ECE 5630: Information Theory for Data Transmission, Security and Machine Learning 2/13/20

Lecture 6: f-Divergences

Lecturer: Prof. Ziv Goldfeld

Scriber: Rami Pellumbi , *Net ID*: rp534

Assistant Editor: Kia Khezeli

6.1 Preliminaries

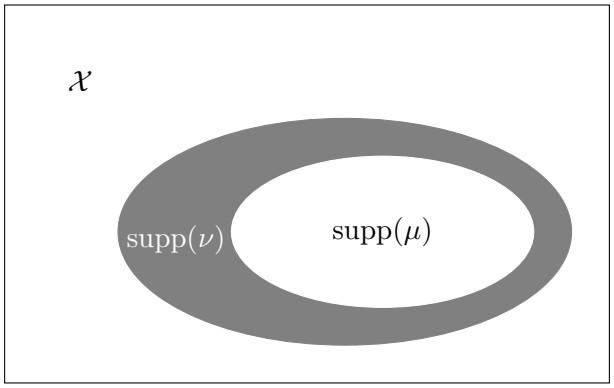
Before discussing f-divergences, we first need the notions of absolute continuity and the Radon-Nikodym theorem. To set these ideas, let $\mathcal{M}_+(\mathcal{X})$ be the set of all non-negative σ-finite measures $\mathcal{X}$ (we leave the σ-algebra implicit for this discussion). A non-negative measure μ is called σ-finite if there exist measurable sets $A_1, A_2, \ldots \subseteq \mathcal{X}$ with $\mu(A_n) < \infty$, for all n, such that $\bigcup_{n=1}^{\infty} A_n = \mathcal{X}$.

Definition 6.1 (Absolutely Continuous Measures) *For two measures $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$, we say that μ is absolutely continuous with respect to ν, denoted by $\mu \ll \nu$, if*

$$\nu(A) = 0 \implies \mu(A) = 0$$

for all measurable A. When $\mu \ll \nu$, we also say that ν dominates μ.

Remark 6.1 (Absolute Continuity and Supports) *If $\mu \ll \nu$, then* $\text{supp}(\mu) \subseteq \text{supp}(\nu)$.



Theorem 1 (Radon-Nikodym) *Let $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$ such that $\mu \ll \nu$. Then there exists a function $f \in L^1(\nu)$ such that for any measurable set A*

$$\mu(A) = \int_A f(x) \mathsf{d}\nu(x).$$

The function f is called the Radon-Nikodym derivative of μ with respect to ν, often denoted by $f = \frac{\mathsf{d}\mu}{\mathsf{d}\nu}$.

Example 6.1 (The Counting Measure) *Let $\nu = \#$ be the counting measure, where*

$$\#(A) = \begin{cases} |A| & , \ |A| < \infty \\ +\infty & , \ \textit{otherwise} \end{cases}$$

for any measurable A. If $\mu \ll \#$, then the $\text{supp}(\mu)$ *is countable and the Radon-Nikodym derivative $p := \frac{\mathsf{d}\mu}{\mathsf{d}\#}$ is the PMF of μ, i.e., $p(x) = \mu(\{x\})$, for all $x \in \text{supp}(\mu)$.*

Example 6.2 (The Lebesgue Measure) *Let $\mathcal{X} = \mathbb{R}^d$ and $\nu = \lambda$ be the Lebesgue measure on $\mathbb{R}^d$. If $\mu \ll \lambda$, then the the Radon-Nikodym derivative $p = \frac{\mathsf{d}\mu}{\mathsf{d}\lambda}$ is the PDF of μ, i.e., $\mu(A) = \int_A p(x) \mathsf{d}x$, for any measurable A.*

6.2 f-Divergences

We focus now on a class of divergences, called f-divergences, constructed from convex functions that satisfy certain conditions. As we shall see in future chapters, f-divergences frequently come up in solutions to operational compression, communication or inference questions. This chapter explores their properties, dual representations and some relation to generative modeling. We start with the definition of an f-divergence.

Definition 6.2 (f-divergence) *Let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a convex function such that:*

(i) f(1) = 0

(ii) f is strictly convex around 1, i.e.,

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)f(y)$$

for all $x, y \in \mathbb{R}_{\geq 0}$ and $\alpha \in [0,1]$ such that $\alpha x + (1-\alpha)y = 1$.

Let $P, Q \in \mathcal{P}(\mathcal{X})$ be two probability measures on $\mathcal{X}$, and let $\lambda \in \mathcal{M}_+(\mathcal{X})$ be a measure that dominates them both, i.e., $P, Q \ll \lambda$.[1] The f-divergence between Q and P is defined as

$$D_f(P\|Q) := \mathbb{E}_Q\left[f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right)\right] = \int_{\mathcal{X}} f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right)\mathsf{d}Q(x)$$

where $\mathsf{d}P/\mathsf{d}\lambda$ and $\mathsf{d}Q/\mathsf{d}\lambda$ are the Radon-Nikodym derivatives of P and Q, respectively, w.r.t λ.

