Distances and Divergences for Probability Distributions

Andrew Nobel

October, 2020

Background

Basic question: How far apart (different) are two distributions P and Q?

- Measured through distances and divergences
- Used to define convergence of distributions
- Used to assess smoothness of parametrizations $\{P_\theta : \theta \in \Theta\}$
- Means of assessing the complexity of a family of distributions
- Key role in understanding the consistency of inference procedures
- Key ingredient in formulating lower and upper bounds on the performance of inference procedures

Kolmogorov-Smirnov Distance

Definition: Let P and Q be probability distributions on $\mathbb{R}$ with CDFs F and G. The Kolmogorov-Smirnov (KS) distance between P and Q is

$$\text{KS}(P, Q) \; = \; \sup_t |F(t) - G(t)|$$

Properties of Total Variation

1. $0 \leq \text{KS}(P, Q) \leq 1$
2. $\text{KS}(P, Q) = 0$ iff $P = Q$
3. KS is a metric
4. $\text{KS}(P, Q) = 1$ iff there exists $s \in \mathbb{R}$ with $P((-\infty, s]) = 1$ and $Q((s, \infty)) = 1$

Total Variation Distance

Definition: Let $\mathcal{X}$ be a set with a sigma-field $\mathcal{A}$. The total variation distance between two probability measures P and Q on $(\mathcal{X}, \mathcal{A})$ is

$$\mathsf{TV}(P, Q) \;=\; \sup_{A \in \mathcal{A}} |P(A) - Q(A)|$$

Properties of Total Variation

1. $0 \leq \mathsf{TV}(P, Q) \leq 1$
2. $\mathsf{TV}(P, Q) = 0$ iff $P = Q$
3. TV is a metric
4. $\mathsf{TV}(P, Q) = 1$ iff there exists $A \in \mathcal{A}$ with $P(A) = 1$ and $Q(A) = 0$

KS, TV, and the CLT

Note: KS(P, Q) and TV(P, Q) can both be expressed in the form

$$\sup_{A \in \mathcal{A}_0} |P(A) - Q(A)|$$

For KS family $\mathcal{A}_0 =$ all intervals $(-\infty, t]$, while for TV family $\mathcal{A}_0 =$ all (Borel) sets

Example: Let $X_1, X_2, \ldots \in \{-1, 1\}$ iid with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. By the standard central limit theorem

$$Z_n \;=\; \frac{1}{n^{1/2}} \sum_{i=1}^{n} X_i \;\Rightarrow\; \mathcal{N}(0, 1)$$

Let $P_n =$ distribution of Z_n and $Q = \mathcal{N}(0, 1)$. Can show that

$$\text{KS}(P_n, Q) \;\leq\; cn^{-1/2} \quad \text{while} \quad \text{TV}(P_n, Q) \equiv 1$$

Total Variation and Densities

Scheffé's Theorem: Let $P \sim f$ and $Q \sim g$ be distributions on $\mathcal{X} = \mathbb{R}^d$. Then

1. TV$(P, Q) \ = \ \frac{1}{2} \int |f(x) - g(x)| \, dx$
2. TV$(P, Q) = 1 - \int \min\{f(x), g(x)\} \, dx$
3. TV$(P, Q) = P(A) - Q(A)$ where $A = \{x : f(x) \geq g(x)\}$

Analogous results hold when $P \sim p(x)$ and $Q \sim q(x)$ are described by pmfs

Upshot: Total variation distance between P and Q is half the L_1-distance between densities or mass functions

Total Variation and Hypothesis Testing

Problem: Observe $X \in \mathcal{X}$ having density f_0 or f_1. Wish to test

$$\text{H}_0 : X \sim f_0 \;\text{ vs. }\; \text{H}_1 : X \sim f_1$$

Any decision rule $d : \mathcal{X} \to \{0, 1\}$ has overall (Type I + Type II) error

$$\text{Err}(d) \;=\; \mathbb{P}_0(d(X) = 1) + \mathbb{P}_1(d(X) = 0)$$

Fact: The optimum overall error among *all* decision rules is

$$\inf_{d:\mathcal{X}\to\{0,1\}} \text{Err}(d) \;=\; \int \min\{f_0(x), f_1(x)\}\, dx \;=\; 1 - \text{TV}(P_0, P_1)$$

Total Variation and Coupling

Definition: A *coupling* of distributions P and Q on $\mathcal{X}$ is a jointly distributed pair of random variables (X, Y) such that $X \sim P$ and $Y \sim Q$

Fact: $\text{TV}(P, Q)$ is the minimum of $\mathbb{P}(X \neq Y)$ over all couplings of P and Q

- If $X \sim P$ and $Y \sim Q$ then $\mathbb{P}(X \neq Y) \geq \text{TV}(P, Q)$
- There is an optimal coupling achieving the lower bound
- Optimal coupling makes X, Y equal as much as possible

Note: If ρ is a metric on $\mathcal{X}$ the Wasserstein distance between distributions P and Q is defined by $\min \mathbb{E}[\rho(X, Y)]$ where the minimum is over all couplings (X, Y) of P and Q.

