Lecture 4: f-divergences

September 4, 2025

Instructor: Shubhanshu Shekhar

In this lecture, we introduce the notion of f-divergences, discuss some of its main instances (relative entropy, total variation, Hellinger distance, and chi-squared divergence), and derive some useful inequalities comparing these divergences. Note: all logs starting with this lecture will be natural logarithms.

1 Definition and Some Examples

Definition 1.1. Let $f : (0, \infty)$ denote a convex function with $f(1) = 0$, and let $f(0) := \lim_{x \downarrow 0} f(x)$. Consider two probability measures P, Q on a common space $(\mathcal{X}, \mathcal{F})$, with $P \ll Q$ (i.e., $Q(E) = 0 \implies P(E) = 0$ for $E \in \mathcal{F}$). Then the f-divergence between P and Q is defined as

$$D_f(P \parallel Q) := \mathbb{E}_{X \sim Q}\left[f\left(\frac{dP}{dQ}(X) \right) \right],$$

where dP/dQ is the Radon-Nikodym derivative of P with respect to Q (here we have used the assumption that $P \ll Q$).

Remark 1.2. The function f associated with a divergence D_f is referred to as its generator. An interesting consequence of the definition is that two distinct function can induce the same divergence. In fact, let $g(x) = f(x) + c(x-1)$. Then, it immediately follows that

$$D_g(P \parallel Q) = \mathbb{E}_Q[g(dP/dQ)] = \mathbb{E}_Q[f(dP/dQ)] + c\mathbb{E}_Q[dP/dQ - 1] = \mathbb{E}_Q[f(dP/dQ)] = D_f(P \parallel Q).$$

Remark 1.3. In the general case of when $P \not\ll Q$, we can use the Lebesgue decomposition of P, that is, $P = P^{(a)} + P^{(s)}$, where $P^{(a)} \ll Q$ is the absolutely continuous part of P (w.r.t. Q), and $P^{(s)} \perp Q$ is the singular part. With this decomposition, the f-divergence between P and Q is defined as

$$D_f(P \parallel Q) := \mathbb{E}_{X \sim Q}\left[f\left(\frac{dP^{(a)}}{dQ}(X) \right) \right] + P^{(s)}(\mathcal{X}) f'(\infty),$$

where $f'(\infty) = \lim_{x \to 0} x f(1/x)$.

To get some intuition behind this definition, suppose P and Q have densities p and q respectively w.r.t. some common dominating measure, ν (say $P + Q$). Let $E_0 = \{x \in \mathcal{X} : q(x) = 0, \text{ and } p(x) > 0\}$. Then, we have

$$\int q f\left(\frac{p}{q}\right) d\nu = \int_{\mathcal{X} \setminus E_0} q f\left(\frac{p}{q}\right) d\nu + \int_{E_0} q f\left(\frac{p}{q}\right) d\nu = \int_{\mathcal{X} \setminus E_0} q f\left(\frac{p}{q}\right) d\nu + \int_{E_0} p \frac{q}{p} f\left(\frac{p}{q}\right) d\nu.$$

Since $q = 0$ and $p > 0$ on E_0, we can write the second integral with $\int_{E_0} p \left(\lim_{x \to 0} x f(1/x)\right) d\nu$, which is equal to $\int p f'(\infty) d\nu = f'(\infty) P(\{Q = 0\})$. The first integral is simply $\mathbb{E}_{X \sim Q}\left[f(dP^{(a)}/dQ) \right]$.

As always, the main objective behind introducing somewhat abstract definitions (such as f-divergence) is to allow a unified study of several, seemingly unrelated, quantities. Let us see how different choices of f give us important statistical divergence/distance measures.

- Relative entropy: with $f(x) = x \log x$, we get $D_{\mathrm{KL}}(P \parallel Q)$, and with $f(x) = -\log(x) + x - 1$, we recover $D_{\mathrm{KL}}(Q \parallel P)$. A symmetric generalization of relative entropy is the Jensen-Shannon Divergence, defined as

$$JS(P \parallel Q) = \frac{1}{2}\left(D_{\mathrm{KL}}\left(P \parallel \frac{P+Q}{2}\right) + D_{\mathrm{KL}}\left(Q \parallel \frac{P+Q}{2}\right)\right).$$

  Exercise: Find an f such that $D_f(P \parallel Q) = JS(P \parallel Q)$. As we will see later, this f does not have to be unique. An important modern application of the Jensen-Shannon divergence is in the design of Generative Adversarial Networks or GANs.

- Chi-squared distance: A popular measure used for instance in nonparametric goodness of fit tests, corresponds to $f(x) = (x-1)^2$ or equivalently $f(x) = x^2 - 1$ (recall Remark 1.2).

