Statistics 612: L_p spaces, metrics on spaces of probabilites, and connections to estimation

Moulinath Banerjee

March 17, 2009

1 L_p spaces and Hilbert spaces

We first formally define L_p spaces. Consider a measure space $(\mathcal{X}, \mathcal{A}, \mu)$, where μ is a (σ-finite) measure. Let $\mathcal{F}$ be the set of all real–valued measurable functions defined on $\mathcal{X}$. The space $L_p(\mu)$ comprises that subset of functions of $\mathcal{F}$ that have finite p'th moments (here $p \geq 1$). In other words:

$$L_p(\mu) = \{f \in \mathcal{F} : \int \mid f(x) \mid^p \, d\mu(x) < \infty\}\,.$$

When the measure μ is a probability P, we obtain the class of all random variables on the probability space, $(\mathcal{X}, \mathcal{A}, P)$ that have finite p'th moments. This is a normed linear space over $\mathbb{R}$ with the norm (length) of the vector f (this is called the L_p norm) being given by

$$\|f\|_p = \left(\int \mid f(x) \mid^p \, d\mu(x)\right)^{1/p}\,.$$

The above norm induces a metric d where $d(f, g) = \|f - g\|_p$. Note that $d(f, g) = 0$ if and only if $f = g$ a.e. μ, in which case we identify f with g. The L_p norm, like all worthy norms, satisfies the triangle inequality:

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p\,;$$

this is precisely Minkowski's inequality. For random variables X, Y defined on the same probability space and having finite p'th moments, Minkowski's inequality states:

$$E(\mid X + Y \mid^p)^{1/p} \leq E(\mid X \mid^p)^{1/p} + E(\mid Y \mid^p)^{1/p}\,.$$

Minkowski's inequality is a consequence of Hölder's inequality which states that for measurable real-valued functions f, g defined on $\mathcal{X}$, we have:

$$\mid \int f(x)\, g(x)\, d\mu(x) \mid \leq \left(\int \mid f(x) \mid^p \, d\mu(x)\right)^{1/p} \left(\int \mid g(x) \mid^q \, d\mu(x)\right)^{1/q}\,.$$

The space $L_p(\mu)$ is a *Banach space* for $p \geq 1$ – this is a normed linear space that is complete – i.e. in which every Cauchy sequence has a limit.

The notion of continuity for real valued functions defined on $L_p(\mu)$ is a natural extension of the usual one for Euclidean spaces. A sequence of functions g_n converges to a function g in $L_p(\mu)$ if $\|g_n - g\|_p \to 0$. A real-valued function ψ defined on $L_p(\mu)$ is said to be continuous if $\psi(g_n)$ converges to $\psi(g)$ as a sequence of real numbers whenever g_n converges to g in $L_p(\mu)$.

Let's concretize to a specific example. Let $\mathcal{X}$ be the space of positive integers $\mathcal{N}$, $\mathcal{A}$ be the power set of $\mathcal{N}$ and μ be counting measure that assigns the cardinality of a subset of $\mathcal{N}$ as its measure. The space $L_p(\mu)$ is then the space of all real valued sequences $\{x_1, x_2, \ldots\}$ that are p'th power summable – i.e. $\sum_{i=1}^{\infty} \mid x_i \mid^p < \infty$. Clearly all sequences that have only finitely many non-zero entries satisfy this condition. This space is referred to as the l_p space. The l_p spaces are infinite-dimensional spaces – i.e. these spaces do not have a finite basis.

Another crucial inequality, which in particular, implies that the function $f \mapsto \|f\|_p$ is a continuous function from $L_p(\mu)$ to the reals is that

$$|\|h_1\|_p - \|h_2\|_p| \leq \|h_1 - h_2\|_p \,.$$

This is once again a consequence of the triangle inequality.

