences and several important properties will be presented. In Section 3, we will go into the details of the complicated relation among f-divergences and understand the pros and cons of them. Finally, in Section 4, applications such as HCR lower bound will be discussed.

The goal of this notes is to explore the useful facts and properties of f-divergences. After studying this materials, readers should have the understanding of the framework of f-divergences, the pros and cons of each f-divergence, and how to use the inequalities and bounds for research.
Most of the contents are followed from the lecture 3-6 of ECE598. Some relevant materials are from [Pol01] and [Tsy09]. For a quick overview, one can see the slides I made [Cho15].

## 2 Definition, common f-divergences, and properties

Before we formally introduce the f-divergences, let's recall the differences between *distance* and *divergence*.
Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $x, y, z \in \mathcal{X}$, non-negativity means $f(x, y) \geq 0$, identity means $f(x, y) = 0 \Leftrightarrow x = y$, symmetry means $d(x, y) = d(y, x)$, triangle inequality means $d(x, y) + d(y, z) \geq d(x, z)$. Note that the differences between distance and divergence are only two properties: symmetry and triangle inequality. However, it does not mean that distances are always better than divergence since sometimes divergence might enjoy better properties.

| | Non-negativity | Identity | Symmetry | Triangle inequality |
|---|---|---|---|---|
| Distance | O | O | O | O |
| Divergence | O | O | X | X |

Table 2: Distance vs. Divergence.

## 2.1 Definition

Given two probability distributions $P$ and $Q$ over alphabet set $\mathcal{X}$, f-divergence uses the probability ratio (relative density) of $P$ and $Q$ to quantify the dissimilarity of them.

**Definition 1 (f-divergence)** *Let two probability distributions $P$ and $Q$ over alphabet set $\mathcal{X}$ such that $P \ll Q$. Consider $f : (0, \infty) \rightarrow \mathbb{R}$ be strictly convex at 1 and $f(1) = 0$. We define the f-divergence of $P$ and $Q$ as $D_f(P||Q) := \mathbb{E}_Q[\frac{dP(X)}{dQ(X)}]$.*

Intuitively, the more point $x \in \mathcal{X}$ such that $\frac{dP(x)}{dQ(x)} = 1$, the two distributions are more similar. To capture the idea of "more points", we simply take expectation over $Q$.

**Remark**:

- The reason why we restrict $f$ to be convex is deeply related to *convex conjugate*, which will be discussed in Section 2.2. For now, one can just think of this requirement as providing nice structure such as joint convexity and data processing inequality.
- Here we require $P \ll Q$, $P$ is absolutely continuous w.r.t. $Q$. Namely, for any set $E$ such that $Q(E) = 0$, then $P(E) = 0$.
- To be more general, we can define the $f-$divergence of $P$ and $Q$ with their *dominating measure* $\nu$ as follow: $D_f(P||Q) := \mathbb{E}_Q[f(\frac{dP/d\nu}{dQ/d\nu})]$. Recall that we say measure $\nu$ dominate $P$ and $Q$ if $P \ll \nu$ and $Q \ll \nu$.

**Exercise 2.1** Consider $L(P||Q) = \int \frac{(dP-dQ)^2}{dP+dQ}$, is it an f-divergence?

Before we see that definition of the four common f-divergences, I want to emphasize four important aspects when you learning these divergences. They are the pros and cons of each divergence.

- Is it a distance or not?
- What's the range of it?
- What's the underlying intuition?
- Any good properties?

Please put these in mind when you learn the following divergences. Once you understand the pros and cons of divergences, you can master them. Now, let's see the four common f-divergences:

| f-divergence | Notation | $f(x)$ | Integration form | Distance? | Range |
|---|---|---|---|---|---|
| Total variation | $d_{TV}(P,Q)$ | $\frac{1}{2}\lvert x-1\rvert$ | $\frac{1}{2}\int \lvert dP-dQ\rvert$ | O | [0,1] |
| KL-divergence | $D(P\|\|Q)$ | $x\log x$ | $\int \log\frac{dP}{dQ}dP$ | No symmetry | $[0,\infty)$ |
| $\chi^2-$divergence | $\chi^2(P\|\|Q)$ | $(x-1)^2$ | $\int \frac{dP^2}{dQ} - 1$ | No $\triangle -$ ineq. | $[0,\infty)$ |
| Hellinger distance | $H^2(P,Q)$ | $(1-\sqrt{x})^2$ | $\int(\sqrt{dP}-\sqrt{dQ})^2$ | O | [0,2] |

Table 3: Four common f-divergences.

Here, I plot their $f$ functions in Figure 2 for comparison.

