ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 4: Total variation/Inequalities between f-divergences

Lecturer: Yihong Wu *Scribe: Matthew Tsao, Feb 8, 2016 [Ed. Mar 22]*

Recall the definition of f-divergences from last time. If a function $f : \mathbb{R}^+ \to \mathbb{R}$ satisfies the following properties:

- f is a convex function.
- $f(1) = 0$.
- f is strictly convex at $x = 1$, i.e. for all x, y, α such that $\alpha x + \overline{\alpha} y = 1$, the inequality $f(1) < \alpha f(x) + \overline{\alpha} f(y)$ is strict.

Then the functional that maps pairs of distributions to $\mathbb{R}_+$ defined by

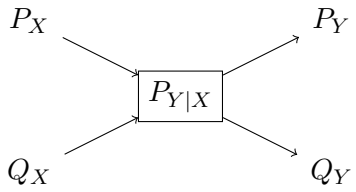
$$D_f(P\|Q) \triangleq \mathbb{E}_Q \left[f\left(\frac{dP}{dQ}\right)\right]$$

is an f-divergence.

4.1 Data processing inequality

Theorem 4.1. *Consider a channel that produces Y given X based on the law $P_{Y|X}$ (shown below). If P_Y is the distribution of Y when X is generated by P_X and Q_Y is the distribution of Y when X is generated by Q_X, then for any f-divergence $D_f(\cdot\|\cdot)$,*

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



One interpretation of this result is that processing the observation x makes it more difficult to determine whether it came from P_X or Q_X.

Proof.

$$\begin{aligned} D_f(P_X\|Q_X) &= \mathbb{E}_{Q_X}\left[f\left(\frac{P_X}{Q_X}\right)\right] \overset{(a)}{=} \mathbb{E}_{Q_{XY}}\left[f\left(\frac{P_{XY}}{Q_{XY}}\right)\right] = \mathbb{E}_{Q_Y}\left[\mathbb{E}_{Q_{X|Y}} f\left(\frac{P_{XY}}{Q_{XY}}\right)\right] \\ \text{Jensen's inequality} \to &\geq \mathbb{E}_{Q_Y}\left[f\left(\mathbb{E}_{Q_{X|Y}}\frac{P_{XY}}{Q_{XY}}\right)\right] \\ &= \mathbb{E}_{Q_Y}\left[f\left(\mathbb{E}_{P_{X|Y}}\frac{P_Y}{Q_Y}\right)\right] \overset{(b)}{=} \mathbb{E}_{Q_Y}\left[f\left(\frac{P_Y}{Q_Y}\right)\right] = D_f(P_Y\|Q_Y). \end{aligned}$$

Note that (a) means $D_f(P_X\|Q_X) = D_f(P_{XY}\|Q_{XY})$; (b) can be alternatively understood by noting that $\mathbb{E}_Q[\frac{P_{XY}}{Q_{XY}}|Y]$ is precisely the relative density $\frac{P_Y}{Q_Y}$, by checking the definition of change of measure, i.e., $\mathbb{E}_P[g(Y)] = \mathbb{E}_Q[g(Y)\frac{P_{XY}}{Q_{XY}}] = \mathbb{E}_Q[g(Y)\mathbb{E}[\frac{P_{XY}}{Q_{XY}}|Y]]$ for any g. □

Remark 4.1. $P_{Y|X}$ can be a deterministic map so that $Y = f(X)$. More specifically, if $f(X) = \mathbf{1}_E(X)$ for any event E, then Y is Bernoulli with parameter $P(E)$ or $Q(E)$ and the data processing inequality gives

$$D_f(P_X\|Q_X) \geq D_f(\text{Bern}(P(E))\|\text{Bern}(Q(E))). \tag{4.1}$$

This is how we prove the converse direction of large deviation.

Example 4.1. If $X = (X_1, X_2)$ and $f(X) = X_1$, then we have $D_f(P_{X_1X_2}\|Q_{X_1X_2}) \geq D_f(P_{X_1}\|Q_{X_1})$. As seen from the proof of Theorem 4.1, this is in fact equivalent to data processing inequality.

Remark 4.2. If $D_f(P\|Q)$ is an f-divergence, then $D_{\tilde{f}}(P\|Q)$ with $\tilde{f}(x) := xf(\frac{1}{x})$ is also an f-divergence and $D_f(P\|Q) = D_{\tilde{f}}(Q\|P)$. Example: $D_f(P\|Q) = D(P\|Q)$ then $D_{\tilde{f}}(P\|Q) = D(Q\|P)$.