For $p = 2$, the space $L_p(\mu)$ has more geometric structure: it becomes a Hilbert space. We introduce Hilbert spaces in some generality. A Hilbert space $\mathcal{H}$ is a normed linear (vector) space (over the field $\mathbb{R}$ in the current discussion, though the general treatment requires the field to be the field of complex numbers) that is complete (with respect to the metric topology induced by the norm) and such that the norm arises from an *inner product*. The inner product $\langle x, y\rangle$ is a (bilinear) map from $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$ satisfying the following properties: (a) $\langle \alpha\, x + \beta\, y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$, (b) $\langle x, y\rangle = \langle y, x\rangle$, and (c) $\langle x, x\rangle = \|x\|^2$. Note that properties (a) and (b) jointly imply linearity in the second co-ordinate as well; i.e. $\langle z, \alpha x + \beta y\rangle = \alpha\, \langle z, x\rangle + \beta\langle z, y\rangle$. Hence bilinearity follows.

It is worthwhile here to recall the fundamental properties of a norm (on any normed linear space). These are (a) $\|x\| = 0$ if and only if $x = 0$, (b) $\|\alpha\, x\| = \mid \alpha \mid \|x\|$, for any scalar α, and (c) $\|x + y\| \leq \|x\| + \|y\|$, the triangle inequality.

The Hilbert spaces that we will be most concerned will be L_2 spaces, but this is not the simplest example of this species. The simplest Hilbert spaces over $\mathbb{R}$ are the Euclidean spaces $\mathbb{R}^k$ for any integer k. In $\mathbb{R}^k$, the inner product is $\langle x, y\rangle = \sum_{i=1}^{k} x_i\, y_i = x^T\, y$(if we write x and y as column vectors). Check that this inner product does induce the usual Euclidean norm. Completeness of $\mathbb{R}^k$ with respect to the usual Euclidean metric follows as a direct consequence of the fact that the real line is complete.

A fundamental relation in Hilbert spaces is the Cauchy-Schwarz inequality which states

that:

$$| \langle x, y \rangle | \leq \|x\| \, \|y\| \, .$$

Simple as this relation is, its ramifications are *profound*. How do we prove this? Assume without loss of generality that neither x nor y is 0 (otherwise, the inequality is trivial). Observe that $\|x - \alpha \, y\|^2$ is necessarily non-negative, for every real α. Expanding this in terms of the inner product, we find:

$$\|x\|^2 - 2 \, \alpha \, \langle x, y \rangle + \alpha^2 \, \|y\|^2 \geq 0 \, .$$

This function is strictly convex in α and is uniquely minimized at $\alpha_0 = \langle x, y \rangle / \|y\|^2$, as can be checked by straightforward differentiation. Plugging in α_0 for α must preserve the inequality above; doing so, and simplifying leads to the fact that

$$\langle x, y \rangle^2 \leq \|x\|^2 \, \|y\|^2 \, .$$

This implies the inequality.

The line of proof above admits a nice geometric intepretation that will prove useful in thinking about inner products. Draw vectors x, y on the ($\mathbb{R}^2$) plane (emanating from the origin), and for simplicity let them lie in the first quadrant (so that the angle θ between the two of them is an acute angle). For each α you can draw the vector $x - \alpha \, y$; it is then easy to see that the the length of this vector ($\|x - \alpha \, y\|$, with the norm here denoting the usual Euclidean norm) is minimized for that α_0 for which the vector $x - \alpha_0 \, y$ is perpendicular to the vector $\alpha_0 \, y$ (and $\alpha_0 \neq 0$, unless $x = y$). Now, by the proof above, α_0 should be $(x_1 \, x_2 + y_1 \, y_2)/(y_1^2 + y_2^2)$. Does this tally with what analytical geometry tells us? We can use the Pythagoras' theorem to verify. We have:

$$\alpha_0^2 (y_1^2 + y_2^2) + (x_1 - \alpha_0 \, y_1)^2 + (x_2 - \alpha_0 \, y_2)^2 = x_1^2 + x_2^2 \, ,$$

which on simplification gives:

$$-2 \, \alpha_0 \, (x_1 \, y_1 + x_2 \, y_2) + 2 \, \alpha_0^2 (y_1^2 + y_2^2) = 0 \, ,$$

and this yields:

$$\alpha_0 = \frac{x_1 \, y_1 + x_2 \, y_2}{y_1^2 + y_2^2} = \frac{\langle x, y \rangle}{\langle y, y \rangle^2} \, .$$

