Chapter 3

Total variation distance between measures

1. Why bother with different distances?

When we work with a family of probability measures, $\{\mathbb{P}_\theta : \theta \in \Theta\}$, indexed by a metric space Θ, there would seem to be an obvious way to calculate the distance between measures: use the metric on Θ. For many problems of estimation, the obvious is what we want. We ask how close (in the metric) we can come to guessing θ_0, based on an observation from $\mathbb{P}_{\theta_0}$; we compare estimators based on rates of convergence, or based on expected values of loss functions involving the distance from θ_0.

When the parametrization is reasonable (whatever that means), distances measured by the Θ metric are reasonable. (What else could I say?) However it is not hard to concoct examples where the Θ metric is misleading.

<1> **Example.** Let $\mathbb{P}_{n,\theta}$ denote the joint distribution for n independent observations from the $N(\theta, 1)$ distribution, with $\theta \in \mathbb{R}$. Under $\mathbb{P}_{n,\theta_0}$, the sample average, $\bar{X}_n$, converges to θ_0 at a $n^{-1/2}$ rate. The parametrization is reasonable.

What happens if we reparametrize, replacing the $N(\theta, 1)$ by a $N(\theta^3, 1)$? We are still fitting the same model—same probability measures, only the labelling has changed. The maximum likelihood estimator, $\bar{X}_n^{1/3}$, still converges at an $n^{-1/2}$ rate if $\theta_0 \neq 0$, but for $\theta_0 = 0$ we get an $n^{-1/6}$ rate, as an artifact of the reparametrization.

More imaginative reparametrizations can produce even stranger behaviour for the maximum likelihood estimator. For example, define the one-to-one reparametriztion

$$\psi(\theta) = \begin{cases} \theta & \text{if } \theta \text{ is rational} \\ \theta + 1 & \text{if } \theta \text{ is irrational} \end{cases}$$

Now let $\mathbb{P}_{n,\theta}$ denote the joint distribution for n independent observations from the $N(\psi(\theta), 1)$ distribution. If θ_0 is rational, the maximum likelihood estimator, $\psi^{-1}(\bar{X}_n)$, gets very confused: it concentrates around $\theta_0 - 1$ as n gets larger. □

You would be right to scoff at the second reparametrization in the Example, yet it does make the point that distances measured in the Θ metric, for some parametrization picked out of the air, might not be particularly informative about the behaviour of estimators. Less ridiculous examples arise routinely in "nonparametric" problems, that is, in problems where "infinite dimensional" parameters enter, making the choice of metric less obvious.

Fortunately, there are intrinsic ways to measure distances between probability measures, distances that don't depend on the parametrizations. The rest of this Chapter will set forth a few of the basic definitions and facts. The

total variation distance has properties that will be familiar to students of the Neyman-Pearson approach to hypothesis testing. The ***Hellinger distance*** is closely related to the total variation distance—for example, both distances define the same topology of the space of probability measures—but it has several technical advantages derived from properties of inner products. (Hilbert spaces have nicer properties than general Banach spaces.) For example, Hellinger distances are very well suited for the study of product measures (Section 5). Also, Hellinger distance is closely related to the concept called Hellinger differentiability (Chapter 6), an elegant alternative to the tradional assumptions of pointwise differentiability in some asymptotic problems. Kullback-Leibler distance, which is also kown as relative entropy, emerges naturally from the study of maximum likelihood estimation. The relative entropy is not a metric, but it is closely related to the other two distances, and it too is well suited for use with product measures. See Section 5 and Chapter 4.

The intrinsic measures of distance are the key to understanding minimax rates of convergence, as you will learn in Chapter 18.

For reasonable parametrizations, in classical finite-dimensional settings, the intrinsic measures usually tell the same story as the Θ metric, as explained in Chapter 6.

2. Total variation and lattice operations

In classical analysis, the total variation of a function f over an interval $[a, b]$ is defined as

$$v(f, [a, b]) := \sup_{\mathfrak{g}} \sum_{i=1}^{k} |f(t_i) - f(t_{i-1})|,$$

where the supremum runs over all finite grids $\mathfrak{g} : a = t_0 < t_1 < \ldots < t_k = b$ on $[a, b]$.

The total variation of a signed measure μ, on a sigma-field $\mathcal{A}$ of subsets of some $\mathcal{X}$, is defined analogously (Dunford & Schwartz 1958, Section III.1):

$$v(\mu) := \sup_{\mathfrak{g}} \sum_{i=1}^{k} |\mu A_i|,$$

where now the supremum runs over all finite partitions $\mathfrak{g} : \mathcal{X} = \sum_{i=1}^{k} A_i$ of $\mathcal{X}$ into disjoint $\mathcal{A}$-sets.

> REMARK. The simplest way to create a signed measure is by taking a difference $\mu_1 - \mu_2$ of two nonnegative measures. In fact, one of the key properties of signed measures with finite total variation is that they can always be written as such a difference.

In fact, there is no need to consider partitions into more than two sets: for if $A = \cup_i \{A_i : \mu A_i \geq 0\}$ then

$$\sum\nolimits_i |\mu A_i| = \mu A - \mu A^c = |\mu A| + |\mu A^c|$$

That is,

$$v(\mu) := \sup_{A \in \mathcal{A}} \left(|\mu A| + |\mu A^c| \right).$$

If μ has a density m with respect to a countably additive, nonnegative measure λ then the supremum is achieved by the choice $A = \{m \geq 0\}$:

$$v(\mu) := \sup_{A \in \mathcal{A}} \left(|\lambda A m| + |\lambda A^c m| \right) = |\lambda \{m \geq 0\} m| + |\lambda \{m < 0\} m| = \lambda |m|$$

That is, $v(\mu)$ equals the $\mathcal{L}^1(\lambda)$ norm of the density $d\mu/d\lambda$, for every choice of dominating measure λ. This fact suggests the notation $\|\mu\|_1$ for the total variation of a signed measure μ.

The total variation $v(\mu)$ is also equal to $\sup_{|f|\le 1} |\mu f|$, the supremum running over all $\mathcal{A}$-measurable functions f bounded in absolute value by 1. Indeed,

$$|\mu f| = \lambda|mf| \le \lambda|m| \qquad \text{if } |f| \le 1,$$

with equality when $f = \{m \ge 0\} - \{m < 0\}$.

When $\mu(\mathcal{X}) = 0$, there are some slight simplifications in the formulae for $v(\mu)$. In that case, $0 = \lambda m = \lambda m^+ - \lambda m^-$ and hence

$$v(\mu) = \|\mu\|_1 = 2\lambda m^+ = 2\lambda m^- = 2\mu\{m \ge 0\} = 2\sup_{A\in\mathcal{A}} \mu A$$

As a special case, for probability measures P_1 and P_2, with densities p_1 and p_2 with respect to λ,

$$\begin{aligned} v(\mu) = \|\mu\|_1 &= 2\lambda(p_1 - p_2)^+ = 2\lambda(p_2 - p_1)^+ \\ &= 2\sup_A (P_1A - P_2A) = 2\sup_A (P_1A - P_2A) \\ &= 2\sup_A |P_1A - P_2A| \end{aligned} \tag{<2>}$$

Many authors, no doubt with the special case foremost in their minds, define the total variation as $\sup_A |P_1A - P_2A|$. An unexpected extra factor of 2 can cause confusion. To avoid this confusion I will abandon the notation $v(\mu)$ altogether and write $\|\mu\|_{\mathrm{TV}}$ for the modified definition.