- The total variation metric between two probability measures is defined using $f(x) = |x - 1|/2$:

$$\begin{aligned} d_{TV}(P, Q) &= \frac{1}{2}\int \left|\frac{p}{q} - 1\right| q d\nu = \frac{1}{2}\left(\int_{p>q} (p-q) d\nu + \int_{p \le q} (q-p) d\nu\right) \\ &= P(p > q) - Q(p > q). \end{aligned}$$

- (Squared) Hellinger metric between P and Q is defined as

$$H^2(P, Q) = \int \left(\sqrt{p} - \sqrt{q}\right)^2 d\nu,$$

  which corresponds to $f(x) = (\sqrt{x} - 1)^2$. We can verify that $H(P, Q)$ is a metric on the space of probability measures.

Remark 1.4. Suppose P and Q have densities p, q, and let $\ell = p/q$ denote their likelihood ratio function. Then, $D_f(P \parallel Q)$ is a measure of how much does ℓ deviate from the value 1, where the penalty for deviating from 1 is characterized by the function f. Different choices of f penalized different behaviors:

- $f(x) = x \log x$ most heavily penalizes $\ell \gg 1$ regimes
- $f(x) = -\log x + x - 1$ penalizes the $\ell \approx 0$ regimes
- TV corresponds to the robust, absolute value influence
- H^2 correspond to L^2 norm of the densities after square-root map

1.1 Properties

We now record some properties of f-divergences under a simplified setting: Throughout, for simplicity, we will assume that P and Q have densities p and q with respect to the Lebesgue measure. We begin with the simplest property that states that f-divergences are nonnegative.

Nonnegativity. This is a simple consequence of the convexity of f. In particular, note that

$$D_f(P \parallel Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx \overset{\text{Jensen's}}{\geq} f\left(\int \frac{p(x)}{q(x)} q(x) dx\right) = f(1) = 0.$$

The last equality uses the requirement that $f(1) = 0$. Furthermore, if f is strictly convex at 1, then it is also true that $D_f(P \parallel Q) = 0$ implies $P = Q$.

Monotonicity. We had proved the following chain rule for relative entropy

$$D_{\mathrm{KL}}(P_{XY} \parallel Q_{XY}) = D_{\mathrm{KL}}(P_X \parallel Q_X) + D_{\mathrm{KL}}(P_{Y|X} \parallel Q_{Y|X} | P_X) \geq D_{\mathrm{KL}}(P_X \parallel Q_X).$$

The first equality above was obtained as a special consequence of the property of logarithms (changes products $p_{XY}(x,y) = p_X(x) p_{Y|X}(y|x)$ into sums), and this is generally not possible to obtain for general f. However, the inequality that establishes a "monotonicity" between marginal and joint relative entropy can still be deduced for general f.

$$\begin{aligned} D_f(P_{XY} \parallel Q_{XY}) &= \int q_X(x) dx \int f\left(\frac{p_{XY}(x,y)}{q_{XY}(x,y)}\right) q_{Y|X}(y|x) dy \\ &\overset{\text{Jensen's}}{\geq} \int q_X(x) f\left(\frac{p_X(x)}{q_X(x)} \int \frac{p_{Y|X}(y|x)}{q_{Y|X}(y|x)} q_{Y|X}(y|x) dy\right) dx \\ &= \int f\left(\frac{p_X(x)}{q_X(x)}\right) q_X(x) dx = D_f(P_X \parallel Q_X). \end{aligned}$$

Data Processing Inequality (DPI). Just as we can infer the DPI for relative entropy from the chain rule, we can also prove a similar result for the case of f-divergences. In particular, suppose P_X and Q_X are two distributions on an alphabet $\mathcal{X}$, and let P_Y and Q_Y denote the distributions obtained by passing P_X, Q_X through a common channel $P_{Y|X}$. Then, observe that

$$D_f(P_{XY} \parallel Q_{XY}) = \int q_X(x) p_{Y|X}(y|x) f\left(\frac{p_X(x) \cancel{p_{Y|X}(y|x)}}{q_X(x) \cancel{p_{Y|X}(y|x)}}\right) dx dy = D_f(P_X \parallel Q_X).$$

On the other hand, by the monotonicity property, we have

$$D_f(P_{XY} \parallel Q_{XY}) \geq D_f(P_Y \parallel Q_Y).$$

Together, these two previous inequalities give us the data processing inequality for f-divergences

$$D_f(P_X \parallel Q_X) \geq D_f(P_Y \parallel Q_Y), \quad \text{where} \quad P_Y = P_X P_{Y|X},\ Q_Y = Q_X P_{Y|X}.$$

Convexity. The joint convexity of D_f in (P, Q) follows by an identical perspective-based argument that we used for relative entropy in Lecture 3. We can also prove the convexity by an analog of the log-sum-inequality for general f.