Now, check (from definitions) that the cosine of the angle between x and y is:

$$\cos \theta = \frac{\alpha_0 \, \|y\|}{\|x\|} = \frac{\langle x, y \rangle}{\|x\| \, \|y\|} \, .$$

But since $\cos \theta$ in absolute magnitude is less than or equal to 1, the Cauchy-Schwarz inequality must hold. The same arguments would work in 3 dimensions, but in any case, the following is clear: Geometrically, the Cauchy-Schwarz inequality is a restatement of the fact that the cosine of the angle between two vectors in a Hilbert space is no greater than 1 in absolute value. Of course, it is

difficult to visualize angles in a general Hilbert space. But taking the cue from 2 or 3 dimensions, one can *define* the angle between two vectors in a Hilbert space to be:

$$\alpha = \cos^{-1} \frac{\langle x, y \rangle}{\|x\| \, \|y\|} \, .$$

Orthogonality of two vectors then corresponds to α being either $\pi/2$ or $3\,\pi/2$ whence the numerator in the argument to $\cos^{-1}$ function should vanish: in other words, vectors x and y in a Hilbert space will be said to be orthogonal to each other (written as $x \perp y$), if $\langle x, y \rangle = 0$.

As a consequence of the Cauchy-Schwarz inequality, it follows that the inner product is a continuous function from $\mathcal{H} \times \mathcal{H} \to \mathbb{R}$; in other words, if $(x_n, y_n) \to (x, y)$ in $\mathcal{H} \times \mathcal{H}$ then $\langle x_n, y_n \rangle \to \langle x, y \rangle$.

We develop the concept of orthogonality in some detail as it plays an important role in probabilistic and statistical computations. A vector x is said to be orthogonal to a non-empty subset S of $\mathcal{H}$ if $x \perp y$ for every vector $y \in S$. The set of all vectors orthogonal to S is denoted by $S^\perp$. Check that the following hold: (i) $\{0\}^\perp = \mathcal{H}$, $\mathcal{H}^\perp = 0$, (ii) $S \cap S^\perp \subset \{0\}$, (iii) $S_1 \subset S_2 \Rightarrow S_2^\perp \subset S_1^\perp$ and (iv) $S^\perp$ is a closed linear subspace of $\mathcal{H}$. By a suspace $\tilde{S}$ of $\mathcal{H}$, we mean a subset that is closed under the formation of finite linear combinatons of elements in $\tilde{S}$ (and hence in particular, contains the 0 vector). A subspace of a Hilbert space is not necessarily a Hilbert space, but a closed subspace of a Hilbert space necessarily is. To show that $S^\perp$ is a closed subspace, it suffices to show that if x and y are in $S^\perp$ so is $\alpha\, x + \beta\, y$. But this follows trivially from the definition of $S^\perp$. We only need to verify that $S^\perp$ is closed. So let x_n be a sequence of elements in $S^\perp$ converging to $x \in \mathcal{H}$. Then by the continuity of the inner product, we have $\langle x_n, y \rangle \to \langle x, y \rangle$ for any $y \in \mathcal{H}$. Now for any $y \in S$, $\langle x_n, y \rangle$ is 0 for all n, showing that $\langle x, y \rangle$ must also be 0. But this shows that $x \in S^\perp$. Check that $S \subset S^{\perp\,\perp} \equiv S^{\perp^\perp}$. If S is a non-empty subset of $\mathcal{H}$, then $S^{\perp\,\perp}$ is the closure of the set of all (finite) linear combinations of vectors in S. Thus $S^{\perp\,\perp}$ is the smallest closed linear subspace of $\mathcal{H}$ that contains S. In particular, if S is a subspace, then $S^{\perp\,\perp}$ is the closure of S (which is also a subspace), and in addition, if S is closed $S^{\perp\,\perp} = S$.