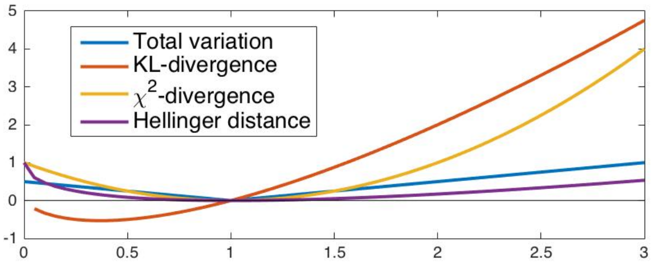

Figure 2: $f$ function.

For me, directly observe the $f$ function is too difficult. Let's take a look at the f-divergences of normal distribution with the same variance.

**Exercise 2.2** Let $P \sim \mathcal{N}(\theta,\sigma^2)$, $Q \sim \mathcal{N}(0,\sigma^2)$. Compute the following f-divergences.

- $d_{TV}(P,Q)$:

- $D_{KL}(P||Q)$:

- $\chi^2(P||Q)$:

- $H^2(P,Q)$:

## 2.2 Variational representation

We cannot get too much information directly from the definitions of these f-divergences. However, most of them have various representations so that we can find some interesting intuitions behind them.

### Variational representation of total variation

Let's start with the total variation. I will list the four variational form of it and left the proofs as exercises.

**Exercise 2.3**

- $d_{TV}(P,Q) = \sup_{E\subseteq\mathcal{X}} P(E) - Q(E)$:
- $d_{TV}(P,Q) = 1 - \int dP \wedge dQ$, where $a \wedge b = \min(a,b)$:
- $d_{TV}(P,Q) = \inf_{P_{X,Y}:P_X=P,P_Y=Q} \mathbb{P}[X \neq Y]$, provided that $\{(x,x)|x \in \mathcal{X}\}$ is measurable:

- $d_{TV}(P,Q)=\frac{1}{2}\sup_{f\in\mathcal{F}}\mathbb{E}_P[f(X)]-\mathbb{E}_Q[f(X)]$:

Total variation is a special f-divergence so that it has so many variational representations. For other f-divergences, most of the time it's not so lucky. However, there's actually a systematic approach via *convex conjugate* to characterize variational representation for f-divergence.

Let me present this idea in an intuitive way.

**Convex conjugate**

Since $f:(0,\infty)$ is a convex function, we know that $f$ is actually the **supremum** of a family of **affine functions**. Namely, for any slope $m$ and feasible intercept $\alpha$, we have $\forall t\in\mathbb{R}, f(t)\geq mt-\alpha$. Observe that

$$\begin{aligned} f(t)\geq mt-\alpha &\Leftrightarrow \alpha\geq mt-f(t),\forall t\in\mathbb{R}\\ &\Leftrightarrow \alpha\geq\sup_{t\in\mathbb{R}} mx-f(t) \end{aligned} \tag{1}$$

It's tempting to find the optimal intercept for a given slope so that we can find the affine function that *touches* function $f$. In fact, this turns out to be the concept of convex conjugate! We define the convex conjugate of $f$ as

$$f^*(m)=\sup_{t\in\mathbb{R}} mt-f(t)$$

Thus, from (1),we have a direct lower bound for $f$ (also known as the Young-Fenchel inequality)

$$f(t)\geq mt-f^*(m),\forall m\in\mathbb{R}$$

**Remark**: If $f$ is *convex* and *lower semi-continuous*, then $f(t)=f^{**}(t)$. That is, $f(t)=\sup_{m\in\mathbb{R}} mt-f^*(m)$.

Finally, let's go back to our variational representation of f-divergences. We can replace the function $f$ term with the supremum of an affine function as follow.

**Theorem 2 (variational representation)**

$$D_f(P||Q)=\sup_{g:\mathcal{X}\to\mathbb{R}}\mathbb{E}_P[g(X)]-\mathbb{E}_Q[f^*(g(X))]$$

**Proof:** Since the function $f$ of each $f-$divergence is both convex and lower semi-continuous, we have

$$\begin{aligned} D_f(P||Q)=\int_{x\in\mathcal{X}}f(\frac{dP(x)}{dQ(x)})dQ(x)&=\int_{x\in\mathcal{X}}[\sup_{m_x} m_x\frac{dP(x)}{dQ(x)}-f^*(m_x)]dQ(x)\\ (*)&=\int_{x\in\mathcal{X}}g(x)dP(x)-\int_{x\in\mathcal{X}}f^*(g(x))dQ(x)\\ &=\mathbb{E}_P[g(X)]-\mathcal{E}_Q[f^*(g(X))] \end{aligned}$$

In (*), $g(x) = \arg\max_{m_x} m_x x - f^*(m_x)$ for any $x \in \mathcal{X}$. ■

One can play with the following examples.