In summary, for a finite signed measure μ on a sigma-field $\mathcal{A}$, with density m with respect to a nonnegative measure λ,

$$\|\mu\|_1 := \lambda|m| = \sup_{|f|\le 1} |\mu f| = \sup_{A\in\mathcal{A}} \left(|\mu A| + |\mu A^c|\right)$$

If $\mu\mathcal{X} = 0$ then

$$\tfrac{1}{2}\|\mu\|_1 = \|\mu\|_{\mathrm{TV}} := \sup_A |\mu A| = \sup_A \mu A = -\inf_A \mu A.$$

The use of an arbitrarily chosen dominating measure also lets us perform lattice operations on finite signed measures. For example, if $d\mu/d\lambda = m$ and $d\nu/d\lambda = n$, with $m, n \in \mathcal{L}^1(\lambda)$, then the measure γ defined by $d\gamma/d\lambda := m \vee n$ has the property that

$$\gamma A = \lambda\left((m \vee n)A\right) \ge \max\left(\mu A, \nu A\right) \qquad \text{for all } A \in \mathcal{A}. \tag{<3>}$$

In fact, γ is the smallest measure with this property. For suppose γ_0 is a nother signed measure with $\gamma_0 A \ge \max\left(\mu A, \nu A\right)$ for all $\mathcal{A}$. We may assume, with no loss of generality, that γ_0 is also dominated by λ, with density g_0. The inequality

$$\lambda\left(mA\{m \ge n\}\right) = \mu A\{m \ge n\} \le \gamma_0 A\{m \ge n\} = \lambda\left(g_0\{m \ge n\}A\right),$$

for all $A \in \mathcal{A}$, implies $g_0 \ge m$ a.e. $[\lambda]$ on the set $\{m \ge n\}$. Similarly $g_0 \ge n$ a.e. $[\lambda]$ on the set $\{m < n\}$. Thus $g_0 \ge m \vee n$ a.e. $[\lambda]$ and $\gamma_0 \ge \gamma$, as measures on $\mathcal{A}$.

Even though γ was defined via a particular choice of dominating measure λ, the setwise properties show that the resulting mesure is the same for every such λ.

<4> **Definition.** *For each pair of finite, signed measures μ and ν on $\mathcal{A}$, there is a smallest signed measure $\mu \vee \nu$ for which*

$$(\mu \vee \nu)(A) \ge \max\left(\mu A, \nu A\right) \qquad \text{for all } A \in \mathcal{A}$$

and a largest signed measure $\mu \wedge \nu$ for which

$$(\mu \wedge \nu)(A) \ge \min\left(\mu A, \nu A\right) \qquad \text{for all } A \in \mathcal{A}$$

If λ is a dominating (nonnegative measure) for which $d\mu/d\lambda = m$ and $d\nu/d\lambda = n$ then

$$\frac{d(\mu \vee \nu)}{d\lambda} = \max(m, n) \qquad \text{and} \qquad \frac{d(\mu \wedge \nu)}{d\lambda} = \min(m, n) \qquad \text{a.e. } [\lambda].$$

In particular, the nonnegative measures defined by $d\mu^+/d\lambda := m^+$ and $d\mu^-/d\lambda := m^-$ are the smallest measures for which $\mu^+ A \geq \mu A \geq -\mu^- A$ for all $A \in \mathcal{A}$.

REMARK. Note that the set function $A \mapsto \max\left(\mu A, \nu A\right)$ is not, in general, a measure because it need not be additive. In general, $(\mu \vee \nu)(A)$ is strictly greater than $\max\left(\mu A, \nu A\right)$.

<5> **Example.** If μ and ν are finite signed measures on $\mathcal{A}$ with densities m and n with respect to a dominating λ then $(\mu - \nu)^+ + (\nu - \mu)^+$ has density $(m - n)^+ + (n - m)^+ = |m - n|$. Thus

$$(\mu - \nu)^+(\mathcal{X}) + (\nu - \mu)^+(\mathcal{X}) = \lambda|m - n| = \|\mu - \nu\|_1.$$

Similarly,

$$(\mu-\nu)^+(\mathcal{X})-(\nu-\mu)^+(\mathcal{X}) = \lambda\left((m - n)^+ - (n - m)^+\right) = \lambda(m-n) = \mu\mathcal{X}-\nu\mathcal{X}.$$

In particular, if $\mu\mathcal{X} = \nu\mathcal{X}$ then $(\mu - \nu)^+(\mathcal{X}) = (\nu - \mu)^+(\mathcal{X}) = \|\mu - \nu\|_{\text{TV}}$. □

<6> **Example.** If μ and ν are finite signed measures on $\mathcal{A}$, then

$$(\mu \wedge \nu)(A) = \inf\{\mu(fA) + \nu(gA) : f + g = 1 \text{ and } f, g \geq 0\}$$

The infimum here runs of all pairs of nonnegative, measurable functions f, g for which $f(x) + g(x) = 1$ everywhere. again the assertion is easy to establish by expressing it in terms of a dominating λ. For f and g as above,

$$\lambda\left((m \wedge n)A\right) \leq \lambda\left((fm + gn)A\right),$$

with equality when $f = \{m < n\}$.

In particular, if μ and ν are nonnegative measures,

$$\mu \wedge \nu(\mathcal{X}) = \inf\{\mu f + \nu g : f + g = 1 \text{ and } f, g \geq 0\} = \|\mu \wedge \nu\|_1,$$

a quantity that is sometimes called the ***affinity*** between μ and ν and denoted by $\alpha_1(\mu, \nu)$. When μ and ν are probability measures,

$$2\alpha_1(\mu, \nu) = 2\lambda(m \wedge n) = \lambda\left(m + n - |m - n|\right) = 2 - \|\mu - \nu\|_1.$$

Equivalently, for probability measures μ and ν,

$$\alpha_1(\mu, \nu) + \|\mu - \nu\|_{\text{TV}} = 1.$$

Some arguments involving total variation distances belong clearer when reexpressed in terms of affinities. □

<7> **Example.** Suppose P and Q are probability measures on $(\mathcal{X}, \mathcal{A})$. If X and Y are random elements of $\mathcal{X}$ with distributions P and Q,then for every A in $\mathcal{A}$,

$$\begin{aligned} |PA - QA| &\leq |\mathbb{P}\{X \in A,\ X = Y\} - \mathbb{P}\{Y \in A,\ X = Y\}| \\ &\qquad + |\mathbb{P}\{X \in A,\ X \neq Y\} - \mathbb{P}\{Y \in A,\ X \neq Y\}| \\ &\leq 0 + \mathbb{P}\{X \neq Y\} \end{aligned}$$

Take the supremum over A to deduce that

<8> $$\|P - Q\|_{\text{TV}} \leq \inf\{\mathbb{P}\{X \neq Y\} : X \sim P, Y \sim Q\}.$$

The infimum runs over all probability spaces $(\Omega, \mathcal{F}, \mathbb{P})$, on which (X, Y) can be defined with the desired marginals, (P, Q).