Proof. First we verify that $\tilde{f}$ has all three properties required for $D_{\tilde{f}}(\cdot\|\cdot)$ to be an f-divergence.

- For $x, y \in \mathbb{R}^+$ and $\alpha \in [0,1]$ define $c = \alpha x + \overline{\alpha}y$ so that $\frac{\alpha x}{c} + \frac{\overline{\alpha}y}{c} = 1$. Observe that

  $$\tilde{f}(\alpha x + \overline{\alpha}y) = cf\left(\frac{1}{c}\right) = cf\left(\frac{\alpha x}{c}\frac{1}{x} + \frac{\overline{\alpha}y}{c}\frac{1}{y}\right) \leq c\frac{\alpha x}{c}f\left(\frac{1}{x}\right) + c\frac{\overline{\alpha}y}{c}f\left(\frac{1}{y}\right) = \alpha\tilde{f}(x) + \overline{\alpha}\tilde{f}(y).$$

  Thus $\tilde{f} : \mathbb{R}^+ \to \mathbb{R}$ is a convex function.

- $\tilde{f}(1) = f(1) = 0$.

- For $x, y \in \mathbb{R}^+$, $\alpha \in [0,1]$, if $\alpha x + \overline{\alpha}y = 1$, then by strict convexity of f at 1,

  $$0 = \tilde{f}(1) = f(1) = f\left(\alpha x\frac{1}{x} + \overline{\alpha}y\frac{1}{y}\right) < \alpha xf\left(\frac{1}{x}\right) + \overline{\alpha}yf\left(\frac{1}{y}\right) = \alpha\tilde{f}(x) + \overline{\alpha}\tilde{f}(y).$$

  So $\tilde{f}$ is strictly convex at 1 and thus $D_{\tilde{f}}(\cdot\|\cdot)$ is a valid f-divergence.

Finally,

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \mathbb{E}_P\left[\frac{Q}{P}f\left(\frac{P}{Q}\right)\right] = \mathbb{E}_P\left[\tilde{f}\left(\frac{Q}{P}\right)\right] = D_{\tilde{f}}(Q\|P).$$ □

4.2 Total variation and hypothesis testing

Recall that the choice of $f(x) = \frac{1}{2}|x-1|$ gives rise to the total variation distance,

$$D_f(P\|Q) = \frac{1}{2}\mathbb{E}_Q\left|\frac{P}{Q} - 1\right| = \frac{1}{2}\int |P - Q|,$$

where $\int |P-Q|$ is a short-hand understood in the usual sense, namely, $\int |\frac{dP}{d\mu} - \frac{dQ}{d\mu}|d\mu$ where μ is a dominating measure, e.g., $\mu = P + Q$, and the value of the integral does not depends on μ.

We will denote total variation by $d_{\mathrm{TV}}(P,Q)$ or $\mathrm{TV}(P,Q)$.

Theorem 4.2. *The following definitions for total variation are equivalent:*

1. $$d_{\mathrm{TV}}(P,Q) = \sup_E P(E) - Q(E), \tag{4.2}$$
   where the supremum is over all measurable set E.
2. $1 - d_{\mathrm{TV}}(P,Q)$ is the minimal sum of Type-I and Type-II error probabilities for testing P versus Q, and[1]
   $$d_{\mathrm{TV}}(P,Q) = 1 - \int P \wedge Q. \tag{4.3}$$
3. Provided the diagonal $\{(x,x) : x \in \mathcal{X}\}$ is measurable,
   $$d_{\mathrm{TV}}(P,Q) = \inf_{\substack{P_{XY}: \\ P_X = P, P_Y = Q}} \mathbb{P}\left[X \neq Y\right]. \tag{4.4}$$
4. Let $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}, \|f\|_\infty \leq 1\}$. Then
   $$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sup_{f\in\mathcal{F}} \mathbb{E}_P f(x) - \mathbb{E}_Q f(x). \tag{4.5}$$

Remark 4.3 (Variational representation)**.** The equation (4.2) and (4.5) provide sup-representation of total variation, which will be extended to general f-divergences (later). Note that (4.4) is an inf-representation of total variation in terms of couplings, meaning total variation is the Wasserstein distance with respect to Hamming distance. The benefit of variational representations is that choosing a particular coupling in (4.4) gives an upper bound on $d_{\mathrm{TV}}(P,Q)$, and choosing a particular f in (4.5) yields a lower bound.