**Exercise 2.4** Compute the convex conjugate of the following f-divergences.

- **KL-divergence**: $f(t) = t\log t$, $f^*(m) = e^{m-1}$
- $\chi^2-$**divergence**: $f(t) = (x-1)^2$, $f^*(m) = m + \frac{m^2}{4}$

## 2.3 Properties

**Property 3 (basic properties of f-divergences)**

- *Identity:* $D_f(P||Q) = 0$ *iff* $\mathbf{P}[P(X) = Q(X)] = 1$.
- *Non-increasing:* $D(P||Q) \geq 0$.
- *Joint convexity:* $(P,Q) \to D_f(P||Q)$ *is jointly convex.*

Most of the time, $P$ and $Q$ are not that simple for us to directly compute their f-divergences. Luckily, f-divergences have several nice properties, which can simplify many common cases and make the computation easier.

**Property 4 (Processing properties of f-divergences)**

- **Conditioning increase f-divergence**: *Let* $P_{Y|X}$ *and* $Q_{Y|X}$ *be two channel rules/processes and* $P_X$ *be a prior distribution. The conditional f-divergence between* $P_{Y|X}$ *and* $Q_{Y|X}$ *given* $P_X$ *is defined as*
  $$D_f(P_{Y|X}||Q_{Y|X}|P_X) := \mathbb{E}_{P_X}[D_f(P_{Y|X||}Q_{Y|X})]$$
  *Moreover, let* $P_Y = P_X \cdot P_{Y|X}$ *and* $Q_Y = P_X \cdot Q_{Y|X}$*, we have*
  $$D_f(P_Y||Q_Y) \leq D_f(P_{Y|X}||Q_{Y|X}|P_X)$$
  **Remark**: *The average f-divergence between channels are at least that of output distributions. See Figure 3.*

- **Data processing inequality***: Consider a channel (process) that produces $Y$ given $X$ via the law $P_{Y|X}$ and $P_X, Q_X$ are two source distributions and $P_Y, Q_y$ are the corresponding destination distributions. We have,*

  $$D_f(P_Y||Q_Y) \leq D_f(P_X||Q_X)$$

  **Remark***: After processing, it becomes more difficult to distinguish two distributions. See Figure 4.*

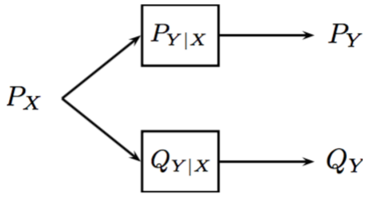

Figure 3: Conditioning increases f-divergence.

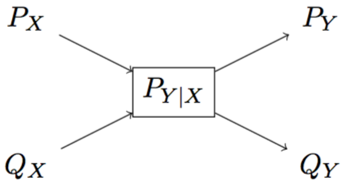

Figure 4: Data processing inequality.

- **Invariance***: For any one-to-one function $g : \mathcal{X} \to \mathcal{Y}$, we have*

  $$D_f(P_{g(X)}||Q_{g(X)}) = D_f(P||Q)$$

  **Remark***: f-divergences are invariant under translation, dilation, or rotation.*

- **Sufficiency***: Let $Y$ be a sufficient statistic for testing $P_X$ and $Q_X$, we have*

  $$D_f(P_Y||Q_Y) = D_f(P_X||Q_X)$$

  **Remark:** *Sufficient statistic is enough.*

- **Dimension reduction***: For any distributions $P_1$, $P_2$, and $Q$, we have*

  $$D_f(P_0 \otimes Q||P_1 \otimes Q) = D_f(P_0||P_1)$$

We can think of *processing* as playing with the input/source data. For example, an estimator is like a conditional distribution $P_{\hat{\theta}|\theta}$ given underlying distribution $P_\theta$ and output an estimator with distribution $P_{\hat{\theta}}$.