Variational Formulation. We have the following for convex and lower-semicontinuous f and any function class $\mathcal{G}$:

$$D_f(P \parallel Q) \geq \sup_{g \in \mathcal{G}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(Y))], \quad \text{where} \quad f^*(y) = \sup_{x \in \mathbb{R}} \left(xy - f(x) \right),$$

with equality if and only if $\partial f(\ell) \cap \mathcal{G} \neq \emptyset$, where $\ell = dP/dQ$.

Following Nguyen et al. (2010), we proceed by observing that for convex lower-semicontinuous f, we have the following:

$$f(x) = \sup_{y \in \mathbb{R}} \left(xy - f^*(y) \right).$$

Hence, we have

$$\begin{aligned} D_f(P \parallel Q) &= \mathbb{E}_Q[f(\ell(X))] = \mathbb{E}_Q\left[\sup_y \left(y\ell(X) - f^*(y) \right)\right] \\ &= \mathbb{E}_Q\left[\sup_g \left(g(X)\ell(X) - f^*(g(X)) \right)\right] \geq \mathbb{E}_Q\left[\sup_{g \in \mathcal{G}} \left(g(X)\ell(X) - f^*(g(X)) \right)\right] \\ &\geq \sup_{g \in \mathcal{G}} \mathbb{E}_Q\left[\left(g(X)\ell(X) - f^*(g(X)) \right) \right] = \sup_{g \in \mathcal{G}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(X))]. \end{aligned}$$

2 Some Properties of Specific f-divergences

The three main properties discussed at the end of the previous section are satisfied by all f-divergences. In this section, we derive some useful relations that are specific to We now record some useful properties of these divergence measures that will be useful in later lectures.

Total Variation: The following are true:

1. The TV distance admits an alternative variational definition:

   $$TV(P, Q) = \sup_{E \text{ measurable}} |P(E) - Q(E)|,$$

   and infact $TV(P, Q) = P(E^*) - Q(E^*)$, where $E^* = \{p > q\}$.

2. $TV(P, Q) = 1 - \int_{\mathcal{X}} (p \wedge q) d\nu$.

3. $0 \leq TV(P, Q) \leq 1$, with $TV(P, Q) = 0$ if and only if $P = Q$, while $TV(P, Q) = 1$ if and only if $P \perp Q$.

The first statement implies that total variation is also an instance of an integral probability metric (IPM), in addition to being an f-divergence. In fact, it is the only statistical distance/divergence that is both an f-divergence and an IPM!

Proof. We proceed as follows:

1. We had already shown that $TV(P,Q) = P(p > q) - Q(p > q)$ where p, q denote the densities of P, Q. This implies that

$$TV(P,Q) \leq \sup_E |P(E) - Q(E)| = \sup_E P(E) - Q(E). \tag{1}$$

To show the other direction, consider any measurable E, and observe that

$$P(E) - Q(E) = \int_E (p-q)d\nu = \int_{E\cap E^*} (p-q)d\nu + \int_{E\cap (E^*)^c} (p-q)d\nu.$$

By definition, $(p-q) \leq 0$ on the set $(E^*)^c$, which means that the second integral in the RHS above is ≤ 0. Hence,

$$\sup_E P(E) - Q(E) \leq \sup_E \int_{E\cap E^*} (p-q)d\nu \leq \int_{E^*} (p-q)d\nu = TV(P,Q). \tag{2}$$

Together, (1) and (2) give the required equality.

2. Observe that

$$TV(P,Q) = \int_{p>q} \frac{p-q}{2} + \int_{p\leq q} \frac{q-p}{2} \quad \text{and} \quad 1 = \int_{p>q} \frac{p+q}{2} + \int_{p\leq q} \frac{p+q}{2}.$$

Subtracting the first equality from the second, we get

$$1 - TV(P,Q) = \int_{p>q} q + \int_{p\leq q} p = \int (p \wedge q),$$

which completes the proof.

3. $TV(P,Q) \geq 0$ follows from the nonnegativity of absolute values. If $TV(P,Q) = 0$, it means that $|P(E) - Q(E)| = 0$ for all $E \in \mathcal{F}_{\mathcal{X}}$, which implies that $P(E) = Q(E)$ for all $E \in \mathcal{F}_{\mathcal{X}}$. This is exactly the defining condition for $P = Q$.