Remark 4.4 (Operational meaning)**.** In the binary hypothesis test for $H_0 : X \sim P$ or $H_1 : X \sim Q$, Theorem 4.2 shows that $1 - d_{\mathrm{TV}}(P,Q)$ is the sum of false alarm and missed detection probabilities. This can be seen either from (4.2) where E is the decision region for deciding P or from (4.3) since the optimal test (for average probability of error) is the likelihood ratio test $\frac{dP}{dQ} > 1$. In particular,

- $d_{\mathrm{TV}}(P,Q) = 1 \Leftrightarrow P \perp Q$, the probability of error is zero since essentially P and Q have disjoint supports.
- $d_{\mathrm{TV}}(P,Q) = 0 \Leftrightarrow P = Q$ and the minimal sum of error probabilities is one, meaning the best thing to do is to flip a coin.

[1] Throughput the course $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$. Here again $\int P \wedge Q$ is a short-hand understood per the usual sense, namely, $\int (\frac{dP}{d\mu} \wedge \frac{dQ}{d\mu})d\mu$ where μ is any dominating measure.

4.3 Motivating example: Hypothesis testing with multiple samples

Observation: "Not all f-divergences are both equal"

1. Different f-divergence has different operational significance. For example, as we saw in Section 4.2, testing two hypothesis boils down to total variation, which determines the fundamental limit (minimum average probability of error). Later in the course we will encounter another f-divergence: $L(P\|Q) = \int \frac{(P-Q)^2}{P+Q}$, which is useful for estimation.
2. Some f-divergence is easier to evaluate than others. For example, for product distributions, Hellinger distance and χ^2-divergence **tensorize** in the sense that they are easily expressible in terms of those of the one-dimensional marginals; however, computing the total variation between product measures is frequently difficult. Another example is that computing the χ^2-divergence between a product distribution and a mixture of product distributions is convenient, which will become useful later in the course.

Therefore the punchline is that it is often fruitful to bound one f-divergence by another and this sometimes leads to tight characterizations. In this section we consider a specific useful example to drive this point home. Then in the next section we develop inequalities between f-divergences systematically.

Consider a binary hypothesis test where data $X = (X_1, X_2, ...X_n)$ are i.i.d drawn from either P or Q and the goal is to test

$$H_0 : X \sim P^{\otimes n} \quad \text{vs} \quad H_1 : X \sim Q^{\otimes n}.$$

As mentioned before, $1 - d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n})$ gives minimal Type-I+II probabilities of error, achieved by the maximum likelihood test. By the data processing inequality, $d_{\mathrm{TV}}(P^{\otimes m}, Q^{\otimes m}) \le d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n})$ for $m < n$. From this we see that $d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n})$ is an increasing sequence in n (and bounded by 1 by definition) and hence converges. One would hope that as $n \to \infty$, $d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n})$ converges to 1 and consequently, the probability of error in the hypothesis test converges to zero. It turns out that if the distributions P, Q are independent of n, then large deviation theory gives

$$d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n}) = 1 - \exp(-nC(P, Q) + o(n)) \tag{4.6}$$

where the constant $C(P, Q) = -\log \inf_{0\le\alpha\le1} \int P^\alpha Q^{1-\alpha}$ is the **Chernoff Information** of P, Q. It is clear from this that $d_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n}) \to 1$ as $n \to \infty$, and, in fact, exponentially fast.

However, as frequently encountered in high-dimensional problems, if the distributions $P = P_n$ and $Q = Q_n$ depend on n, then the large-deviation approach that leads to (4.6) is no longer valid. In such a situation, total variation is still relevant for hypothesis testing, but its behavior as $n \to \infty$ is not obvious nor easy to compute. In this case, understanding how a more computationally tractable f-divergence is related to total variation may give insight on hypothesis testing without needing to directly compute the total variation. It turns out Hellinger distance is precisely suited for this task – see Theorem 4.3 below.

Recall that the squared Hellinger distance, $H^2(P, Q) = \mathbb{E}_Q\left[\left(1 - \sqrt{\frac{P}{Q}}\right)^2\right]$ is an f-divergence with $f(x) = (1 - \sqrt{x})^2$, which provides a sandwich bound for total variation

$$0 \le \frac{1}{2}H^2(P, Q) \le d_{\mathrm{TV}}(P, Q) \le H(P, Q)\sqrt{1 - \frac{H^2(P, Q)}{4}} \le 1. \tag{4.7}$$

The proof of this statement will explained in the next lecture. A few observations which are direct consequences of these inequalities:

- $H^2(P,Q) = 2$, if and only if $d_{\mathrm{TV}}(P,Q) = 1$.
- $H^2(P,Q) = 0$ if and only if $d_{\mathrm{TV}}(P,Q) = 0$.
- Hellinger consistency $\Leftrightarrow$ TV consistency, namely $H^2(P_n,Q_n) \to 0 \Leftrightarrow d_{\mathrm{TV}}(P_n,Q_n) \to 0$.