**Property 5 (Product properties of f-divergences)**

- **Tensorization***: For any distributions $P_1, \dots, P_k$ and $Q_1, \dots, Q_k$, we have*

  $$d_{TV}(\prod_{i=1}^k P_i, \prod_{i=1}^k Q_i) \leq \sum_{i=1}^k d_{TV}(P_i, Q_i)$$

  $$D(\prod_{i=1}^k P_i|| \prod_{i=1}^k Q_i) = \sum_{i=1}^k D(P_i||Q_i)$$

  $$\chi^2(\prod_{i=1}^k P_i|| \prod_{i=1}^k Q_i) = \prod_{i=1}^k (1 + \chi^2(P_i||Q_i)) - 1$$

  $$H^2(\prod_{i=1}^k P_i, \prod_{i=1}^k Q_i) = 2 - 2\prod_{i=1}^k (1 - \frac{H^2(P_i, Q_i)}{2})$$

**Remark**: *Here* $\prod_{i=1}^{k} P_i$ *is the tensor product of distributions* $P_1, \dots, P_k$*, which is simply the joint distribution with independent marginal distributions* $P_i$*.*

Product of probability distribution is directly related to multiple samples in statistics. By Property 5, we can simplify the computation for the f-divergence of underlying joint distribution. Please do the following exercises with the help of Property 4 and Property 5.

**Exercise 2.5** Assume $\Sigma$ is positive semidefinite matrix. Let $P \sim \mathcal{N}(\theta, \Sigma^2)$, $Q \sim \mathcal{N}(0, \Sigma^2)$. Compute the following f-divergences.

- $\chi^2(P||Q)$:

- $H^2(P, Q)$:

# 3 Inequalities between f-divergences

It's time for playing with these four f-divergences! Let me first give you an overview for the relation among them in Figure 5.

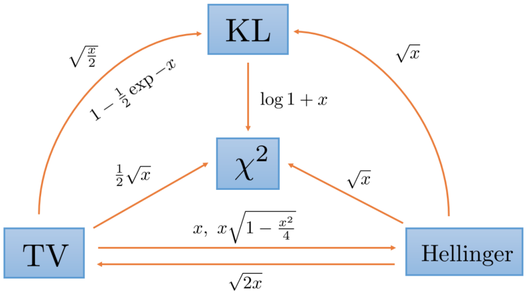

Figure 5: Inequalities between f-divergences: $D_1 \xrightarrow{f(x)} D_2$ means $D_1 \leq f(D_2)$.

In the following, I will take two inequalities as concrete examples and then present a general method called: *joint range* to show how to derive sharp inequalities.

## 3.1 Pinsker's inequality: $D_{KL}(P||Q) \geq 2d_{TV}^2(P, Q)$

Pinsker's inequality used KL-divergence to upper bound total variation. However, its constant term is not sharp. Thus, people still try to improve it in different settings. Here, I present the original

Pinsker's inequality and a simple proof.

**Theorem 6 (Pinsker's inequality)** $D_{KL}(P||Q) \geq 2d_{TV}^2(P,Q)$

**Proof:** We prove it in three steps: First, show that by data processing inequality, we can reduce to Bernoulli case. Next, prove the Bernoulli version of Pinsker's inequality. Finally, use the supremum representation of total variation to show the final result.

(i) Let $E$ be arbitrary event on $\mathcal{X}$ and take random variable $Y = \mathbf{1}_{X\in E}$. We have two reduced distribution from $P, Q$ to $Y$: $P_Y = Bern(P(E))$ and $Q_Y = Bern(Q(E))$. For simplicity, we denote $p = P(E)$ and $q = Q(E)$. By the data processing inequality, we have

$$D_{KL}(P||Q) \geq D_{KL}(P_Y||Q_Y) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

(ii) Observe that

$$\begin{aligned}
D_{KL}(P_Y||Q_Y) &= p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \\
&= p[\log p - \log q] + (1-p)[\log(1-p) - \log(1-q)] \\
&= p\int_q^p \frac{dt}{t} + (p-1)\int_q^p \frac{dt}{1-t} = \int_q^p \frac{p-t}{t(1-t)}dt \\
(\because \frac{1}{t(1-t)} \geq 4) &\geq 4\int_q^p (p-t)dt = 2(p-q)^2 \\
&= 2d_{TV}^2(P_Y, Q_Y)
\end{aligned}$$

(iii) Take $E^* = \arg\max_{E\subseteq\mathcal{X}} P(E) - Q(E)$ and denote the corresponding $Y$ as $Y^*$, by the supremum representation of total variation we know that $d_{TV}(P,Q) = d_{TV}(P_{Y^*}, Q_{Y^*})$. As a result,

$$D_{KL}(P||Q) \geq D_{KL}(P_{Y^*}||Q_{Y^*}) \geq 2d_{TV}^2(P_{Y^*}, Q_{Y^*}) = 2d_{TV}^2(P,Q)$$

■

**Exercise 3.1** Check whether the case $P \sim \mathcal{N}(\theta, \sigma^2)$, $Q \sim \mathcal{N}(0, \sigma^2)$ satisfies Pinsker's inequality? Is it tight?