Since $P(E), Q(E) \in [0,1]$ for all $E \in \mathcal{F}_{\mathcal{X}}$, we must have $P(E) - Q(E) \in [-1,1]$, which implies that $|P(E) - Q(E)| \leq 1$. Taking the supremum over all $E \in \mathcal{F}_{\mathcal{X}}$ implies that $TV(P,Q) \leq 1$. The case of $TV(P,Q) = 1$ implies that for every $\epsilon > 0$, there exists an E_ϵ such that $P(E_\epsilon) - Q(E_\epsilon) \geq 1 - \epsilon$.

Formally, for any $n \geq 1$, define E_n to be the set with $\epsilon = 1/n$. Then, $E^* := \cup_{n=1}^\infty E_n = \lim_{n\to\infty} \cup_{m\leq n} E_m$ satisfies $P(E^*) = 1$, and $Q(E^*) = 0$. Here, we have used the fact that $E^* \in \mathcal{F}_{\mathcal{X}}$ as it is a countable union of elements of $\mathcal{F}_{\mathcal{X}}$.

This completes the proof. $\square$

Hellinger Distance. The Hellinger distance satisfies the following properties:

1. $0 \leq H^2(P, Q) \leq 2$
2. If $P = \otimes_{i=1}^n P_i$, and $Q = \otimes_{i=1}^n Q_i$, then we have

$$H^2(P, Q) = 2\left(1 - \prod_{i=1}^n \left(1 - \frac{H^2(P_i, Q_i)}{2}\right)\right). \tag{3}$$

Proof. We proceed as follows:

1. The nonnegativity is follows by definition. For the upper bound, note that $(\sqrt{p} - \sqrt{q})^2 = p + q - 2\sqrt{pq} \leq p + q$. Furthermore, note that the upper bound holds with equality if and only if $pq = 0$ almost-surely; that is, when P and Q are singular.

2. To see (3):

$$\begin{aligned} H^2(P, Q) &= \int \left(2 - 2\sqrt{\prod p_i(x_i) q_i(x_i)}\right) \prod dx_i = 2 - 2\prod \int \sqrt{p_i(x_i) q_i(x_i)} dx_i \\ &= 2 - 2\prod_{i=1}^n \left(1 - \frac{H^2(P_i, Q_i)}{2}\right) = 2\left(1 - \prod_{i=1}^n \left(1 - \frac{H^2(P_i, Q_i)}{2}\right)\right). \end{aligned}$$

$\square$

The second property of Hellinger distance makes it more tractable for analyzing distances between two product measures as compared to TV which does not admit such a decomposition into marginals.

Chi-Squared Distance. We record a similar decomposition equality for the case of chi-squared divergence. Suppose $P = \otimes_{i=1}^n P_i$ and $Q = \otimes_{i=1}^n Q_i$. Then, we have

$$\chi^2(P \parallel Q) = \prod_{i=1}^n \left(1 + \chi^2(P_i \parallel Q_i)\right) - 1.$$

Proof. The proof of this result follows a similar path as the Hellinger case.

$$\begin{aligned} \chi^2(P \parallel Q) &= \int \left(\frac{\prod_i p_i(x_i)}{\prod_i q_i(x_i)} - 1\right)^2 \prod_i q_i(x_i) dx_i = -1 + \int \frac{\prod_i p_i^2(x_i)}{\prod_i q_i(x_i)} \prod_i dx_i \\ &= -1 + \prod_i \int \frac{p_i^2(x_i)}{q_i(x_i)} dx_i = -1 + \prod_{i=1}^n \left(1 + \chi^2(P_i \parallel Q_i)\right). \end{aligned}$$

The last equality simply uses the fact that

$$\chi^2(P_i \parallel Q_i) = \int \frac{p_i^2(x_i)}{q_i(x_i)} dx_i - 1 \implies \int \frac{p_i^2(x_i)}{q_i(x_i)} dx_i = 1 + \chi^2(P_i \parallel Q_i).$$

$\square$

2.1 Some Inequalities

TV and Relative Entropy. Let P, Q denote two distributions on the same alphabet. Then, we have the following (with relative entropy in log with base e)

$$TV(P,Q) \leq \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \parallel Q)} \qquad (Pinsker)$$
$$TV(P,Q) \leq 1 - e^{-D_{\mathrm{KL}}(P \| Q)}. \qquad (Bretagnolle - Huber)$$