Theorem 4.3. *For any sequence of distributions P_n and Q_n, as $n \to \infty$,*[2]

$$d_{\mathrm{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0 \Leftrightarrow H^2(P_n,Q_n) = o\left(\frac{1}{n}\right)$$
$$d_{\mathrm{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1 \Leftrightarrow H^2(P_n,Q_n) = \omega\left(\frac{1}{n}\right)$$

Proof. Because the observations $X = (X_1, X_2, ...X_n)$ are i.i.d, the joint distribution factors

$$\begin{aligned} H^2(P_n^{\otimes n}, Q_n^{\otimes n}) &= 2 - 2\mathbb{E}_{Q_n^{\otimes n}}\left[\sqrt{\prod_{i=1}^n \frac{P_n}{Q_n}(X_i)}\right] \\ \text{By independence} \to &= 2 - 2\prod_{i=1}^n \mathbb{E}_{Q_n}\left[\sqrt{\frac{P_n}{Q_n}(X_i)}\right] = 2 - 2\left(\mathbb{E}_{Q_n}\left[\sqrt{\frac{P_n}{Q_n}}\right]\right)^n \\ &= 2 - 2\left(1 - \frac{1}{2}H^2(P_n,Q_n)\right)^n \end{aligned}$$

$d_{\mathrm{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0$ which happens precisely when $\left(1 - \frac{1}{2}H^2(P_n,Q_n)\right)^n \to 1$, which happens when $H^2(P_n,Q_n) = o(\frac{1}{n})$.

Similarly, $d_{\mathrm{TV}}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \to 2$ which happens precisely when $\left(1 - \frac{1}{2}H^2(P_n,Q_n)\right)^n \to 0$, if and only if $H^2(P_n,Q_n) = \omega(\frac{1}{n})$. □

Remark 4.5. The proof of Theorem 4.3 relies on two ingredients:

1. Sandwich bound (4.7).
2. Tensorization properties of Hellinger:

$$H^2\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) = 2 - 2\prod_{i=1}^n \left(1 - \frac{H^2(P_i,Q_i)}{2}\right) \tag{4.8}$$

Note that there are other f-divergences that are also tensorizable, e.g., χ^2-divergences:

$$\chi^2\left(\prod_{i=1}^n P_i, \prod_{i=1}^n Q_i\right) = \prod_{i=1}^n \left(1 + \chi^2(P_i,Q_i)\right) - 1; \tag{4.9}$$

however, no sandwich inequality like (4.7) exists for d_{TV} and χ^2 and hence there is no χ^2-version of Theorem 4.3. Asserting the non-existence of such inequalities requires understanding the relationship between these two f-divergences.

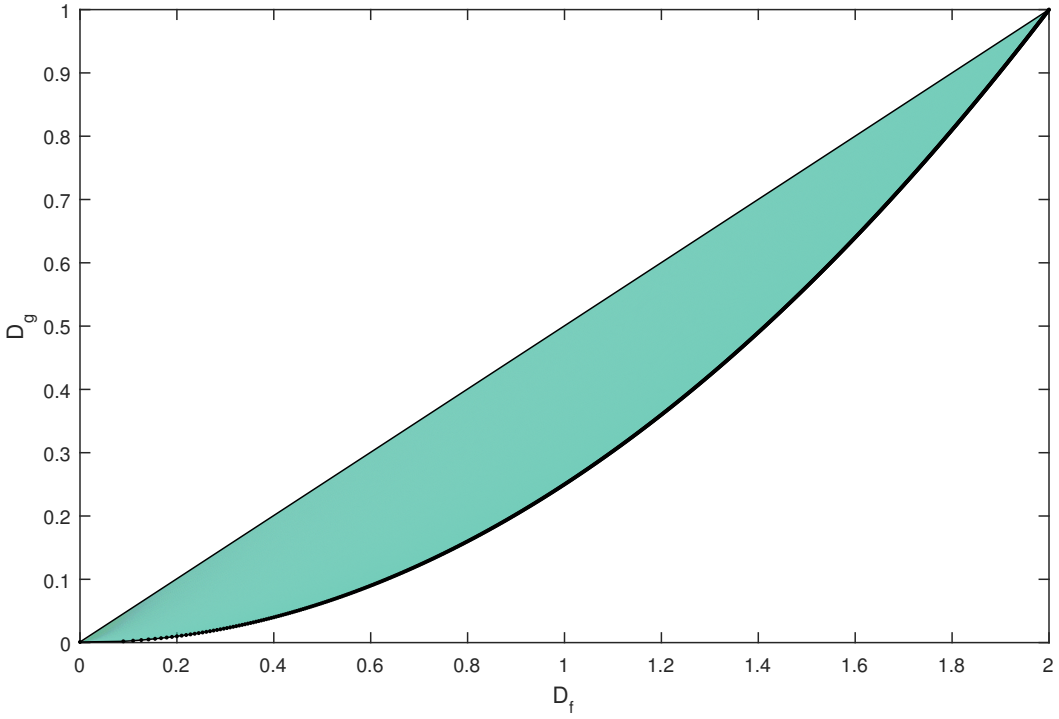
[2] For positive sequences $\{a_n\}, \{b_n\}$, we say $a_n = \omega(b_n)$ if $b_n = o(a_n)$.

