MOVE MORE STUFF HERE FROM THE DIVERGENCE SPACES DOCUMENT
OR MAYBE i want to dedicate the next section to only the total variation - and move everything else to the following section - “total variation norm” - “other divergences for measures”

0.1 Supremum metrics

Consider a set of functions with domain $\mathcal{X}$. We will say that a subset S of the domain *separates* functions if $x = y \Leftrightarrow [x(a) = y(a)\,\forall a \in S]$. That is, if any two functions agree on S, then they must agree everywhere. Recall, for instance, that if two measures agree on a generating class, then they agree on the measure that they assign to all sets (i.e. they are the same measure). So any generating class separates measures.

Let x and y be real-valued functions with the same domain, and assume that they are bounded on a subset S that separates functions. d is called a *supremum metric* if it is defined by

$$d(x, y) := \sup_{a\in S} |x(a) - y(a)|$$

1. We’ve already established in section (CITE) that such a d will satisfy the metric properties when the supremum is taken over all of $\mathcal{X}$. Show that if S separates functions, we still get the property $d(x, y) = 0 \Rightarrow x = y$.

0.1.1 Total variation distance

There are a number of important supremum metrics on the set of probability measures. (SIDENOTE: In the context of probability measures, S is often called the set of *test functions*. Recall that sets can be identified with their indicator functions.) The total variation distance (in addition to being an f-divergence!) is the supremum metric that uses the set of all measurable sets as its test functions.

$$d_{\mathrm{TV}}(P, Q) = \sup_{A\in\mathcal{B}} |PA - QA| \tag{1}$$

(SIDENOTE: Some authors define total variation distance to be the supremum over all possible partitions of the sample space of $\sum |PA_i - QA_i|$ summing over all A_i in the partition. This definition is nicer if you want to extend the distance to the vector space of all finite signed measures. In the case of probability measures, it’s simply equal to two times our definition, as the maximizing partition splits the space into two pieces: $\{p < q\}$ and $\{p \geq q\}$ which contribute equally.) $d_{\mathrm{TV}}(P, Q)$ is an upper bound for the difference in the probabilities that P and Q assign to any event. This formulation makes it clear that d_{TV} separates points.

2. Explain why $d^2_{\mathrm{TV}} \leq d_{\mathrm{TV}}$.

3. Show that the two formulations (**??**) and (3) of d_{TV} coincide.

The *Kolmogorov distance* d_{K} is also an important supremum metric on probability distributions on $(\mathbb{R}, \mathcal{B})$ by the test functions $S := \{(-\infty, t] : t \in \mathbb{R}\}$. Recall that this S is a generating set for $\mathcal{B}$, so it distinguishes measures (CITE SECTION). For probability measures P and Q, we can express d_{K} in terms of the cdfs F_P and F_Q.

$$\begin{aligned} d_K(P, Q) &:= \sup_{t\in\mathbb{R}} |P(-\infty, t] - Q(-\infty, t]| \\ &= \sup_{t\in\mathbb{R}} |F_P(t) - F_Q(t)| \end{aligned}$$

Two other notable supremum metrics on distributions on $(\mathbb{R}, \mathcal{B})$ are *Wasserstein distance* (with test functions Lip(1)) and *bounded Wasserstein distance* (with test functions Lip(1) $\cap$ Bdd(1)). (MAKE SURE I’ve defined Lipschitz functions somewhere! And that i’ve introduced the notation Bdd(a) for the set of functions bounded in absolute value by a.... ARE all the functions in Lip(1) measurable? OR do I need to specify the measurable subset of this?)

HOW do I know that Lip(1) $\cap B(1)$ is “big enough” to separate the finite signed measures?? Figure this out. Then maybe make it an Exercise.

1 Hellinger distance

OUT OF PLACE

4. Show that *Hellinger distance*

$$H(P, Q) := \sqrt{2[1 - A(P, Q)]}$$

is equal to the L_2 norm distance between the square root densities $\sqrt{p}$ and $\sqrt{q}$.

5. Use the identity (DOES THIS require non-negative a, b, or does it work fine for all of $\mathbb{R}$ by using complex roots?)

$$|a - b| = |\sqrt{a} - \sqrt{b}||\sqrt{a} + \sqrt{b}|$$

to prove that $d_{\mathrm{TV}}(P, Q) \leq H(P, Q)$. Then prove that

$$A(P, Q) \leq \sqrt{1 - d^2_{\mathrm{TV}}(P, Q)}$$

Squared Hellinger distance can be upper bounded by $2d_{\mathrm{TV}}$.

$$\begin{aligned} H^2(P, Q) &= \mu\,(\sqrt{p} - \sqrt{q})^2 \\ &\leq \mu\,[p + q - (p \wedge q) \\ &= \mu\,|p - q| \\ &= 2d_{\mathrm{TV}}(P, Q) \end{aligned}$$

(CITE RAMAMOORTHI Prop 1.2.1) Together with the inequality $d_{\mathrm{TV}}(P, Q) \leq H(P, Q)$ from Exercise (NUMBER), this tells us that convergence in Hellinger distance is equivalent to convergence in total variation distance. It also shows (along with Exercise (NUMBER)) that Hellinger affinity is squeezed between two functions of d_{TV}.

$$A(P, Q) \geq 1 - d_{\mathrm{TV}}$$

POINT out that it is a true metric, unlike many of the divergence that we’ve studied. SQUARED Hellinger distance is an f-divergence - not Hellinger distance itself (OF COURSE, because d_{TV} is the only f-divergence that is also a metric!)
Emphasize the aspects related to inner product.
NOT AN INNER PRODUCT SPACE for probability measures, because that’s not a vector space! What about for the set of all finite signed measures that are absolutely continuous with respect to the defining reference measure?
According to wikipedia, there is another f function for which squared Hellinger distance is an f-divergence. It is $f(t) = (\sqrt{t} - 1)^2$. Can I verify this? Make it an exercise? Does something like that extend to the general squared λ-Hellinger divergences, or is this a special case?