Strictly speaking, we would need to assume that the diagonal $\Delta := \{(x, y) \in \mathcal{X}^2 : x = y\}$ is $\mathcal{A} \otimes \mathcal{A}$-measurable, to ensure that $\{\omega : X(\omega) \neq Y(\omega)\}$

belongs to $\mathcal{F}$. Under such an assumption, the inequality in <8> is an equality. Indeed, the assertions trivial if $\|P - Q\|_{\mathrm{TV}} = 0$ because then $P = Q$.

If $c := \|P - Q\|_{\mathrm{TV}} > 0$, let μ be the measure $P \wedge Q$ lifted up by the map $x \mapsto (x, x)$ to live on Δ. The nonnegative measure

$$\mathbb{P} := \mu + (P - Q)^+ \otimes (Q - P)^+ / c$$

has, by Example <5> and <6>, total mass

$$\|P \wedge Q\|_1 + \|(P - Q)^+\|_1 \|(Q - P)^+\|_1 / c = (1 - c) + c^2/c = 1.$$

That is, $\mathbb{P}$ is a probability measure on $\mathcal{A} \otimes \mathcal{A}$. It has marginals

$$(P - Q)^+ (P \wedge Q) = P \qquad \text{and} \qquad (Q - P)^+ (P \wedge Q) = Q$$

and mass $(P \wedge Q)(\mathcal{X}) = 1 - c$ on Δ. The coordinate maps X and Y have distributions P and Q and $\mathbb{P}\{X \neq Y\} = \mathbb{P}\Delta^c = \|P - Q\|_{\mathrm{TV}}$. □

3. Some examples of total variation distances

In a few cases it is possible to calculate the exact total variation distance between two measures.

<9> **Example.** Let P_1 denote the $N(\theta_1, I_n)$ multivariate normal distribution and P_2 denote the $N(\theta_2, I_n)$, with $\theta_1 \neq \theta_2$. Define $\tau := |\theta_1 - \theta_2|/2$ and $u = (\theta_2 - \theta_1)/|\theta_2 - \theta_1|$. Note that $u'(x - \theta_1)$ has a $N(0, 1)$ distribution under P_1, and a $N(2\tau, 1)$ distribution under P_2. The density of $P_1 - P_2$ with respect to Lebesgue measure is nonnegative in the halfspace $A_0 = \{x : u'(x - \theta_1) \leq \tau\}$. Thus

$$\begin{aligned} \|P_1 - P_2\|_1 &= 2(P_1 - P_2)A_0 \\ &= 2\mathbb{P}\{N(0, 1) \leq \tau\} - 2\mathbb{P}\{N(2\tau, 1) \leq \tau\} \\ &= 2\mathbb{P}\{|N(0, 1)| \leq \tau\} \\ &= \frac{4\tau}{\sqrt{2\pi}} + O(\tau^2) \qquad \text{as } \tau \to 0 \end{aligned}$$

When θ_1 is close to θ_2, so that $\tau = |\theta_1 - \theta_2|/2 \approx 0$, the total variation distance is approximately $\sqrt{2/\pi}|\theta_1 - \theta_2|$.

The rate of convergence in this Example is typical. Consider for example a family of probability measures $\{P_\theta : \theta \in \mathbb{R}^k\}$ with densities $\{f_\theta\}$ with respect to a measure λ. Supose the family of densities is ***differentiable in*** $\mathcal{L}^1(\lambda)$ ***norm*** at θ. That is, suppose there is an integrable function $\dot{f}_\theta$ for which

$$\lambda |f_{\theta+t} - f_\theta - t' \dot{f}_\theta| = o(|t|) \qquad \text{as } t \to 0$$

Then, writing u for the unit vector $t/|t|$, we have

$$\frac{\|P_{\theta+t} - P_\theta\|_1}{|t|} = \frac{\lambda |f_{\theta+t} - f_\theta|}{|t|} = \lambda |u' \dot{f}_\theta| + o(1)$$

For the $N(\theta, I_n)$ densities, $\phi(x - \theta)$, the pointwise derivative $(x - \theta)\phi(x - \theta)$ is also the derivative in $\mathcal{L}^1$ norm, which gives

$$\begin{aligned} \frac{\|N(\theta + t, I_n) - N(\theta, I_n)\|_1}{|t|} &= o(1) + \int |u'(x - \theta)\phi(x - \theta)| \\ &= o(1) + \mathbb{P}|u' N(0, I_n)| \\ &= o(1) + \frac{2}{\sqrt{\pi}} \end{aligned}$$

because $u' N(0, I_n)$ is $N(0, 1)$ distributed, and $\mathbb{P}|N(0, 1)| = \sqrt{2/\pi}$. □

Exact calculation of total variation distances in closed form can be difficult, if not impossible. Bounds and approximations often suffice. The next Section compares two methods by means of an application to a well known approximation.

<10> **Example.** Let P_n denote the $\text{Bin}(n,\theta)$ distribution and Q_n denote the Poisson$(n\theta)$ distribution, for a fixed $\theta > 0$. Example <8> gives a bound on the distance $\|P_n - Q_n\|_{\text{TV}}$. For the simple case with $n = 1$, the measure $P_1 \wedge Q_1$ puts mass only on the points 0 and 1:

$$\begin{aligned} \min(1-\theta, e^{-\theta}) &= 1-\theta \qquad \text{at } 0 \\ \min(\theta, \theta e^{-\theta}) &= \theta e^{-\theta} \qquad \text{at } 1 \end{aligned}$$

Thus

$$\|P_1 - Q_1\|_{\text{TV}} = 1 - (P_1 \wedge Q_1)(\mathbb{R}) = \theta(1-e^{-\theta}) \le \theta^2$$

Construct, on the same probability space, independent pairs (X_i, Y_i) with $X_i \sim P_1$ and $Y_i \sim Q_1$ and $\mathbb{P}\{X_i \ne Y_1\} \le \theta^2$ for $i = 1, 2, \ldots, n$. Then $X := \sum_i X_i$ is distributed as P_n and $Y_n := \sum_i$ is distributed as Q_n. By Example <8>,

$$\|P_n - Q_n\|_{\text{TV}} \le \mathbb{P}\{X \ne Y\} \ne \sum\nolimits_i \mathbb{P}\{X_i \ne Y_i\} \le n\theta^2.$$

This bound makes precise the idea that Q_n is a reasonable approximation to P_n if $n\theta$ is not too big and θ is small. □

The bound from the previous Example can be strengthened for $n\theta \ge 1$ by a more direct calculation.

<11> **Example.** For a fixed n, abbreviate the P_n and Q_n from the previous Example to P and Q. A more delicate argument will show that $\|P - Q\|_{\text{TV}} \le \theta$.