### 3.2 Le Cam's inequality: $\frac{1}{2}H^2(P,Q) \leq d_{TV}(P,Q) \leq H(P,Q)\sqrt{1 - \frac{H(P,Q)}{4}}$

As we have seen before, total variation does not have a nice tensorization property. When it comes to high-dimension statistics, we must find some way to characterize the behavior of tensor product of total variation. Here, Le Cam's inequality used Hellinger distance to give both upper bound and lower bound for total variation.

**Theorem 7 (Le Cam's inequality)** $\frac{1}{2}H^2(P,Q) \leq d_{TV}(P,Q) \leq H(P,Q)\sqrt{1-\frac{H(P,Q)}{4}}$

**Proof:** The proof can be easily shown with Cauchy-Schwarz and two observations: for any $a,b>0$, $|a-b| = |\sqrt{a}-\sqrt{b}|(\sqrt{a}+\sqrt{b})$, and $|a-b| = a+b-2a\wedge b$.

- $\frac{1}{2}H^2(P,Q) \leq d_{TV}(P,Q)$:

$$\begin{aligned} d_{TV}(P,Q) &= \frac{1}{2}\int |P-Q| = \frac{1}{2}\int P+Q-2P\wedge Q \\ &\geq \frac{1}{2}\int P+Q-2\sqrt{PQ} = \frac{1}{2}\int(\sqrt{P}-\sqrt{Q})^2 \\ &= \frac{1}{2}H^2(P,Q) \end{aligned}$$

- $d_{TV}(P,Q) \leq H(P,Q)\sqrt{1-\frac{H(P,Q)}{4}}$:

$$\begin{aligned} d_{TV}(P,Q) &= \frac{1}{2}\int |P-Q| = \frac{1}{2}\int |\sqrt{P}-\sqrt{Q}|(\sqrt{P}+\sqrt{Q}) \\ &\leq \frac{1}{2}\Big(\int |\sqrt{P}-\sqrt{Q}|^2\Big)^{1/2}\Big(\int (\sqrt{P}+\sqrt{Q})^2\Big)^{1/2} \\ &= \frac{1}{2}H(P,Q)\Big(\int P+Q+2\sqrt{PQ}\Big)^{1/2} = \frac{1}{2}H(P,Q)\Big(4-\int P+Q-2\sqrt{PQ}\Big)^{1/2} \\ &= \frac{1}{2}H(P,Q)\sqrt{4-H^2(P,Q)} = H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}} \end{aligned}$$

■

With Le Cam's inequality, one can tightly bound the extreme case of total variation with Hellinger distance and vice versa. Note that, this is very useful in convergence analysis.

**Corollary 8 (bound for extreme case)**

- $H^2(P,Q)=2 \Leftrightarrow d_{TV}(P,Q)=1$.
- $H^2(P,Q)=0 \Leftrightarrow d_{TV}(P,Q)=0$.
- $H^2(P_n,Q_n)\to 0 \Leftrightarrow d_{TV}(P_n,Q_n)\to 0$.

By the nice tensorization property of Hellinger distance, one can find a sharp characterization for the convergence of tensor product of total variation.

**Corollary 9 (bound for convergence)** *For any sequence of distributions $P_n$ and $Q_n$, as $n\to\infty$*

$$\begin{aligned} d_{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0 &\Leftrightarrow H^2(P_n,Q_n) = o(\frac{1}{n}) \\ d_{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1 &\Leftrightarrow H^2(P_n,Q_n) = \omega(\frac{1}{n}) \end{aligned}$$

**Exercise 3.2** Check whether the case $P \sim \mathcal{N}(\theta, \sigma^2)$, $Q \sim \mathcal{N}(0, \sigma^2)$ satisfies Le Cam's inequality? Is it tight?

## 3.3 Joint range

We can see that Pinsker's inequality and Le Cam's inequality are ad hoc method for finding inequalities among f-divergences. Namely, their techniques cannot be generalized. Luckily, there's an interesting techniques: *joint range method*, which is a general approach for finding optimal inequality among f-divergences.

The idea is simple, for two f-divergences, say $D_f$ and $D_g$, we construct the following mapping: $(P, Q) \to (D_f(P||Q), D_g(P||Q))$ for any probability distribution $P$ and $Q$. Then, we have a 2-dimensional figure where each point corresponds to a possible $(D_f, D_g)$ pairs. Intuitively, this figure presents the relation among these two distributions. For example, when we fix $D_f = \alpha$, then we can find all the possible $D_g$ and yield upper and lower bound for $D_g$ given $D_f = \alpha$. As we let $\alpha$ moves along the range of $D_f$, we can characterize the relation between them. Specifically, this relation will be tight! Take a look at Figure 6 for example.