Remark 6.2 (Conventions and Simplification) *We mention the following regarding the above definition:*

1. *It uses conventions $f(0) = f(0^+)$ and $0f(\frac{0}{0}) = 0$;*
2. *If $P \ll Q$, then $D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$, where $\frac{\mathsf{d}P}{\mathsf{d}Q}$ is the Radon-Nikodym derivative of P w.r.t. Q.*

Example 6.3 (Discrete or Continuous Distributions) *We specialize the Definition 6.2 to discrete or absolutely continuous (w.r.t. Lebesgue) distributions (see Item 2 in Remark 6.2):*

1. *<u>Discrete:</u> If $P \ll Q \ll \#$, where $\#$ is the counting measure, then*

$$D_f(P\|Q) = \sum_{x\in\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right)q(x)$$

with p and q as the PMFs of P and Q, respectively.

2. *<u>Continuous:</u> If $P \ll Q \ll \lambda$, where λ is the Lebesgue measure, then*

$$D_f(P\|Q) = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right)q(x)\mathsf{d}x$$

with p and q as the PDFs of P and Q, respectively.

[1] Notice that such λ always exists, e.g., $\lambda = P + Q$.

6.3 Important f-Divergences

We now focus on some important f-divergences that are commonly found in the literature. In what follows, all probability measures are defined over the same space $\mathcal{X}$.

6.3.1 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence (also sometimes referred to as relative entropy or information divergence) is the f-divergence induced by $f(x) = x \log x$. Namely, the KL divergence of Q from P is

$$D_{\mathsf{KL}}(P\|Q) = D_{x\log x}(P||Q) = \mathbb{E}_Q\left[f\left(\frac{dP/d\lambda}{dQ/d\lambda}\right)\right] = \mathbb{E}_Q\left[\frac{dP/d\lambda}{dQ/d\lambda}\log\left(\frac{dP/d\lambda}{dQ/d\lambda}\right)\right] = \mathbb{E}_P\left[\log\left(\frac{dP/d\lambda}{dQ/d\lambda}\right)\right].$$

Remark 6.3 (Comments) *We note the following:*

1. *If $P \ll Q \ll \#$, where $\#$ is the counting measure, then*

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

   with p and q as the PMFs of P and Q, respectively.

2. *If $P \ll Q \ll \lambda$, where λ is the Lebesgue measure, then*

$$D_{\mathsf{KL}}(P\|Q) = \int_{\mathcal{X}} \log\left(\frac{p(x)}{q(x)}\right) p(x)\mathsf{d}x$$

   with p and q as the PDFs of P and Q, respectively.

3. *If P, Q are two probability measures such that $P \not\ll Q$, then*

$$D_{\mathsf{KL}}(P\|Q) = \infty.$$

4. *Considering $f(x) = -\log x$, which is a convex function satisfying Items (i) and (ii) in Definition 6.2, we obtain*

$$D_{-\log x}(P||Q) = \mathbb{E}_Q\left[f\left(\frac{dP/d\lambda}{dQ/d\lambda}\right)\right] = \mathbb{E}_Q\left[-\log\frac{dP/d\lambda}{dQ/d\lambda}\right] = \mathbb{E}_Q\left[\log\frac{dQ/d\lambda}{dP/d\lambda}\right] = D_{\mathsf{KL}}(Q\|P).$$

6.3.2 Total Variation Distance

The Total Variation (TV) distance, is the f-divergence induced by $f(x) = \frac{1}{2}|x-1|$. Namely, the TV distance between Q and P is

$$\delta_{\mathsf{TV}}(P,Q) = D_{\frac{1}{2}|x-1|}(P||Q) = \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda} - 1\right|\right] = \frac{1}{2}\int_{\mathcal{X}}\left|\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda} - 1\right|\mathsf{d}Q = \frac{1}{2}\int_{\mathcal{X}}\left|\frac{\mathsf{d}P}{\mathsf{d}\lambda} - \frac{\mathsf{d}Q}{\mathsf{d}\lambda}\right|.$$

Remark 6.4 (Comments) *We make note of the following:*

1. *If $P \ll Q \ll \#$, where $\#$ is the counting measure, then*

$$\delta_{\mathsf{TV}}(P,Q) = \frac{1}{2}\sum_{x\in\mathcal{X}}|p(x) - q(x)| = \frac{1}{2}\big\|p(x) - q(x)\big\|_1.$$

   with p and q as the PMFs of P and Q, respectively.