Define

$$\begin{aligned} b(k) &:= P\{k\} = \binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad \text{for } k = 0, 1, \ldots, n \\ p(k) &:= Q_k\{k\} = e^{-n\theta}(n\theta)^k/k! \qquad \text{for } k = 0, 1, \ldots \end{aligned}$$

To avoid trivial cases, assume $0 < \theta < 1$. For $k = 0, 1, \ldots, n$ define

$$g(k) := p(k)/b(k) = e^{-n\theta}(1-\theta)^{x-n}\prod_{i=1}^{x-1}\left(1-\frac{i}{n}\right)^{-1}$$

Then

$$\|P - Q\|_{\text{TV}} = \sum\nolimits_{k=0}^{n} (p(k) - q(k))^+ = 2P\,(1 - g(k))^+$$

It suffices to show that $g(k) \ge 1 - (k/n)$, for then the last expected value is bounded by $P(k/n) = \theta$.

The lower bound for $g(k)$ is trivial when $k = n$. For other values of k note that

$$\log\frac{g(k)}{1-(k/n)} = -n\theta - (n-k)\log(1-\theta) - \sum\nolimits_{i=1}^{k} h'(i/n)$$

where $h'(t) := -\log(1-t)$ and

$$h(t) := t + (1-t)\log(1-t) = \sum\nolimits_{k=2}^{\infty}\frac{t^k}{k(k-1)} \qquad \text{for } 0 \le t \le 1.$$

Note that h is convex and h' is increasing. For each integer k with $1 \le k \le n-1$,

<12> $$h\,(k/n) - h(0) = \int_0^{k/n} h'(t)\,dt = \sum\nolimits_{i=1}^{k}\int_{(i-1)/n}^{i/n} h'(t)\,dt \le \frac{1}{n}\sum_{i=1}^{k} h'\,(i/n)\,.$$

The inequality is also valid when $k = 0$, for then both sides equal zero.

REMARK. The inequality is fairly sharp, because

$$\frac{1}{n}\sum_{i=1}^{k} h'(i/n) \le \sum_{i=1}^{k} \int_{i/n}^{(i+1)/n} h'(t)\,dt = h((k+1)/n) - h(1/n) \qquad <13>$$

From <12>,

$$\log \frac{g(k)}{1-(k/n)} \ge -n\theta - (n-k)\log(1-\theta) + nh(k/n) \qquad \text{for } k = 0, 1, \ldots, n-1.$$

Replace k in the right-hand side by a continuous variable x in $[0, n]$. The lower bound becomes a convex function of x that achieves its minimum when $h'(x/n) = \log(1-\theta)$, that is, when $x = n\theta$. The lower bound is everywhere greater than than

$$-n\theta - n(1-\theta)\log(1-\theta) + nh(\theta) = -nh(\theta) + nh(\theta) = 0.$$

□ It follows that $g(k) \ge 1 - (k/n)$, as asserted.

The bound from Example <11> can be refined further (Le Cam 1965) by exploitation of properties of the ratio

$$R(k+1) := \frac{g(k+1)}{g(k)} = \frac{(1-\theta)n}{n-k}$$

The ratio decreases as k increases from 0 to n, with $R(k+1) \ge 1$ if and only if $k \ge n\theta$. Thus $g(k)$ achieves its minimum value at k_0, the smallest integer for which $k_0 \ge n\theta$. Not coincidentally, both $b(k)$ and $p(k)$ achieve their maxima near k_0. Indeed,

$$\frac{b(k+1)}{b(k)} = \frac{(n-k)\theta}{(k+1)(1-\theta)} \ge 1 \qquad \text{if and only if } \theta \ge \frac{k+1}{n+1},$$

and

$$\frac{p(k+1)}{p(k)} = \frac{n\theta}{k+1} \ge 1 \qquad \text{if and only if } \theta \ge \frac{k+1}{n},$$

Let me ignore these small differences by assuming that $k_0 = n\theta$ and that both $b(k)$ and $p(k)$ are maximized at k_0. By Stirling's formula,

$$b(k_0) \approx \frac{n^{n+1/2}\theta^{k_0}(1-\theta)^{n-k_0}}{\sqrt{2\pi}\,k_0^{k_0+1/2}(n-k_0)^{n-k_0+1/2}} \approx \left(2\pi k_0(1-\theta)\right)^{-1/2}$$

and

$$p(k_0) \approx \left(2\pi k_0\right)^{-1/2}$$

which gives

$$g(k) \ge g(k_0) \approx \sqrt{1-\theta} \qquad \text{for } k = 0, 1, \ldots, n.$$

REMARK. JAH explained to me that the $\sqrt{1-\theta}$ comes from the fact that the normal approximation to P has standard deviation $\sqrt{n\theta(1-\theta)}$ whereas the normal approximation to Q has standard deviation $\sqrt{n\theta}$. It would be worthwhile to compare with $\|N(0,\sigma_1^2) - N(0,\sigma_2^2)\|$. Presumably this distance behaves like $1 - (\sigma_2/\sigma_1)$ if $\sigma_1 > \sigma_2$.

The set $A_0 := \{k : b(k) \ge p(k)\}$ is of the form $\{k : k_1 \le k \le k_2\}$. Of course $k_0 \in A$. Treating the approximations as equalities, we have

$$\begin{aligned}
\|P - Q\|_{\mathrm{TV}} &= \sum_{k\in A}\left(\gamma b(k) + b(k)\sqrt{1-\theta} - p(k)\right) \qquad \text{where } \gamma := 1 - \sqrt{1-\theta} \\
&\le \gamma \sum_{k\in A} b(k) \qquad \text{because } p(k) \ge b(k)\sqrt{1-\theta} \text{ on } A \\
&= \gamma\, PA
\end{aligned}$$

Thus, $\|P - Q\|_{\text{TV}}$ should be (approximately) smaller than

$$1 - \sqrt{1-\theta} = \tfrac{1}{2}\theta + O(\theta^2) \qquad \text{near } \theta = 0.$$

With a more careful accounting of errors, Le Cam (1965, page 187) got a bound 2θ for the total variation distance. He noted that more elaborate calculations by Prohorov (1961) gave an even better result,

$$\|P - Q\|_{\text{TV}} = \theta\left(\lambda_0 + O(1 \wedge (n\theta)^{-1/2}\right) \qquad \text{with } \lambda_0 = 1/\sqrt{2e\pi} \approx 0.242.$$

Check

Actually, this is the form reported by Barbour, Holst & Janson (1992, page 2), who corrected a minor error in the original.