Hellinger Distance

Definition: Let $P \sim f$ and $Q \sim g$ be probability measures on $\mathbb{R}^d$. The Hellinger distance between P and Q is given by

$$\mathsf{H}(P,Q) = \left[\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx\right]^{1/2}$$

Properties of Total Variation

1. $\mathsf{H}(P,Q)$ is just the L_2 distance between $\sqrt{f}$ and $\sqrt{g}$
2. $\mathsf{H}^2(P,Q) = 2\left(1 - \int\sqrt{f(x)g(x)}\,dx\right)$, therefore $0 \leq \mathsf{H}^2(P,Q) \leq 2$
3. $\mathsf{H}(P,Q) = 0$ iff $P = Q$
4. H is a metric
5. $\mathsf{H}^2(P,Q) = 2$ iff there exists $A \in \mathcal{A}$ with $P(A) = 1$ and $Q(A) = 0$

Hellinger Distance vs. Total Variation

Fact: For any pair of densities f, g we have the following inequalities

$$\int \min(f,g)\,dx \;\geq\; \frac{1}{2}\left(\int \sqrt{fg}\,dx\right)^2 \;=\; \frac{1}{2}\left(1 - \frac{1}{2}\mathsf{H}^2(f,g)\right)^2$$

Fact: For any distributions P and Q

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

- $\mathsf{H}^2(P,Q) = 0$ iff $\mathsf{TV}(P,Q) = 0$ and $\mathsf{H}^2(P,Q) = 2$ iff $\mathsf{TV}(P,Q) = 1$
- $\mathsf{H}(P_n, Q_n) \to 0$ iff $\mathsf{TV}(P_n, Q_n) \to 0$

Kullback-Liebler (KL) Divergence

Definition: The *KL-divergence* between distributions $P \sim f$ and $Q \sim g$ is given by

$$\mathrm{KL}(P : Q) \;=\; \mathrm{KL}(f : g) \;=\; \int f(x) \log \frac{f(x)}{g(x)} \, dx$$

Analogous definition holds for discrete distributions $P \sim p$ and $Q \sim q$

- The integrand can be positive or negative. By convention

$$f(x) \log \frac{f(x)}{g(x)} \;=\; \begin{cases} +\infty & \text{if } f(x) > 0 \text{ and } g(x) = 0 \\ 0 & \text{if } f(x) = 0 \end{cases}$$

- KL divergence is not symmetric, and is not a metric. Note that

$$\mathrm{KL}(P : Q) \;=\; \mathbb{E}_f \left[\log \frac{f(X)}{g(X)} \right]$$

First Properties of KL Divergence

Fact: The integral defining KL$(P : Q)$ is well defined. Letting $u_- = \max(-u, 0)$,

$$\int \left(f(x) \log \frac{f(x)}{g(x)} \right)_- dx \; < \; \infty$$

Key Fact:

- Divergence KL$(P : Q) \geq 0$ with equality if and only if $P = Q$
- KL$(P : Q) = +\infty$ if there is a set A with $P(A) > 0$ and $Q(A) = 0$

Notation: When pmfs or pdfs clear from context, write KL$(p : q)$ or KL$(f : g)$

KL Divergence Examples

Example: Let p and q be pmfs on $\{0,1\}$ with

$$p(0) = p(1) = 1/2 \quad \text{and} \quad q(0) = (1-\epsilon)/2,\ q(1) = (1+\epsilon)/2$$

Then we have the following exact expressions, and bounds

- KL$(p:q) = -\frac{1}{2}\log(1-\epsilon^2) \leq \epsilon^2$ when $\epsilon \leq \frac{1}{\sqrt{2}}$
- KL$(q:p) = \frac{1}{2}\log(1-\epsilon^2) + \frac{\epsilon}{2}\log(\frac{1-\epsilon}{1+\epsilon}) \leq 2\epsilon^2$