With these idea in mind, we can define the joint range of two f-divergences as follow.

**Definition 10 (joint range)** *Consider two f-divergences $D_f(\cdot||\cdot)$ and $D_g(\cdot||\cdot)$, we define their joint range as*

$$\mathcal{R} := \{(D_f(P||Q), D_g(P||Q)|P,\ Q \text{ are arbitrary probability measures on some measurable space}\}$$
$$\mathcal{R}_k := \{(D_f(P||Q), D_g(P||Q)|P,\ Q \text{ are arbitrary probability measures on } [k]\}$$

*where $\mathcal{R}$ is the joint and $\mathcal{R}_k$ is the joint range with dimension $k$ for $k = 2, 3, \dots$.*

As what we discussed previously, our goal is to find the **boundary** of $\mathcal{R}$. But as you can see, it's not very easy to do so straightforwardly. Luckily again, we have a theorem to ease out computation from infinite-dimensional probability space to 2-dimensional probability space.

**Theorem 11 (Harremoës-Vajda)**

$$\mathcal{R} = \boldsymbol{co}(\mathcal{R}_2)$$

Although the proof is really awesome, for the sake of conciseness, I omit the details and only present the key lemma, which is useful for us to write down the close form of $\mathcal{R}_2$.

**Lemma 12** *Consider two f-divergences $D_f(\cdot||\cdot)$ and $D_g(\cdot||\cdot)$, their joint range is*

$$\mathcal{R} = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]) \end{pmatrix} | \ X \geq 0, \mathbb{E}[X] \leq 1 \right\}$$

$$\mathcal{R}_k = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]) \end{pmatrix} | \ X \geq 0, \mathbb{E}[X] \leq 1, |\mathcal{X} = k-1| \ or \ X \geq 0, \mathbb{E}[X] = 1.|\mathcal{X}| = k \right\}$$

For example, consider the joint range of Hellinger distance and total variation.

**Example 13** We know that $H(P,Q) = \int(\sqrt{P}-\sqrt{Q})^2 = 2-\int 2\sqrt{PQ}$ and $d_{TV}(P,Q) = \frac{1}{2}\int|P-Q|$. From Lemma 12, we have

$$\mathcal{R}_2 = \{(2(1-\sqrt{pq}-\sqrt{(1-p)(1-q)}), \ |p-q|)| \ 0 \leq p \leq 1, 0 \leq q \leq 1\}$$

We can visualize in Figure 6.

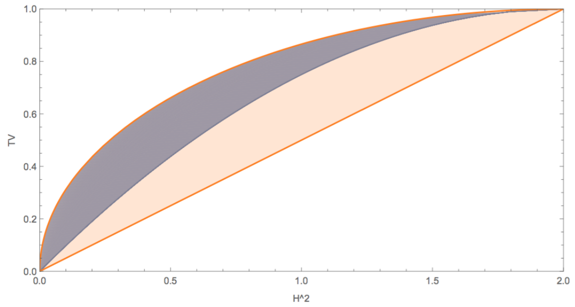

Figure 6: Joint range of Hellinger distance and total variation. The gray area is $\mathcal{R}_2$ and the entire area is $\mathcal{R}$.

**Exercise 3.3** Can you plot the joint range of $H^2$ versus $L$ where $L(P||Q) = \int \frac{(P-Q)^2}{P+Q}$?

## 4 Applications

Here, I will focus on two applications of f-divergences: *hypothesis testing* and *estimation lower bound*.

## 4.1 Hypothesis testing and total variation

Recall that in Section 2.2, we've seen that $d_{TV}(P,Q) = 1 - \int P \wedge Q$. Now, let's take a deep look at the term $\int P \wedge Q$ in Figure 7.

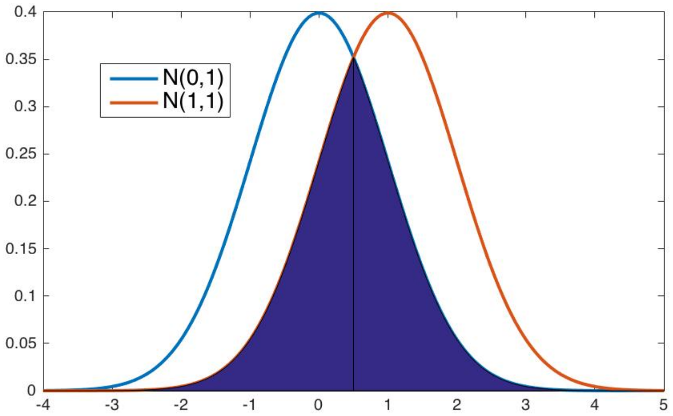

Figure 7: $\int P \wedge Q$: The dark blue area.