4. Total variation and minimax rates of convergence

The basic idea relating performance of an estimator to total variation distance is due to Le Cam (1973). Suppose $\widehat{\theta}$ is an estimator of a parameter θ in a metric space such that

<14> $$\mathbb{P}_\theta\{d(\widehat{\theta}, \theta) \geq \delta\} \leq \epsilon \qquad \text{for all } \theta.$$

In particular, suppose such an inequality holds for two values θ_0 and θ_1 for which $d(\theta_0, \theta_1) \geq 2\delta$. If we define $A_0 := \{\omega : d(\widehat{\theta}(\omega), \theta_0) < \delta\}$ then

$$\mathbb{P}_{\theta_0} A_0 \geq 1 - \epsilon \qquad \text{and} \qquad \mathbb{P}_{\theta_1} A_0 \leq \epsilon$$

from which it follows that

$$\|\mathbb{P}_{\theta_0} - \mathbb{P}_{\theta_1}\|_{\text{TV}} \geq \mathbb{P}_{\theta_0} A_0 - \mathbb{P}_{\theta_1} A_0 \geq 1 - 2\epsilon.$$

Conversely, if we wish to show that assertion <14> cannot be true, we can try to frind a pair θ_0 and θ_1 for which $d(\theta_0, \theta_1) \geq 2\delta$ but $\|\mathbb{P}_{\theta_0} - \mathbb{P}_{\theta_1}\|_{\text{TV}} < 1 - 2\epsilon$.

<15> **Example.** Consider the problem of estimation of a parameter θ, based on a sample $\{X_1, \ldots, X_n\}$ of size n from the Uniform$[0, \theta]$ distribution. More formally, let $\mathbb{P}_{n,\theta}$ be the uniform distribution on the cube $[0, \theta]^n$ and the $\{X_i\}$ be the coordinate maps on $\mathbb{R}^n$.

The estimator $M_n := \max\{X_1, \ldots, X_n\}$ lies within $O_p(1/n)$ of θ:

$$\begin{aligned}\mathbb{P}_{n,\theta}\{|M_n - \theta| \geq C/n\} &= \mathbb{P}_{n,\theta}\{X_i \leq \theta - C/n \quad \text{for } i = 1, \ldots, n\} \\ &= (1 - C/n\theta)^n \\ &\to \exp(-C/\theta) \qquad \text{as } n \to \infty.\end{aligned}$$

More precisely, for each $\theta_0 > 0$, there exists a constant $C_0 = C_0(\epsilon, \theta_0)$ such that

<16> $$\mathbb{P}_{n,\theta}\{|M_n - \theta| \geq C_0/n\} \leq \epsilon \qquad \text{for all } n \text{ and all } \theta \in (0, \theta_0].$$

The $O_p(1/n)$ rate of convergence for $\{M_n\}$ is uniform over bounded subsets of the parameter space.

In general, an estimator sequence $\{\widehat{\theta}_n\}$ for which a rate of convergence is required to hold uniformly over even a small neighborhood of an particular θ_0 cannot do better than $O_p(1/n)$. For simplicity, suppose $\theta_0 = 1$. Consider $\theta_n = 1 - 2C_\epsilon/n$, for some (suitably large) positive C_ϵ. Calculate total variation distance using densities with respect to Lebesgue measure on $\mathcal{B}(\mathbb{R}^n)$.

$$\begin{aligned}\|\mathbb{P}_{n,\theta_n} - \mathbb{P}_{n,\theta_0}\| &= \int \left(\theta_n^{-n}\{\max_i x_i \leq \theta_n\} - \{\max_i x_i \leq 1\}\right)^+ d\mathbf{x} \\ &= (\theta_n^{-n} - 1)\int \{\max_i x_i \leq \theta_n\}\, d\mathbf{x} \\ &= 1 - (1 - 2C_\epsilon/n)^n \\ &\to 1 - \exp(-2C_\epsilon).\end{aligned}$$

If $C_\epsilon < \frac{1}{2}\log(1/2\epsilon)$ then $1-(1-2C_\epsilon/n)^n < 1-2\epsilon$ for $n \geq n_\epsilon$. For no estimator $\widehat{\theta}_n$ can we have

$$\sup_{n\geq n_\epsilon} \mathbb{P}_{n,\theta}\{|\widehat{\theta}_n - \theta| \geq C_\epsilon/n\} \leq \epsilon \qquad \text{for both } \theta = 1 \text{ and } \theta = 1 - 2C_\epsilon/n.$$

We certainly cannot find an estimator that converges at a rate faster than $O_p(1/n)$ uniformly over θ near 1. □

Traditional theoretical statistics is much concerned with estimation based on independent observations. By calculating bounds on the total variation distance between two product measures, we can deduce lower bounds on the local minimax rates of convergence of estimators. For these calculations, it is sometimes easier to work with distances for which the calculation reduces to calculation of distances between ("one-dimensional") marginals. The Hellinger and Kullback-Leibler distances, the key properties of which will be summarized in the next Sections, have this desirable property. However, as shown in Section 6, there are also tractable bounds based on calculations with $\mathcal{L}^2$ distances.

The simple argument from the start of the section will be generalized in Chapter 18 to show that total variation distance between convex hulls of sets of measures provide better lower bounds for minimax rates. Unfortunately, even with independent observations, the taking of convex hulls destroys much of the convenience of working with Hellinger or Kullback-Leibler distances. We will need to explore other methods for controlling total variation distance. The method from Section 6 has a simple generalization to handle convex hulls.

5. Helinger and Kullback-Leibler distances

Let P and Q be probability measures with densities p and q with respect to a dominating measure λ. The square roots of the densities, $\sqrt{p}$ and $\sqrt{q}$ are both square integrable; they both belong to $\mathcal{L}^2(\lambda)$. The Hellinger distance between the two measures is defined as the $\mathcal{L}^2$ distance between the square roots of their densities,

$$\begin{aligned} H(P,Q)^2 &= \lambda(\sqrt{p}-\sqrt{q})^2 \\ &= \lambda(p+q-2\sqrt{pq}) \\ &= 2 - 2\lambda\sqrt{pq}. \end{aligned} \tag{17}$$

It is easy to show that the integral defining the Hellinger distance does not depend on the choice of dominating measure (Problem [2]). The quantity $\lambda\sqrt{pq}$ is called the ***Hellinger affinity*** between the two measures. It is denoted by $\alpha_2(P,Q)$.

The Hellinger distance satisfies the inequality $0 \leq H(P,Q) \leq \sqrt{2}$. Some authors prefer to have an upper bound of 1; they include an extra factor of a half in the definition of $H(P,Q)^2$. The equality at 0 occurs when $\sqrt{p} = \sqrt{q}$ almost surely mod[λ], that is, when $P = Q$ as measures on $\mathcal{A}$. Equality at $\sqrt{2}$ occurs when the Hellinger affinity is zero, that is, when $pq = 0$ almost surely mod[λ], which is the condition that P and Q be supported by disjoint subsets of $\mathcal{X}$. For example, discrete distributions (concentrated on a countable set) are always at the maximum Hellinger distance from nonatomic distributions (zero mass at each point).

The ***Kullback-Leibler "distance"*** (also known as the relative entropy) between P and Q is defined as $D(P\|Q) = \lambda(p\log(p/q))$. Again $D(P\|Q)$ does not depend on the choice of dominating measure. In fact $D(\cdot\|\cdot)$ is not a metric—for one thing, it is not symmetric—but it does have analogous

properties, which makes it a useful substitute for total variation distance. It is not hard to show, via Jensen's inequality, that $D(P\|Q)$ is always nonnegative, achieving the value zero if and only if $P = Q$. If $\lambda\{p > 0 = q\} > 0$, which happens when P is not dominated by Q, then $D(P\|Q) = \infty$. When P is dominated by Q we can choose λ equal to Q, giving $D(P\|Q) = Q(p_0 \log p_0)$, where $p_0 := dP/dQ$.