Proof. Due to DPI, it suffices to prove Pinsker's inequality for Bernoulli random variables. In particular, observe that for any measurable set E,

$$D_{\mathrm{KL}}(P \parallel Q) \geq d_{KL}(P(E) \parallel Q(E)).$$

To show Pinsker's inequality for Bernoulli distributions is an exercise in calculus. We will follow the argument in Canonne (2022), by introducing $f(x) = p \log x + \bar{p} \log \bar{x}$. Then, observe that

$$\begin{aligned} d_{KL}(p \parallel q) = f(p) - f(q) = \int_p^q f'(x)dx = \int_p^q \left(\frac{p}{x} - \frac{1-p}{1-x}\right) dx = \int_p^q \frac{p-x}{x(1-x)} dx \\ \geq 4 \int_p^q (p-x)dx = 2(p-q)^2. \end{aligned}$$

Thus, we have

$$D_{\mathrm{KL}}(P \parallel Q) \geq \sup_E d_{KL}(P(E) \parallel Q(E)) \geq 2 \sup_E |P(E) - Q(E)|^2 = 2TV(P,Q)^2.$$

□

TV and Hellinger. The following inequalities establish a tight connection between TV and Hellinger.

$$\frac{1}{2} H^2(P,Q) \leq TV(P,Q) \leq H(P,Q) \sqrt{1 - \frac{H^2(P,Q)}{4}}.$$

Hellinger and Relative Entropy. The relative entropy (in base e) between two distribution is always larger than the squared Hellinger distance:

$$H^2(P,Q) \leq D_{\mathrm{KL}}(P \parallel Q).$$

Proof. The result follows essentially from the inequality that $\log(x) \leq x - 1$ for all $x > 0$. Suppose P and Q have densities p, q. Then, $D_{\mathrm{KL}}(P \parallel Q) = \int p \log(p/q) = -2 \int p \log(\sqrt{q/p}) \geq -2 \int p(\sqrt{q/p} - 1) = -2 \int (\sqrt{pq} - 1) = H^2(P,Q)$. □

Relative Entropy and Chi-Squared. Finally, we can show that

$$D_{\mathrm{KL}}(P \parallel Q) \leq \log(1 + \chi^2(P \parallel Q)) \leq \chi^2(P \parallel Q).$$

Proposition 2.1. *To summarize, here is the general (though suboptimal) of inequalities to remember:*

$$\frac{1}{2}H^2(P,Q) \leq TV(P,Q) \leq H(P,Q) \leq \sqrt{D_{KL}(P \parallel Q)} \leq \sqrt{\chi^2(P \parallel Q)}.$$

3 Application: Generative Adversarial Networks

One possible generator function for the Jensen-Shannon divergence is

$$f(x) = \frac{1}{2}\left[x \log x - (x+1)\log\left(\frac{x+1}{2}\right)\right], \quad \text{with} \quad f^*(y) = -\log(2 - e^x), \quad \text{for} \quad x < \log 2.$$

Hence, we can obtain the following variational representation of $JS(P \parallel Q)$ as

$$JS(P \parallel Q) = \sup_{g:\mathcal{X}\to(-\infty,\log 2)} \mathbb{E}_P[g(X)] + \mathbb{E}_Q[\log(2 - e^{g(X)})],$$

which on reparametrizing $h(x) = (1/2)e^{g(x)}$ simplifies to

$$JS(P \parallel Q) = \log 2 + \sup_{h:\mathcal{X}\to(0,1)} \mathbb{E}_P[\log h(X)] + \mathbb{E}_Q[\log(1 - h(X))].$$

Idea behind GANs: Let P denote some distribution we want to estimate, and let Q denote a model distribution; for example, obtained by passing $Z \sim P_Z = N(0, I_d)$ through a neural network G (called the "generator network"). Ignoring $\log 2$, the goal of generator is to find the best G from a family that minimizes the JS divergence between P and $G(Z)$:

$$\min_{G\in\mathcal{G}} \sup_{h:\mathcal{X}\to(0,1)} \mathbb{E}_P[\log h(X)] + \mathbb{E}_Z[\log(1 - h(G(Z)))].$$

In practice, these expectations are often in very high dimensional spaces (images, videos, etc.), and P is only known through a finite dataset X^n. Furthermore, it is usually computationally infeasible to evaluate the "sup" over all measurable $(0,1)$ valued functions. Instead, the sup is taken over another class of neural networks, called the discriminator D:

$$\inf_G \sup_D \frac{1}{n}\sum_{i=1}^n \log D(X) + \frac{1}{m}\sum_{i=1}^m \log(1 - D(G(Z_i))).$$

References

C. L. Canonne. A short note on an inequality between kl and tv. arXiv preprint arXiv:2202.07198, 2022.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. IEEE Transactions on Information Theory, 56(11): 5847–5861, 2010.