Solution

1. If the supremum is zero, then $|x(a) - y(a)|$ must be zero for all a. By the assumption that S separates functions, this tells us that x and y must be equal.

2. Formulation (3) and its accompanying interpretation tells us that d_{TV} must be in $[0, 1]$. So its squared value can be no larger.

3. We’ll start by considering ().

$$\begin{aligned} d_{\mathrm{TV}}(P, Q) &= \sup_{A\in\mathcal{B}} |PA - QA| \\ &= \sup_{A\in\mathcal{B}} |\mu\,pA - \mu\,qA| \\ &= \sup_{A\in\mathcal{B}} |\mu\,(p - q)A| \end{aligned}$$

Clearly the supremum of this quantity will correspond to A being either $\{q < p\}$ or $\{p < q\}$; otherwise there would be positive and negative parts of the integral fighting against each other. In fact, if $P\Omega = Q\Omega$ (for instance, if P and Q are probability distributions), these sets give the same result:

$$\begin{aligned} |\mu\,(p - q)\{q < p\}| &= \mu\,(p - q)\{q < p\} \qquad \text{non-negative integrand} \\ &= \mu\,(p - q)[1 - \{p \leq q\} \\ &= \mu\,(p - q)[1 - \{p < q\} \\ &= \underbrace{P\Omega - Q\Omega}_{0} + \mu\,(q - p)\{p < q\} \\ &= \mu\,|p - q|\{p < q\} \\ &= |\mu\,(p - q)\{p < q\}| \end{aligned}$$

So $d_{\mathrm{TV}}(P, Q) = |\mu\,(p - q)\{q < p\}| = |\mu\,(p - q)\{p < q\}|$.

Now let’s expand the other end (**??**), and see if we can make the strands meet.

$$\begin{aligned} \frac{1}{2}\mu\,|p - q| &= \frac{1}{2}[\mu\,|p - q|\{q < p\} + \mu\,|p - q|\{p \leq q\}] \\ &= \frac{1}{2}[\mu\,|p - q|\{q < p\} + \mu\,|p - q|\{p < q\}] \\ &= \frac{1}{2}[|\mu\,(p - q)\{q < p\}| + |\mu\,(p - q)\{p < q\}|] \end{aligned}$$

This is an average of two quantities that are equal to each other, so it is equal to both of them. And our derivation above shows that each of them is equal to $d_{\mathrm{TV}}(P, Q)$

4. Notice that the Hellinger affinity $A(P, Q) := \mu\,\sqrt{p}\sqrt{q}$ is the L_2 inner product between the square root densities. We’ll begin at the end, by simplifying $\|\sqrt{p} - \sqrt{q}\|^2$ and find that it takes us right to $H^2(P, Q)$.

$$\begin{aligned} \|\sqrt{p} - \sqrt{q}\|^2 &= \mu\,(\sqrt{p} - \sqrt{q})^2 \\ &= \mu\,[p + q - 2\sqrt{p}\sqrt{q}] \\ &= 2 - 2A(P, Q) \\ &= 2(1 - A(P, Q)) \\ &= H^2(P, Q) \end{aligned}$$

Taking the square root on both sides gives the desired result.

5. The identity in the problem statement is a hint to start with the half-L_1 distance formulation of d_{TV} then use Cauchy-Schwarz and invoke the result from Exercise (NUMBER).

$$\begin{aligned} d_{\mathrm{TV}}(P, Q) &= \frac{1}{2}\mu\,|p - q| \\ &= \frac{1}{2}\mu\,|\sqrt{p} - \sqrt{q}||\sqrt{p} + \sqrt{q}| \\ &\leq \frac{1}{2}\underbrace{\|\sqrt{p} - \sqrt{q}\|}_{H(P,Q)}\,\|\sqrt{p} + \sqrt{q}\| \\ &= H(P, Q)\frac{1}{2}[\mu\,(p + q + 2\sqrt{p}\sqrt{q})]^{1/2} \\ &= H(P, Q)\frac{1}{2}\sqrt{2[1 + A(P, Q)]} \\ &\leq H(P, Q) \end{aligned}$$

The last inequality used the fact that Hellinger affinities are bounded by 1.

In the above derivation, if we had substituted the definition of Hellinger distance before the final inequality, we would have gotten

$$\begin{aligned} d_{\mathrm{TV}}(P, Q) &\leq H(P, Q)\frac{1}{2}\sqrt{2[1 + A(P, Q)]} \\ &= \frac{1}{2}\sqrt{2[1 - A(P, Q)]}\sqrt{2[1 + A(P, Q)]} \\ &= \sqrt{1 - A^2(P, Q)} \\ \Rightarrow A(P, Q) &\leq \sqrt{1 - d^2_{\mathrm{TV}}(P, Q)} \end{aligned}$$

CITE RAMAMOORTHI Prop 1.2.1 and Corollary 1.2.1