The following properties of Hellinger distance are directly relevant to calculation of minimax rates. See Chapter 5 for a more systematic treatment of Hellinger distance and the related concept of Hellinger differentiability.

(i) $\frac{1}{2}\|P - Q\|_1 \le H(P, Q) \le \|P - Q\|_1^{1/2}$. See Problem [5].
(ii) $H(P_1 \otimes \ldots \otimes P_n, Q_1 \otimes \ldots \otimes Q_n)^2 \le \sum_{i=1}^n H(P_i, Q_i)^2$. See Problem [6].
(iii) $H^2(P, Q) \le D(P\|Q)$. See Problem [7]

The following properties of Kullback-Leibler distance are directly relevant to calculation of minimax rates. See Chapter 4 for a more systematic treatment of Kullback-Leibler distance.

(iv) $\frac{1}{2}\|P - Q\|_1^2 \le D(P\|Q)$. See Problem [8].
(v) $D(P_1 \otimes \ldots \otimes P_n\|Q_1 \otimes \ldots \otimes Q_n) = \sum_{i \le n} D(P_i\|Q_i)$. See Problem [9]

If P and $\{P_n\}$ are probability measures then inequality (i) implies

$$\|P_n - P\|_1 \to 0 \qquad \text{if and only if} \qquad H(P_n, P) \to 0.$$

Inequality (iv) implies that

$$\|P_n - P\|_1 \to 0 \qquad \text{if} \qquad \min\big(D(P_n\|P), D(P\|P_n)\big) \to 0,$$

but the implication in the opposite direction does not hold.

6. Second-moment bounds on total variation distance

Particularly for probability measures P and Q that are close, we often need only upper bounds on total variation distance. If both measures are dominated by a probability measure λ, with densities p and q, then

<18> $$\|P - Q\|_1^2 \le \lambda(p - q)^2$$

Notice that the right-hand side depends on the choice of λ, whereas the left-hand side does not. Often it will be convenient to choose $\lambda = \mathbb{P}$ or $\lambda = (\mathbb{P} + \mathbb{Q})/2$.

The second-moment upper bound is often of the correct order of magnitude for a well chosen λ. For example, suppose $dQ/dP = 1 + \Delta$, with Δ small enough to justify integration of the expansion

$$(1 + \Delta)^{1/2} = 1 + \tfrac{1}{2}\Delta - \tfrac{1}{4}\Delta^2 + \ldots$$

to give $P(1 + \Delta)^{1/2} \approx 1 - P\Delta^2/4$. (Of course $P\Delta = 0$ because $1 = Q1 = P1 + P\Delta$.) Then

$$H^2(P, Q) = 2 - 2P(1 + \Delta)^{1/2} \approx \tfrac{1}{2}P\Delta^2$$

More precisely, if there exists a constant C such that $\sqrt{p} + \sqrt{q} \le C$ everywhere then

$$H^2(P, Q) = \lambda \frac{|p - q|^2}{|\sqrt{p} + \sqrt{q}|^2} \ge \frac{\lambda|p - q|^2}{C^2}$$

and if there exists a constant c such that $\sqrt{p}+\sqrt{q} \geq c$ everywhere then

$$H^2(P, Q) \leq \frac{\lambda|p-q|^2}{c^2}$$

In particular, if $\lambda = (P+Q)/2$ then $p+q=2$, so that $\sqrt{2} \leq \sqrt{p}+\sqrt{q} \leq 2\sqrt{2}$ and

<19> $$\frac{\lambda|p-q|^2}{4} \leq H^2(P, Q) \leq \frac{\lambda|p-q|^2}{2} \qquad \text{if } \lambda = \frac{P+Q}{2}$$

See Problem [10] for a comparison between P and $(P+Q)/2$ as dominating measures.

<20> **Example.** As shown in Example <9>, and the explanation that follows that Example, the total variation distance between the $N(\theta, 1)$ and the $N(0, 1)$ distributions decreases like $\sqrt{2/\pi}|\theta|$ as $\theta \to 0$. More precisely,

$$\phi(x-\theta) = \phi(x) + \theta x \phi(x) + \tfrac{1}{2}\theta^2(x^2-1)\phi(x) + \ldots$$

so that

$$\int |\phi(x-\theta) - \phi(x)|\, dx = |\theta| \int |x|\phi(x)\, dx + O(\theta^2)$$

A similar argument suggests that the mixture $P_\theta = \frac{1}{2}N(\theta, 1) + \frac{1}{2}N(-\theta, 1)$ converges to the $N(0, 1)$ at an even faster rate:

$$\tfrac{1}{2}\left(\phi(x-\theta) + \phi(x+\theta)\right) = \phi(x) + \tfrac{1}{2}\theta^2(x^2-1)\phi(x) + \ldots$$

so that

$$\int \left|\tfrac{1}{2}\phi(x-\theta) + \tfrac{1}{2}\phi(x+\theta) - \phi(x)\right| dx = \tfrac{1}{2}\theta^2 \int |x^2-1|\phi(x)\, dx + O(\theta^4)$$

Integration by parts gives $\frac{1}{2}\int |x^2-1|\phi(x)\, dx = 2\phi(1) \approx 0.48$.

It is not too difficult to make these calculations rigorous. The second moment bound gives the same rate of convergence even more easily.

$$\begin{aligned}
\|P_\theta - P_0\|_1^2 \leq P_0\left|\frac{dP_\theta}{dP_0} - 1\right|^2 &= P_0\left|\frac{dP_\theta}{dP_0}\right|^2 - 1 \\
&= \tfrac{1}{4}P_0\left|\exp\left(\theta x - \frac{\theta^2}{2}\right) + \exp\left(-\theta x - \frac{\theta^2}{2}\right)\right|^2 - 1 \\
&= \tfrac{1}{2}\left(\exp(\theta^2) + \exp(-\theta^2)\right) - 1 \\
&= \frac{\theta^4}{2!} + \frac{\theta^8}{4!} + \ldots
\end{aligned}$$

The bound on the distance $\|P_\theta - P_0\|_1$ decreases like $\theta^2/\sqrt{2}$, an overestimate by a constant factor of approximately 1.5. □

Often the second moment method reduces calculations of bounds on total variation distances to calculations of variances and covariances.

<21> **Lemma.** *If $\mathbb{P} = \prod_{i\leq n} P_i$ and $\mathbb{Q} = \prod_{i\leq n} Q_i$ are finite products of probability measures such that Q_i has density $1+\Delta_i(x_i)$ with respect to P_i, then*

$$\|\mathbb{P} - \mathbb{Q}\|_1^2 \leq \prod_{i\leq n}\left(1 + P_i\Delta_i^2\right) - 1 \leq \exp\left(\sum_i P_i\Delta_i^2\right) - 1.$$

Proof. We may assume each $P_i\Delta_i^2$ finite, for otherwise the asserted inequality is trivial.

