Theoretical Statistics, STOR 655

Total Variation Distance and Kullback-Liebler Divergence

Andrew Nobel

February 2023

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 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 KS distance

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 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 sup over all intervals $(-\infty, t]$, while for TV sup over 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

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

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

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

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

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

Coupling and Total Variation

Fact: Let P and Q be distributions on $\mathcal{X}$. Then

$$\mathsf{TV}(P, Q) \;=\; \min_{(X,Y)} \mathbb{P}(X \neq Y)$$

where the minimum is over all joint distributions (X, Y) such that $X \sim P$ and $Y \sim Q$. A joint distribution of this sort is called a *coupling*

Corollary

- If $X \sim P$ and $Y \sim Q$ are defined on the same probability space then $\mathbb{P}(X = Y) \leq 1 - \mathsf{TV}(P, Q)$
- There is an optimal coupling achieving the upper bound, which makes X and Y equal as much as possible

Kullback-Liebler (KL) Divergence

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

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

Analogous definition 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

First Properties of KL Divergence

Fact: Divergence $\mathrm{KL}(P : Q)$ is well defined: if $u_- = \max(-u, 0)$ then

$$\int \left(f(x) \log \frac{f(x)}{g(x)} \right)_- dx \;\leq\; 1$$

Key Fact:

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

Notation: When pmfs/pdfs clear from context, write $\mathrm{KL}(p : q)$ or $\mathrm{KL}(f : g)$

KL Divergence Examples

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

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

where $\epsilon \in (0,1)$. Then we have

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

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

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

KL Divergence and Inference

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

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

is just the expected log likelihood ratio under H_0

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

$$\hat{\theta}_n^{\text{MLE}}(X_1^n) \;\approx\; \operatorname*{argmin}_{\theta \in \Theta} \text{KL}(f(\cdot|\theta_0) : f(\cdot|\theta))$$

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

Data Processing Inequality

- Measurable spaces $(\mathcal{X}, \mathcal{A})$ with measures P and Q
- Measurable function $f : \mathcal{X} \to \mathcal{Y}$ from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$
- Map f pushes P and Q forward to measures $\tilde{P}$ and $\tilde{Q}$ on $(\mathcal{Y}, \mathcal{B})$ where

$$\tilde{P}(B) = P(f^{-1}B) \text{ and } \tilde{Q}(B) = Q(f^{-1}B)$$

Data Processing Inequality: Application of f reduces divergence, namely

$$\text{KL}(\tilde{P} : \tilde{Q}) \leq \text{KL}(P : Q)$$

Result extends to stochastic transformations (transition kernels) from $\mathcal{X}$ to $\mathcal{Y}$

Variational Formulation and Convexity

Fact: Let P and Q be distributions on $(\mathcal{X}, \mathcal{A})$. Then

$$\mathsf{KL}(P : Q) \;=\; \sup_f \left[\int f \, dP - \log \left(\int e^f \, dQ \right) \right]$$

where the supremum is over all functions $f : \mathcal{X} \to \mathbb{R}$ such that $\int e^f \, dQ$ is finite

Corollary: For each distribution Q on $(\mathcal{X}, \mathcal{A})$ the function $\mathsf{KL}(\cdot : Q)$ is convex: if P_1, P_2 are distributions and $\alpha \in (0, 1)$ then

$$\mathsf{KL}(\alpha P_1 + (1-\alpha) P_2 : Q) \;\leq\; \alpha \mathsf{KL}(P_1 : Q) + (1-\alpha) \mathsf{KL}(P_2 : Q)$$

Product Densities (Tensorization)

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

Tensorization: Let $P_1, \ldots, P_n$ and $Q_1, \ldots, Q_n$ be distributions on $\mathcal{X}$ with densities $f_1, \ldots, f_n$ and $g_1, \ldots, g_n$, respectively. Then

1. $\text{KS}(\otimes_{i=1}^n P_i, \otimes_{i=1}^n Q_i) \leq \sum_{i=1}^n \text{KS}(P_i, Q_i)$
2. $\text{TV}(\otimes_{i=1}^n P_i, \otimes_{i=1}^n Q_i) \leq \sum_{i=1}^n \text{TV}(P_i, Q_i)$
3. $\text{KL}(\otimes_{i=1}^n P_i : \otimes_{i=1}^n Q_i) = \sum_{i=1}^n \text{KL}(P_i, Q_i)$

Kullback Liebler vs Total Variation

Pinsker's Inequality: For any distributions P and Q on $(\mathcal{X}, \mathcal{A})$,

$$\mathsf{KL}(P : Q) \;\geq\; 2\,\mathsf{TV}(P : Q)^2$$