2. *If $P \ll Q \ll \lambda$, where λ is the Lebesgue measure, then*

$$\delta_{\mathsf{TV}}(P,Q) = \frac{1}{2}\big\|p(x) - q(x)\big\|_{L^1(\mathbb{R}^d)},$$

where

$$\|f\|_{L^1(\mathbb{R}^d)} = \int |f| \mathsf{d}\lambda(x)$$

and p and q are the PDFs of P and Q, respectively.

3. *$\delta_{\mathsf{TV}}(P,Q)$ is a metric on $\mathcal{P}(\mathcal{X})$ (the immediately follows from the norm representation when the probability measure have PMFs or PDFs).*

4. *If $\operatorname{supp}(P) \cap \operatorname{supp}(Q) = \emptyset$, then $\delta_{\mathsf{TV}}(P,Q) = 1$.*

6.3.3 χ^2-Divergence

The χ^2-divergence is the f-divergence induced by $f(x) = (x-1)^2$. Namely, the χ^2 divergence between Q from P is

$$\chi^2(P\|Q) = D_{(x-1)^2}(P||Q) = \mathbb{E}_Q\left[\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda} - 1\right)^2\right].$$

Expanding the above reveals that there is not one-to-one correspondence between functions f and induced f-divergences. Indeed:

$$\mathbb{E}_Q\left[\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda} - 1\right)^2\right] = \int\left(\frac{\mathsf{d}P}{\mathsf{d}Q}\right)^2 \mathsf{d}Q - 2\int \frac{\mathsf{d}P}{\mathsf{d}Q}\mathsf{d}Q + \int \mathsf{d}Q = \mathbb{E}_Q\left[\left(\frac{\mathsf{d}P}{\mathsf{d}Q}\right)^2 - 1\right].$$

Remark 6.5 (Comments) *We make note of the following:*

1. *As shown above, the mapping $f \mapsto D_f$ is* not *injective.*

2. *If $P \not\ll Q$, then $\chi^2(P\|Q) = \infty$.*

6.4 Properties of f-Divergences

Having seen some popular f-divergences we now state some important properties.

Proposition 6.1 (Properties of f-Divergences) *For any $P, Q \in \mathcal{P}(\mathcal{X})$ dominated by a common measure $P, Q \ll \lambda$, and an f-divergence D_f, we have:*

1. ***Non-Negativity:*** *$D_f(P\|Q) \geq 0$ with equality if and only if $P = Q$.*

2. ***Convexity:*** *The mapping $(P,Q) \mapsto D_f(P\|Q)$ is jointly convex. Consequently, $P \mapsto D_f(P\|Q)$ is convex for fixed Q, and $Q \mapsto D_f(P\|Q)$ is convex for fixed P.*

3. ***Conditioning Increases f-Divergences:*** *Let $P_{Y|X}$ and $Q_{Y|X}$ be two transition kernels and $P_X \in \mathcal{P}(\mathcal{X})$. Define the conditional f-divergence by*

$$D_f\left(P_{Y|X}\big\|Q_{Y|X}\big|P_X\right) := \int_{\mathcal{X}} D_f\left(P_{Y|X}(\cdot|x)\big\|Q_{Y|X}(\cdot|x)\right) dP_X(x) = \mathbb{E}_{P_X}\left[D_f\left(P_{Y|X}(\cdot|X)\big\|Q_{Y|X}(\cdot|X)\right)\right]$$

and recall that

$$P_Y(\cdot) = \mathbb{E}_{P_X}\left[P_{Y|X}(\cdot|X)\right]$$
$$Q_Y(\cdot) = \mathbb{E}_{P_X}\left[Q_{Y|X}(\cdot|X)\right].$$

We have

$$D_f(P_Y\|Q_Y) \leq D_f\left(P_{Y|X}\big\|Q_{Y|X}\big|P_X\right).$$

4. ***Joint vs. Marginal:*** *If* $P_{XY} = P_X P_{Y|X}$ *and* $Q_{XY} = Q_X P_{Y|X}$*, then*

$$D_f\left(P_{XY}\|Q_{XY}\right) = D_f\left(P_X\|Q_X\right).$$

Proof: 1. By the definition of f-divergence we have

$$D_f\left(P\|Q\right) = \mathbb{E}_Q\left[f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right)\right] \geq f\left(\mathbb{E}_Q\left[\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right]\right) \geq f(1) = 0$$

where the first inequality follows from Jensen's inequality and the second from convexity of f. To prove equality, first assume that $P = Q$. Then,

$$D_f\left(P\|Q\right) = \mathbb{E}_Q\left[f(1)\right] = 0.$$

Now assume $D_f\left(P\|Q\right) = 0$. Then $f\left(\frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda}\right) = 0 \implies \frac{\mathsf{d}P/\mathsf{d}\lambda}{\mathsf{d}Q/\mathsf{d}\lambda} = 1$ since f is strongly convex at 1. It follows that $P = Q$.