Thanks to JAH for neater proof.

Note that $L := d\mathbb{Q}/d\mathbb{P} = \prod_i (1 + \Delta_i(x_i))$. Also $\mathbb{P}\Delta_i = P_i\Delta_i = 0$ because both P_i and Q_i are probabilities. Thus $\mathbb{P}L = 1$. From <18>,

$$\begin{aligned}\|\mathbb{P} - \mathbb{Q}\|_1^2 &\le \mathbb{P}(L-1)^2 \\ &= \mathbb{P}L^2 - 1 \\ &= \prod_i P_i(1 + 2\Delta_i + \Delta_i^2) - 1 \\ &= \prod_i (1 + 0 + P_i\Delta_i^2) - 1,\end{aligned}$$

□ as asserted.

The upper bound in the Lemma decreases like $\sum_i P_i\Delta_i^2$ when the sum is small. In situations where $P_i\Delta_i^2$ behaves like $H^2(P_i, Q_i)^2$, the second moment bound is comparable to the the analogous bound for Hellinger distance:

$$H^2(\mathbb{P}, \mathbb{Q}) \le \sum_{i=1}^n H(P_i, Q_i)^2$$

As you will see in Chapter 18, the second moment method also works for situations where it becomes exceedingly difficult to calculate Hellinger distances directly.

7. Pointwise estimation of densities

The literature on density estimation is filled with assertions like "If a density f has an eighth order derivative satisfying a Lipschitz condition near a point x_0 then the optimal rate of convergence for estimators of $f(x_0)$ is $O(n^{-\beta})$", where β turns out to be some weird fractional power that depends on the smoothness assumptions (and on the dimension of the underlying Euclidean space in multidimensionalresults). Where do such strange results come from? How does one know if an unknown density really does satisfy all the smoothness assumptions? What happens if the assumptions are violated?

Apart from its obvious virtues as a means of keeping mathematical statisticians employed, density estimation is worthy of study as an illustration of the ideas behind the minimax calculations from the previous Chapter. Many fancy smoothing ideas, particularly those that chew up immense numbers of computing cycles, have their roots in the density estimation literature.

Most of the ideas will be plain enough for the one-dimensional case, with the origin as the point at which the density is to be estimated. Typically density estimation requires smoothness assumptions about the underlying f in a neighborhood of a point, which lets us borrow data from a neighborhood of 0 in order to make inferences about $f(0)$. As a heuristic, smoothness lets us pretend that f is like a polynomial of fixed degree near 0. The minimax bounds will emerge when we discover the most untrustworthy density with a given degree of smoothness, as quantified by formal definition of the following type.

<22> **Definition.** *Let s and δ be strictly positive. Let k be the largest integer strictly less than s and let $s = k + \alpha$, with $0 < \alpha \le 1$. For a function f defined and k times differentiable at least on the interval $(-\delta, \delta)$, define the norm $\|f\|_{s,\delta}$ as the smallest constant C (possibly infinite) for which*

$$\sup_{|t|<\delta} |f^{(i)}(t)| \le C \qquad \text{for } i = 0, 1, \ldots, k$$

$$|f^{(k)}(t_1) - f^{(k)}(t_2)| \le C|t_1 - t_2|^\alpha \qquad \text{for } |t_1|, |t_2| < \delta.$$

Call f locally s-smooth if $\|f\|_{s,\delta}$ is finite for some positive δ. Write $\mathfrak{S}_s(0, \delta, C)$ for the class of such functions with $\|f\|_{s,\delta} \le C$.

Notice that finiteness of $\|f\|_{1,\delta}$ requires only a Lipschitz condition on f near the origin.

Let f_0 denote the uniform density on $[-\frac{1}{2}, \frac{1}{2}]$. For a fixed small $\delta > 0$ and a constant $C > 1$, let us work with a smoothness class $\mathfrak{S}_s = \mathfrak{S}_s(0, \delta, C)$ for the remainder of the Section. Of course $f_0 \in \mathfrak{S}_s$. Untrustworthy densities near f_0 can be manufactured in a systematic fashion, by means of small perturbations. Let $\mathfrak{h}$ be an infinitely differentiable function with the following properties.

Hall ISI 57, 1989

(i) $\mathfrak{h}(x) = 0$ for $x \notin [-1, +1]$

(ii) $\mathfrak{h}(0) \neq 0$

(iii) $\int_{-1}^{+1} \mathfrak{h}(x)\, dx = 0$.

For example, one could splice together several rescaled and translated pieces of a function like

$$\exp(-1/x^2)\left(1 - \exp(-1/(1-x)^2)\right).$$

Construct a small perturbation for small positive ϵ by

$$f_\epsilon(x) = f_0(x) + \epsilon^s \mathfrak{h}(x/\epsilon).$$

The ith derivative is bounded by $\epsilon^{s-i} \sup_x |\mathfrak{h}^{(i)}(x)|$, which satisfies the requisite assumptions for $\mathfrak{S}_s$ if the constant C is large enough. Notice that it is only the Lipshitz condition on the kth derivative that might cause any trouble; for the lower-order derivatives there is a positive power of a small ϵ to keep the function small.

More detail needed?

The choice of the uniform distribution for $\mathbb{P}$ makes calculation of the second moment quantities child's play.

$$\mathbb{P}\left|\frac{f_\epsilon}{f} - 1\right|^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} \epsilon^{2s} \mathfrak{h}(x/\epsilon)^2 dx \le \epsilon^{2s+1} \int_{-\infty}^{\infty} \mathfrak{h}(y)^2 dy.$$

That is,

Argue that ϵ equal to a small multiple of $n^{-1/(1+2s)}$ gives models that are not well separated in the total variation sense, which leads to $n^{-s/(1+2s)}$ as the minimax rate for estimation of $f(0)$.

8. Problems

[1] Suppose P_1 and P_2 are probability measures with densities p_1 and p_2 with respect to a dominating measures λ. Let ν be another dominating measure. Write ℓ for the density of λ with respect to $\lambda + \nu$.

(i) Show that P_i has density $p_i \ell$ with respect to $\lambda + \nu$.

(ii) Show that $(\lambda + \nu)(\sqrt{p_1 \ell} - \sqrt{p_2 \ell})^2 = \lambda(\sqrt{p_1} - \sqrt{p_2})^2$.

(iii) Deduce that the integral that defines the Hellinger distance $H(P_1, P_2)$ does not depend on the choice of dominating measure.

[2] Let P and Q be probability measures with densities p and q with respect to a sigma-finite measure λ. For fixed $\alpha \ge 1$, show that $\Delta_\alpha(P, Q) := \lambda |p^{1/\alpha} - q^{1/\alpha}|^\alpha$ does not depend on the choice of dominating measure. Hint: Let μ be another sigma-finite dominating measure. Write ψ for the density of λ with respect to $\lambda + \mu$. Show that $dP/d(\lambda + \mu) = \psi p$ and $dQ/d(\lambda + \mu) = \psi q$. Express $\Delta_\alpha(P, Q)$ as an integral with respect to $\lambda + \mu$. Argue similarly for μ.