Example: If $P \sim \mathcal{N}_d(\mu_0, \Sigma_0)$ and $Q \sim \mathcal{N}_d(\mu_1, \Sigma_1)$ with $\Sigma_0, \Sigma_1 > 0$ then

$$2\,\mathrm{KL}(P:Q) \;=\; \mathrm{tr}(\Sigma_1^{-1}\,\Sigma_0) + (\mu_1 - \mu_0)^t \Sigma_1^{-1}(\mu_1 - \mu_0) + \ln(|\Sigma_1|/|\Sigma_0|) - d$$

KL Divergence and Inference

Ex 1. (Testing) Consider testing $\mathrm{H}_0 : X \sim f_0$ vs. $\mathrm{H}_1 : X \sim f_1$. The divergence

$$\mathrm{KL}(f_0 : f_1) \; = \; \mathbb{E}_0 \left(\log \frac{f_0(X)}{f_1(X)} \right) \; \geq \; 0$$

is just the expected log likelihood ratio under H_0

Ex 2. (Estimation) Suppose $X_1, X_2, \ldots$ iid with $X_i \sim f(x|\theta_0)$ in $\mathcal{P} = \{f(x|\theta) : \theta \in \Theta\}$. Under suitable assumptions, when n is large,

$$\hat{\theta}_{\mathrm{MLE}}(x) \; \approx \; \operatorname*{argmin}_{\theta \in \Theta} \mathrm{KL}(f(\cdot|\theta_0) : f(\cdot|\theta))$$

In other words, MLE is trying to find θ minimizing KL divergence with true distribution.

KL Divergence vs Total Variation and Hellinger

Fact: For any distributions P and Q we have

(1) $\mathsf{TV}(P,Q)^2 \leq \mathsf{KL}(P:Q)/2$ (Pinsker's Inequality)

(2) $\mathsf{H}(P,Q)^2 \leq \mathsf{KL}(P:Q)$

Log Sum Inequality

Log-Sum Inequality: If $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ are non-negative then

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \;\geq\; \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality iff all the ratios a_i/b_i are equal

Corollary: If $P \sim p$ and $Q \sim q$ are distributions, then for every event B

$$\sum_{x \in B} p(x) \log \frac{p(x)}{q(x)} \;\geq\; P(B) \log \frac{P(B)}{Q(B)}$$

with equality iff $p(x)/q(x)$ is constant for $x \in B$

Product Densities (Tensorization)

Recall: Given distributions $P_1, \ldots, P_n$ on $\mathcal{X}$ with densities $f_1, \ldots, f_n$ the product distribution $P = \otimes_{i=1}^n P_i$ on $\mathcal{X}^n$ has density $f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$

Tensorization: Let $P_1, \ldots, P_n$ and $Q_1, \ldots, Q_n$ be distributions on $\mathcal{X}$

- $\mathsf{TV}(\otimes_{i=1}^n P_i, \otimes_{i=1}^n Q_i) \le \sum_{i=1}^n \mathsf{TV}(P_i, Q_i)$
- $\mathsf{H}^2(\otimes_{i=1}^n P_i, \otimes_{i=1}^n Q_i) \le \sum_{i=1}^n \mathsf{H}^2(P_i, Q_i)$
- $\mathsf{KL}(\otimes_{i=1}^n P_i, \otimes_{i=1}^n Q_i) = \sum_{i=1}^n \mathsf{KL}(P_i, Q_i)$

Distinguishing Coins

Given: Observations $X = X_1, \ldots, X_n \in \{0, 1\}$ iid $\sim$ Bern(θ) with $\theta \in \{\theta_0, \theta_1\}$

Goal: Find a decision rule $d : \{0, 1\}^n \to \{0, 1\}$ such that

- $\star$ $\mathbb{P}_0(d(X) = 1) \leq \alpha$
- $\star$ $\mathbb{P}_1(d(X) = 0) \leq \alpha$

Question: How large does the number of observations n need to be?

Fact: Let $\Delta = |\theta_0 - \theta_1|$. Then there exists a decision procedure achieving performance $(\star)$ and requiring number of observations

$$n = \frac{2 \log(1/\alpha)}{\Delta^2}$$

Identifying Fair and Biased Coins

Suppose now that $\theta_0 = 1/2$ and $\theta_1 = 1/2 + \epsilon$ for some fixed $\epsilon \in (0, 1/4)$

Fact: For every event $A \subseteq \{0,1\}^n$

$$|\mathbb{P}_0(X \in A) - \mathbb{P}_1(X \in A)| \; = \; |P_0(A) - P_1(A)| \; \leq \; \epsilon\sqrt{2n}$$

Fact: If $d : \{0,1\}^n \to \{0,1\}$ is any decision rule achieving ($\star$) then

$$n \; \geq \; \frac{1 - 2\alpha}{\epsilon^2}$$