It turns out that $\int P \wedge Q$, which is the dark blue area in Figure 7, is exactly the minimal sum of type 1 and type 2 error of hypothesis testing over $P$ and $Q$!
Combine with Pinsker's inequality or Le Cam's inequality plus the tensorization of f-divergences, we can construct upper bound for the asymptotic error of hypothesis testing with multiple samples.

### Chernoff information

When we consider $n$ iid samples from either $P^{\otimes n}$ or $Q^{\otimes n}$, the minimal hypothesis testing error is $1 - d_{TV}(P^{\otimes n}, Q^{\otimes n})$. Surprisingly, the asymptotic behavior of this error is highly connected with the Chernoff theorem under Bayesian setting.

**Theorem 14 (Chernoff theorem)** *Let $P_\pi := \inf_{\phi_n:\mathcal{X}^n\to[0,1]} \pi_0\alpha_{\phi_n}^{(n)} + \pi_1\beta_{\phi_n}^{(n)}$, where $\pi$ is the prior of hypothesis testing and $\alpha_{\phi_n}^{(n)}, \beta_{\phi_n}^{(n)}$ are type I and type II error of $\phi_n$ respectively. We have*

$$\lim_{n\to\infty} \frac{-1}{n} \log P_\pi = C(P,Q)$$

*,where $C(P,Q) = -\log \inf_{\alpha\in[0,1]} \int dP^\alpha dQ^{1-\alpha}$ which is the* **Chernoff information**.

**Proof:** Proof can be found in Information theory textbook. ■

It turns out that the minimal hypothesis testing error also scales with Chernoff information. Thus, we have $d_{TV}(P^{\otimes n}, Q^{\otimes n}) = 1 - \exp(-nC(P,Q) + o(\frac{1}{n}))$.

## 4.2 HCR bound and $\chi^2-$divergence

Hammersley, Chapman, and Robbins showed an estimation lower bound with the variational representation of of $\chi^2-$divergence. For now, I cannot find good intuition to explain why this is the case. However, we can go through the proof and get some feelings about it.

**Theorem 15 (Hammersley-Chapman-Robbins (HCR) lower bound)** *Consider quadratic loss function, for any estimator $\hat{\theta}$, we have*

$$Var_\theta[\hat{\theta}] \geq \sup_{\theta' \neq \theta} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'}||P_\theta)}$$

*for all $\theta \in \Theta$. As long as $R_\theta = \mathbb{E}_\theta^2[\hat{\theta} - \theta] + Var_\theta[\hat{\theta}] \geq Var_\theta[\hat{\theta}]$ we have a lower bound for the risk.*

To prove Theorem 15, we need the following lemma.

**Lemma 16**

$$\chi^2(P||Q) \geq \frac{(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2}{Var_Q[X]}$$

**Proof:** Consider the supremum representation of $\chi^2-$divergence

$$\chi^2(P||Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h^2(X)] - 1$$

We restrict $h$ to be a linear function: $h(x) = ax+b$. Thus, we have a lower bound for $\chi^2-$divergence.

$$\chi^2(P||Q) \geq \sup_{a,b\in\mathbb{R}} \left\{2(a\mathbb{E}_P[X] + b) - \mathbb{E}_Q[(aX+b)^2] - 1\right\}$$

Compute the supremum by setting the partial derivatives to 0, we get

$$\begin{cases} a &= \frac{\mathbb{E}_P[X] - \mathbb{E}_Q[X]}{Var_Q[X]} \\ b &= \frac{\mathbb{E}_Q[X^2] - \mathbb{E}_P[X]\mathbb{E}_Q[X]}{Var_Q[X]} \end{cases}$$

Plug into the inequality and we yield the desired result. ■

Now, we can prove Theorem 15.