Proposition: Let M be a closed linear subspace of S. Then any $h \in \mathcal{H}$ can be written uniquely as $h_1 + h_2$ where $h_1 \in M$ and $h_2 \in M^\perp$. The map $h \mapsto h_1$ is called the orthogonal projection into the subspace M (denoted by π_M) and is a (continuous) linear operator. The map $h \mapsto h_2$ is the orthogonal projection onto $M^\perp$ (written as $\pi_{M^\perp}$).

We will not prove this proposition. For a proof, see for example pages 247–251 of Simmons (Introduction to Topology and Modern Analysis), or a standard textbook in Functional Analysis. Note that

$$\pi_M \, h = \operatorname{argmin}_{w \in M} \|h - w\|^2 \, .$$

That this is the case follows on noting that:

$$\|h - w\|^2 = \|h - \pi_M \, h\|^2 + \|\pi_M \, h - w\|^2 + 2 \, \langle h - \pi_M \, h, \pi_M \, h - w \rangle \, ;$$

now the third term vanishes, since $h - \pi_M h \in M^\perp$ and $\pi_M h - w \in M$. The result follows.

We now return to $L_2(\mu)$ spaces. We have already seen that this is a normed linear (Banach) space. Define an inner product on $L_2(\mu)$ by:

$$\langle f, g \rangle = \int f(x)\, g(x)\; d\,\mu(x)\,.$$

It is not difficult to see that this is well defined, satisfies the criteria for an inner product and induces the $L_2(\mu)$ norm. Hence $L_2(\mu)$ is a Hilbert space with respect to this inner product. For the moment, specialize to the case where μ is a probability measure P and consider the space $L_2(P)$ (with underlying sample space $(\mathcal{X}, \mathcal{A})$. Let Y and X be random variables in $L_2(P)$.

Regression: Consider the regression problem of regressing Y on X. This is the problem of finding that function ϕ of X that minimizes $E(Y - \phi(X))^2$. This can be posed in the following manner: Let $\sigma(X)$ be the sub sigma-field of $\mathcal{A}$ generated by X. Let $\tilde{S}$ denote the set of all random variables in $L_2(P)$ that are measurable with respect to $\sigma(X)$. This is a closed linear subspace of $L_2(P)$. The problem now is one of determining that random variable W in the subspace $\tilde{S}$ that minimizes $\|Y - W\|_2$ where $\| \ \|_2$ denotes the $L_2(P)$ norm. By the general theory of Hilbert spaces, this will be the orthogonal projection of Y onto the subspace $\tilde{S}$. Now note that $Y = E(Y \mid X) + (Y - E(Y \mid X))$, where $E(Y \mid X) \equiv E(Y \mid \sigma(X))$. Clearly $E(Y \mid X)$ lies in $\tilde{S}$. We will show that $Y - E(Y \mid X)$ lies in $\tilde{S}^\perp$, whence it will follows that $E(Y \mid X)$ is the orthogonal projection of Y onto $\tilde{S}$ and hence minimizes $\|Y - W\|_2$ over all W in $\tilde{S}$. Thus, for any $W \in \tilde{S}$ we need to show that $\langle W, Y - E(Y \mid X) \rangle = 0$. Now,

$$\begin{aligned}
\langle W, Y - E(Y \mid X) \rangle &= E\left[W\,(Y - E(Y \mid X))\right] \\
&= E\left[E\left[W\,(Y - E(Y \mid X)) \ \mid \sigma(X)\right]\right] \\
&= E\left[W\left[E(Y \mid X) - E(Y \mid X)\right]\right] \\
&= 0\,.
\end{aligned}$$