2. To prove convexity of D_f consider the perspective function of f. Namely, the perspective of $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a function $g_f : \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \to \mathbb{R}$ defined by

$$g_f(x, y) = y f\left(\frac{x}{y}\right), \qquad \mathsf{dom}(g_f) = \left\{(x, y) \in \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} : \frac{x}{y} \in \mathsf{dom}(f)\right\}$$

For a convex function f, the perspective of f is also convex. That is,

$$\begin{aligned} g_f(\alpha(x_1, y_1) + (1-\alpha)(x_2, y_2)) &\leq \alpha g_f(x_1, y_1) + (1-\alpha) g_f(x_2, y_2) \\ &= \alpha y_1 f\left(\frac{x_1}{y_1}\right) + (1-\alpha) y_2 f\left(\frac{x_2}{y_2}\right) \end{aligned}$$

for all $\alpha \in [0, 1]$ and each (x_i, y_i), $i = 1, 2$, in $\mathsf{dom}(g_f)$.
Denote $\tilde{D}_f\left(P, Q\right) = D_f\left(P\|Q\right)$. Then for $(P_1, Q_1), (P_2, Q_2) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$ such that $P_1, P_2, Q_1, Q_2 \ll \lambda$

$$\begin{aligned} \tilde{D}_f\left(\alpha(P_1, Q_1) + (1-\alpha)(P_2, Q_2)\right) &= \int_{\mathcal{X}} f\left(\frac{\alpha \frac{\mathsf{d}P_1}{\mathsf{d}\lambda}(x) + (1-\alpha)\frac{\mathsf{d}P_2}{\mathsf{d}\lambda}(x)}{\alpha \frac{\mathsf{d}Q_1}{\mathsf{d}\lambda}(x) + (1-\alpha)\frac{\mathsf{d}Q_2}{\mathsf{d}\lambda}(x)}\right) \mathsf{d}\lambda \\ &\leq \int \alpha f\left(\frac{\frac{\mathsf{d}P_1}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q_1}{\mathsf{d}\lambda}(x)}\right) \mathsf{d}\lambda + (1-\alpha) f\left(\frac{\frac{\mathsf{d}P_2}{\mathsf{d}\lambda}(x)}{\frac{\mathsf{d}Q_2}{\mathsf{d}\lambda}(x)}\right) \mathsf{d}\lambda \\ &= \alpha \tilde{D}_f(P_1, Q_1) + (1-\alpha)\tilde{D}_f(P_2, Q_2) \end{aligned}$$

Thus the mapping $(P, Q) \mapsto D_f\left(P\|Q\right)$ is convex.

3. For $P_{XY} := P_X P_{Y|X}$ and $Q_{XY} := P_X Q_{Y|X}$, we have

$$\begin{aligned} D_f\left(P_Y\|Q_Y\right) &= D_f\left(\mathbb{E}_{P_X}\left[P_{Y|X}(\cdot|X)\right]\middle\|\mathbb{E}_{P_X}\left[Q_{Y|X}(\cdot|X)\right]\right) \\ &\leq \mathbb{E}_{P_X}\left[D_f\left(P_{Y|X}(\cdot|X)\middle\|Q_{Y|X}(\cdot|X)\right)\right] \\ &= D_f\left(P_{Y|X}\middle\|Q_{Y|X}|P_X\right), \end{aligned}$$

where the inequality uses convexity of D_f and Jensen's Inequality.

4. For $P_{XY} := P_X P_{Y|X}$ and $Q_{XY} := Q_X P_{Y|X}$, we have

$$\begin{aligned} D_f\left(P_{XY} \| Q_{XY}\right) &= \mathbb{E}_{Q_{XY}}\left[f\left(\frac{\mathsf{d}P_{XY}}{\mathsf{d}Q_{XY}}\right)\right] \\ &= \mathbb{E}_{Q_{XY}}\left[f\left(\frac{\mathsf{d}P_X P_{Y|X}}{\mathsf{d}Q_X P_{Y|X}}\right)\right] \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} f\left(\frac{\mathsf{d}P_X P_{Y|X}}{\mathsf{d}Q_X P_{Y|X}}\right) \mathsf{d}Q_{X,Y}(x,y) \\ &= \int_{\mathcal{X}} f\left(\frac{\mathsf{d}P_X}{\mathsf{d}Q_X}\right) \mathsf{d}Q_X(x) \\ &= \mathbb{E}_{Q_X}\left[f\left(\frac{\mathsf{d}P_X}{\mathsf{d}Q_X}\right)\right] \\ &= D_f\left(P_X \| Q_X\right) \end{aligned}$$

■