**Proof of Theorem 15:** Let $P = P_{\theta'}$, $Q = P_\theta$, and $P_{\hat{\theta}|\theta}$ denote the conditional distribution of estimator. We have $P_{\hat{\theta}} = P_{\hat{\theta|\theta}} \cdot P$ and $Q_{\hat{\theta}} = P_{\hat{\theta}|\theta} \cdot Q$. By data processing inequality, we have

$$\chi^2(P||Q) \geq \chi^2(P_{\hat{\theta}}||Q_{\hat{\theta}})$$

Moreover, by Lemma 16,

$$\chi^2(P_{\hat{\theta}}||Q_{\hat{\theta}}) \geq \frac{(\mathbb{E}_{P_{\hat{\theta}}}[\hat{\theta}] - \mathbb{E}_{Q_{\hat{\theta}}}[\hat{\theta}])^2}{Var_{Q_{\hat{\theta}}}[\hat{\theta}]} = \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{Var_\theta[\hat{\theta}]}$$

Combine the above two inequalities, we get the desired result. ■

With Theorem 15, we can easily derive the Cramér-Rao lower bound.

**Corollary 17 (Cramér-Rao (CR) lower bound)** *For any unbiased estimator $\hat{\theta}$ and $\theta \in \Theta$, we have*

$$Var_\theta[\hat{\theta}] \geq \frac{1}{I(\theta)}$$

*where $I(\theta)$ is the Fisher information defined as $I(\theta) := \mathbb{E}_\theta[(\frac{d\log P_\theta}{d\theta})^2]$.*

**Proof:** The proof is left for exercise. Hint: rewrite the Fisher information and apply Taylor's expansion on the $\chi-$divergence term.

■

**Exercise 4.1** Prove the Cramér-Rao (CR) lower bound.

# References

[Cho15] Chi-Ning Chou. Probability metrics and their inequalities. `http://prezi.com/s0unagzhjhwl/?utm_campaign=share&utm_medium=copy`, 2015.

[Pol01] David Pollard. Distances and affinities between measures. *Asymptopia*, 2001.

[Tsy09] Alexandre B Tsybakov. Introduction to nonparametric estimation. *Springer Series in Statistics (*, 2009.

## A Answer to exercises

**Answer to Exercise 2.1**
**Answer to Exercise 2.2**
**Answer to Exercise 2.3**

-
-
- From the previous variational form, it suffices to show that $\sup_{P_{X,Y}:P_X=P,P_Y=Q}\mathbb{P}[X=Y]=\int dP\wedge dQ$. Here, we show the result with two steps:

  (i) ($\sup_{P_{X,Y}:P_X=P,P_Y=Q}\mathbb{P}[X=Y]\leq\int dP\wedge dQ$)
  Observe that $\mathbb{P}[X=Y]=\int_z dP_{X,Y}(z,z)$. As $dP_{X,Y}(z,z)\leq dP_X(z)$ and $dP_{X,Y}(z,z)\leq dP_Y(z)$, we have $\mathbb{P}[X=Y]=\int_z dP_{X,Y}(z,z)$ for every $P_{X,Y}$.

(ii) ($\exists P^*_{X,Y}$ such that $\mathbb{P}[X=Y] = \int dP \wedge dQ$)
Take

$$P^*_{X,Y}(x,y) := \begin{cases} dP_X(x) \wedge dP_Y(y) & \text{, if } x = y \\ \frac{\max\{P_X(x)-P_Y(x),0\}\cdot\max\{P_Y(y)-P_X(x),0\}}{d_{TV}(P_X,P_Y)} & \text{, otherwise} \end{cases}$$

By checking the validity of this joint distribution and verifying the achievability, the coupling property of total variation is proved.

-

**Answer to Exercise 2.4**
**Answer to Exercise 2.5**
**Answer to Exercise 3.1**
**Answer to Exercise 3.2**
**Answer to Exercise 3.3**
**Answer to Exercise 4.1** From Theorem 15, we have $Var_\theta[\hat{\theta}] \geq \sup_{\theta' \neq \theta} \frac{(\mathbb{E}_\theta[\hat{\theta}]-\mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'||\theta})}$. As we assume $\hat{\theta}$ is unbiased, we have $\mathbb{E}_\theta[\hat{\theta}] = \theta$ and $\mathbb{E}_{\theta'}[\hat{\theta}] = \theta'$. Apply Taylor's expansion on $\chi^2(P_{\theta'}||P_\theta)$

$$\begin{aligned}\chi^2(P_\theta||P_{\theta'}) &= \int \frac{(dP_\theta - dP_{\theta'})^2}{dP_\theta} = \int \frac{[(\theta-\theta')\frac{\partial(dP_\theta)}{\partial\theta} + o((\theta-\theta')^2)]^2}{dP_\theta} \\ &= (\theta-\theta')^2 \int \frac{(\frac{\partial(dP_\theta)}{\partial\theta})^2}{dP_\theta} = (\theta-\theta')^2 I(\theta)\end{aligned}$$

Combine these two, we have $Var_\theta[\hat{\theta}] \geq \frac{1}{I(\theta)}$.