Regression problems in statistics can be viewed as those of computing orthogonal projections on appropriate subspaces of L_2 spaces. In linear regression one projects the response variable Y onto the finite dimensional linear subspace of linear combinations of the covariate random variables. This is smaller than the subspace of all random variables that are measurable with respect to the sigma-field generated by the covariate variables. The mean squared error in regression: $E(Y - E(Y \mid X))^2$ can be viewed as the squared length of the projection of Y onto $\tilde{S}^\perp$, the orthogonal complement of $\tilde{S}$.

Exercise 1: Let S_1 and S_2 be two closed linear subspaces of the Hilbert space $\mathcal{H}$ and suppose that S_1 is strictly contained in S_2. For any vector h consider $\pi_{S_1} h$ and $\pi_{S_2} h$. How does the length (norm) of $\pi_{S_1} h$ compare to that of $\pi_{S_2} h$? What is the relation between $\pi_{S_1} h$ and $\pi_{S_1}\, \pi_{S_2} h$?

Exercise 2: Consider a probability space $(\mathcal{X}, \mathcal{A}, P)$. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be sub-sigma fields of $\mathcal{A}$, with $\mathcal{G}_1 \subset \mathcal{G}_2$. Let Y be a random variable defined on $(\mathcal{X}, \mathcal{A})$. From the theory of conditional expectations we know that $E[[Y \mid \mathcal{G}_2] \mid \mathcal{G}_1] = E[Y \mid \mathcal{G}_1]$. Can you derive this from your considerations in Exercise 1? Also, provide a geometric interpretation to the inequality:

$$E\left[Y - E[Y \mid \mathcal{G}_2]\right]^2 \leq E\left[Y - E[Y \mid \mathcal{G}_1]\right]^2 .$$

2 Metrics on spaces of probability measures

Consider the space of all probability measures on a probability space $(\mathcal{X}, \mathcal{A})$. We discuss how this space can be metrized. Metrization of probability measures (i.e. defining a notion of distance) is important, since in statistics one is often concerned about convergence of estimates based on finite samples to the true parameter (which is often a probability measure) and fundamental to a definition of convergence is the notion of a distance.

Two widely used metrics are: (a) Total variation (TV) and (b) Hellinger distance. The TV distance between probability measures P and Q is defined as:

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{A}} \mid P(A) - Q(A) \mid .$$

The Hellinger metric is defined as:

$$H^2(P, Q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, d\mu ,$$

where μ is any measure that dominates both P and Q and p and q are the densities of P and Q with respect to μ. Note that this definition of the Hellinger distance ostensibly depends on the dominating measure μ. However, as we show below, the quantity on the left side of the above display is independent of μ (and the densities p and q). To this end, let $\mu_0 = P_0 + Q_0$, and define:

$$p_0 = \frac{d\,P_0}{d\,\mu_0} \text{ and } q_0 = \frac{d\,P_0}{d\,\mu_0} .$$

Notice that μ_0 dominates both P and Q, so the above derivatives exist. Furthermore, note that μ_0 is dominated by μ, so $d\,\mu_0/d\,\mu$ exists. Also $p = dP/d\,\mu$ and $q = dQ/d\,\mu$. Now,