[3] Adapt the argument from the previous Problem to show that the relative entropy $D(P\|Q)$ does not depend on the choice of dominating measure.

[4] Let P be the standard Cauchy distribution on the real line, and let Q be the standard normal distribution. Show that $D(P\|Q) = \infty$, even though P and Q are mutually absolutely continuous.

[5] For probability measures P and Q with densities p and q with respect to λ, show that

$$\|P - Q\|_1 \le H(P, Q)\sqrt{4 - H(P, Q)^2} \le 2H(P, Q) \le 2\|P - Q\|_1^{1/2}$$

Hint: Show that

$$(\sqrt{p} - \sqrt{q})^2 = |\sqrt{p} - \sqrt{q}|\ \left(\sqrt{p} + \sqrt{q}\right) = |p - q|.$$

Invoke the Cauchy-Schwarz inequality when integrating the middle term.

[6] For probability measures $\{P_i\}$ and $\{Q_i\}$, show that

$$H(P_1 \otimes \ldots \otimes P_n, Q_1 \otimes \ldots \otimes Q_n)^2 = 2 - 2\prod_{i\le n}\left(1 - \tfrac{1}{2}H(P_i, Q_i)^2\right)$$
$$\le \sum_{i=1}^{n} H(P_i, Q_i)^2$$

Hint: Write y_i for $H(P_i, Q_i)^2/2$. For the final inequality you need to show that the function

$$G_n(y_1, \ldots, y_n) := \textstyle\sum_{i=1}^n y_i + \prod_{i=1}^n (1 - y_i) - 1$$

is nonnegative for all $0 \le y_i \le 1$. Check the case $n = 1$ directly. The lower bound of 0 is achieved when $n = 1$. For fixed $y_1, \ldots, y_{n-1}$, show that G_n achieves its minimum at either $y_n = 0$ or $y_n = 1$. Also show that

$$\min\left(G_n(y_1, \ldots, y_{n-1}, 0), G_n(y_1, \ldots, y_{n-1}, 1)\right) \ge G_{n-1}(y_1, \ldots, y_{n-1})$$

[7] For probabilities P and Q on the same space, show that $D(P\|Q) \ge H^2(\mathbb{P}, \mathbb{Q})$, by the following steps.

(i) Dispose of the case where P is not dominated by Q.

(ii) Define $\eta := \sqrt{p} - 1$ where $p = dP/dQ$. Show that $Q\eta^2 = H^2(P, Q)$ and $2Q\eta = -H^2(P, Q)$. Hint: Consider $Q(1 + \eta)^2$.

(iii) Show that

$$D(P\|Q) = 2Q\left((1 + \eta)^2 \log(1 + \eta)\right) \ge 2Q\left((1 + \eta)^2 \frac{\eta}{1 + \eta})\right).$$

[8] (Due to Csiszar (1967), Kullback (1967), and Kemperman (1969). Also Pinsker?). For probabilities P and Q on the same space, show that $D(P\|Q) \ge \frac{1}{2}\|P - Q\|_1^2$ by these steps. Remember that

Check citation

$$\psi(x) := \frac{(1 + x)\log(1 + x) - x}{x^2/2} \ge (1 + x/3)^{-1} \qquad \text{for } x \ge -1.$$

(i) Dispose of the case where P is not dominated by Q.

(ii) Define $\delta := p - 1$ where $p = dP/dQ$. Show that $Q\delta = 0$ and

$$D(P\|Q) = \tfrac{1}{2}Q\left(\delta^2\psi(\delta)\right) \ge \tfrac{1}{2}Q\left(\frac{\delta^2}{1 + \delta/3}\right)Q(1 + \delta/3).$$

(iii) Invoke the Cauchy-Schwarz inequality to bound the last product from below by $\frac{1}{2}\left(Q|\delta|\right)^2$.

[9] Show that $D(P_1 \otimes \ldots \otimes P_n \| Q_1 \otimes \ldots \otimes Q_n) = \sum_{i \le n} D(P_i \| Q_i)$. Hint: Suppose $dP_i/dQ_i = p_i$. Consider $\log\left(\prod_i p_i\right)$.

[10] Suppose a probability measure $\mathbb{Q}$ has density $1+\Delta$ with respect to a probability measure $\mathbb{P}$. Define $\mathbb{M} = (\mathbb{P}+\mathbb{Q})/2$. Write p and q for the densities of $\mathbb{P}$ and $\mathbb{Q}$ with respect to $\mathbb{M}$.

This problem needs checking.

(i) Show that $p = (1+\Delta/2)^{-1} = 2 - q$.

(ii) Deduce that $\mathbb{M}|p-q|^2 \le 2\mathbb{P}\Delta^2$. Hint: $\Delta \ge -1$.

(iii) If $\Delta/2$ is bounded above by a constant C, show that $\mathbb{M}|p-q|^2 \ge \mathbb{P}\Delta^2/(1+C)$.

9. Notes

I adapted the results on the total variation and relative entropy distances between Binomial and Poisson distributions from Reiss (1993, p 25). he credited Barbour & Hall (1984) with the first result, and Falk & Reiss (1992) with the second result.

Check Barbour and Hall

Barbour et al. (1992) have devoted a whole book to the topic of Poisson approximation.

The idea for Section 7 is adapted from Hall (1989).

References

Barbour, A. D. & Hall, P. (1984), 'On the rate of Poisson convergence', *Proceedings of the Cambridge Philosophical Society* **95**, 473–480.

Barbour, A. D., Holst, L. & Janson, S. (1992), *Poisson Approximation*, Oxford University Press.

Csiszar, I. (1967), 'Information-type measures of difference of probability distributions and indirect observations', *Studia Scientarium Mathematicarum Hungarica* **2**, 299–318.

Dunford, N. & Schwartz, J. T. (1958), *Linear Operators, Part I: General Theory*, Wiley.

Falk, M. & Reiss, R.-D. (1992), 'Poisson approximation of empirical processes', *Statist. Prob. Letters* **14**, 39–48.

Hall, P. (1989), 'On convergence rates in non-parametric problems', *International Statistical Review* **57**, 45–58.

Kemperman, J. H. B. (1969), On the optimum rate of transmitting information, *in* 'Probability and Information Theory', Springer-Verlag. Lecture Notes in Mathematics, 89, pages 126–169.

Kullback, S. (1967), 'A lower bound for discrimination information in terms of variation', *IEEE Transactions on Information Theory* **13**, 126–127.

Le Cam, L. (1965), On the distribution of sums of independent random variables, *in* J. Neyman & L. L. Cam, eds, 'Bernouilli, Bayes, Laplace', Springer-Verlag, New York, pp. 179–202. Proceedings of a research seminar, UC Berkley 1963.

Le Cam, L. (1973), 'Convergence of estimates under dimensionality restrictions', *Annals of Statistics* **1**, 38–53.

Prohorov, Yu. V. (1961), 'Asymptotic behavior of the binomial distribution', *Selected Translations in Mathematical Statistics and Probability* **1**, 87–95.

Reiss, R.-D. (1993), *A Course on Point Processes*, Springer-Verlag.