4.4 Inequalities between f-divergences

We will discuss two methods for finding inequalities between f-divergences.

- ad hoc approach: case-by-case proof using results like Jensen's inequality, max $\leq$ mean $\leq$ min, Cauchy-Schwarz, etc.
- systematic approach: **joint range** of f-divergences.

Definition 4.1. The *joint range* between two f-divergences $D_f(\cdot\|\cdot)$ and $D_g(\cdot\|\cdot)$ is the range of the mapping $(P, Q) \mapsto (D_f(P\|Q), D_g(P\|Q))$, i.e., the set $\mathcal{R} \subset \mathbb{R}_+ \times \mathbb{R}_+$ where $(x, y) \in \mathcal{R}$ if there exist distributions P, Q on some common measurable space such that $x = D_f(P\|Q)$ and $y = D_g(P\|Q)$.



The green region in the above figure shows what a joint range between $D_f(\cdot\|\cdot)$ and $D_g(\cdot\|\cdot)$ might look like. By definition of $\mathcal{R}$, the lower boundary gives the sharpest lower bound of D_g in terms of D_f, namely:

$$D_f(P\|Q) \geq V(D_g(P\|Q)), \quad \text{where } V(t) \triangleq \inf\{D_f(P\|Q) : D_g(P\|Q) = t\};$$

similarly, the upper boundary gives the best upper bound. As will be discussed in the next lecture, the sandwich bound (4.7) correspond to precisely the lower and upper boundaries of the joint range of H^2 and d_{TV}, therefore not improvable. It is important to note, however, that $\mathcal{R}$ may be an unbounded region and some of the boundaries may not exist, meaning it is impossible to bound one by the other, such as χ^2 versus d_{TV}.

To gain some intuition, we start with the ad hoc approach by proving *Pinsker's inequality*, which bounds total variation from above by the KL divergence.

Theorem 4.4 (Pinsker's inequality)**.**

$$D(P\|Q) \geq 2d_{\mathrm{TV}}^2(P,Q). \tag{4.10}$$

Proof. First we show that, by the data processing inequality, it suffices to prove the result for Bernoulli distributions. For any event E, let $Y = \mathbf{1}\{X \in E\}$ which is Bernoulli with parameter $P(E)$ or $Q(E)$. By data processing inequality, $D(P\|Q) \geq d(P(E)\|Q(E))$. If Pinsker's inequality is true for all Bernoulli random variables, we have

$$\sqrt{\frac{1}{2}D(P\|Q)} \geq d_{\mathrm{TV}}(\mathrm{Bern}(P(E)), \mathrm{Bern}(Q(E)) = |P(E) - Q(E)|$$

Taking the supremum over E gives $\sqrt{\frac{1}{2}D(P\|Q)} \geq \sup_E |P(E) - Q(E)| = d_{\mathrm{TV}}(P,Q)$, in view of Theorem 4.2.

The binary case follows easily from Taylor's theorem:

$$d(p\|q) = \int_q^p \frac{p-t}{t(1-t)}dt \geq 4\int_q^p (p-t)dt = 2(p-q)^2$$

and $d_{\mathrm{TV}}(\mathrm{Bern}(p), \mathrm{Bern}(q)) = |p-q|$. □

Remark 4.6. Pinsker's inequality is known to be sharp in the sense that the constant "2" in (4.10) is not improvable, i.e., there exist $\{P_n, Q_n\}$ such that $\frac{\mathrm{LHS}}{\mathrm{RHS}} \to 2$ as $n \to \infty$. (Why?) Nevertheless, this does not mean that (4.10) itself is not improvable because it might be possible to subtract some higher-order term from the RHS. This is indeed the case and there are many refinements of Pinsker's inequality. But what is the best inequality? Settling this question rests on characterizing the joint range and the lower boundary. This is the topic of next lecture.