$$\begin{aligned}
\int (\sqrt{p} - \sqrt{q})^2 \, d\mu &= \int \left[\sqrt{\frac{d\,P}{d\,\mu}} - \sqrt{\frac{d\,Q}{d\,\mu}}\right]^2 d\,\mu \\
&= \int \left[\sqrt{\frac{d\,P}{d\,\mu_0}\,\frac{d\,\mu_0}{d\,\mu}} - \sqrt{\frac{d\,Q}{d\,\mu_0}\,\frac{d\,\mu_0}{d\,\mu}}\right]^2 d\,\mu \\
&= \int \left[\sqrt{\frac{d\,P}{d\,\mu_0}} - \sqrt{\frac{d\,Q}{d\,\mu_0}}\right]^2 \frac{d\,\mu_0}{d\,\mu}\, d\,\mu \\
&= \int \left[\sqrt{\frac{d\,P}{d\,\mu_0}} - \sqrt{\frac{d\,Q}{d\,\mu_0}}\right]^2 d\,\mu_0 .
\end{aligned}$$

This shows the invariance of the Hellinger distance to the choice of the dominating measure μ.

The TV distance can also be characterized in terms of densities p and q with respect to some dominating measure μ. We have:

$$d_{TV}\,(P,Q) = \sup_{A\in\mathcal{A}} \mid P(A) - Q(A) \mid = \frac{1}{2}\int \mid p-q \mid \, d\,\mu\,.$$

Thus the TV distance can be thought of as the natural distance between the densities p and q in $L_1(\mu)$, whereas the Hellinger distance is the natural distance between $\sqrt{p}$ and $\sqrt{q}$ in $L_2(\mu)$. To establish the above display, note that, since $\int (p-q)\,d\,\mu = 0$ we have:

$$\int_{\{p>q\}} (p-q)\,d\mu = \int_{\{p>q\}} \mid p-q \mid \, d\,\mu = \int_{\{q>p\}} (q-p)\,d\mu = \int_{\{q>p\}} \mid q-p \mid \, d\,\mu\,.$$

Let $A_0 = \{p > q\}$ and $B_0 = \{q > p\}$. Then, the above display implies that

$$P(A_0) - Q(A_0) = \mid P(A_0) - Q(A_0) \mid = Q(B_0) - P(B_0) = \mid P(B_0) - Q(B_0) \mid \,.$$

Now, for any set A, we have:

$$\begin{aligned}
P(A) - Q(A) &= \int_A p\,d\,\mu - \int_A q\,d\,\mu \\
&= \int_{A\cap\{p>q\}} (p-q)\,d\,\mu + \int_{A\cap\{p<q\}} (p-q)\,d\,\mu \\
&\leq \int_{A\cap\{p>q\}} (p-q)\,d\,\mu \\
&\leq \int_{\{p>q\}} (p-q)\,d\,\mu \\
&= P(A_0) - Q(A_0) \\
&= \mid P(A_0) - Q(A_0) \mid \,.
\end{aligned}$$

Similarly, we deduce that for any set A:

$$Q(A) - P(A) \leq Q(B_0) - P(B_0) = \mid Q(B_0) - P(B_0) \mid \,.$$

It follows that:

$$\sup_{A\in\mathcal{A}} \mid P(A) - Q(A) \mid \leq \mid P(A_0) - Q(A_0) \mid = \mid Q(B_0) - P(B_0) \mid \,.$$

But

$$\begin{aligned}
\frac{1}{2}\int \mid p-q \mid \, d\,\mu &= \frac{1}{2}\int_{\{p>q\}} (p-q)\,d\mu + \frac{1}{2}\int_{\{q>p\}} (q-p)\,d\mu \\
&= \frac{\mid P(A_0) - Q(A_0) \mid + \mid P(B_0) - Q(B_0) \mid}{2} \\
&= \mid P(A_0) - Q(A_0) \mid \\
&= \mid P(B_0) - Q(B_0) \mid \,.
\end{aligned}$$

Exercise (a): Show that $H^2(P,Q) = 1 - \rho(P,Q)$ where the affinity ρ is defined as $\rho(P,Q) = \int \sqrt{p\,q}\,d\,\mu$.

Exercise (b): Show that $d_{TV}\,(P,Q) = 1 - \int\, p \wedge q\; d\,\mu$.

Exercise (c): Show that the following inequalities hold:

$$H^2(P,Q) \leq d_{TV}\,(P,Q) \leq H(P,Q)\,(1+\rho(P,Q))^{1/2} \leq \sqrt{2}\,H(P,Q)\,.$$

We now introduce an important discrepancy measure that crops up frequently in statistics. This is the Kullback-Leibler distance, though the term "distance" in this context needs to be interpreted with a dose of salt, since the Kullback-Leibler divergence does not satisfy the properties of a distance. For probability measures P, Q dominated by a probability measure μ and densities p and q respectively with respect to μ, we define the Kullback-Leibler divergence as:

$$K(P,Q) = \int \log\,\frac{p}{q}\,p\;d\,\mu \equiv \int \log\,\frac{p}{q}\;dP \equiv E_P\,[\log\,p/q].$$

The above is, at this point, only a formal definition, since we do not know whether the above expectation is meaningfully defined. Recall that any random variable X can be written as $X^+ - X^-$, with $X^+ = \max\{Z,0\}$ and $X^- = \max\{-Z,0\}$ and for EX to exist as a finite quantity both EX^+ and EX^- need to be finite, and then $EX = EX^+ - EX^-$. If both are infinite EX does not exist, and in the case that only one is finite EX is either ∞ or $-\infty$. We show below that $K(P,Q)$ always exists and is either finite or ∞. To that end, we first need a brief discussion on Jensen's inequality.

Jensen's Inequality: Let X be a random variable with $P(X \in I) = 1$ where I is an open sub-interval of $\mathbb{R}$. Assume that EX (exists and) is finite. Let ϕ be a real-valued convex function defined on I. Then $E(\phi(X)) \geq \phi(EX)$.

Note that we *do not require* $\phi(X)$ *to have finite expectation* in the above inequality. As the proof will show, the only two possibilities are for $E(\phi(X))$ to be finite or equal ∞ and the inequality holds in either case (trivially for the latter). By hypothesis $\phi(\lambda\,x + (1-\lambda)\,y) \leq \lambda\,\phi(x) + (1-\lambda)\,\phi(y)$ for $x, y \in I$ and $0 \leq \lambda \leq 1$. This is equivalent to the condition that for all $s < t < u$ with $s, t, u \in I$,

$$\frac{\phi(t) - \phi(s)}{t-s} \leq \frac{\phi(u) - \phi(t)}{u-t}\,.$$

Now, EX in I. Setting $t = EX$ and letting β be the supremum of the quotients on the left side of the above display for $s \in I, s < t$, we obtain: $\phi(u) \geq \phi(EX) + \beta\,(u - EX)$ for all $u > EX$. Similarly, letting γ be the infimum of the quotients on the right side of the above display for $u \in I, u > t$, we obtain: $\phi(s) \geq \phi(EX) + \gamma\,(s - EX)$ for all $s < EX$.(Note that both β and γ are finite numbers.) It follows that with probability 1:

$$\phi(X) \geq \phi(EX) + (\beta \wedge \gamma)\,(X - EX)\,.$$

The right side of the preceding display is an integrable random variable and therefore $E(\phi(X))$ is either finite or ∞. (If $X \geq Y$ and Y has finite expectation, then EX is either finite or equals infinity. This is because $X^- \leq Y^-$ and $E\,Y^- < \infty$. Note that X can be allowed to be an extended real valued random variable.) If $\phi(X)$ has finite expectation, Jensen's inequality follows by taking expectations on either side of the above display. If $E(\phi(X)) = \infty$, Jensen's inequality is trivial.

Proposition: Let ϕ be a map defined on $[0, \infty)$, such that ϕ is real-valued and convex on $(0, \infty)$ and $\phi(0) = \infty$. Let X be a non-negative random variable with a finite expectation. Then $E(\phi(X)) \geq \phi(EX)$.

The proof of this proposition runs along the lines of that of Jensen's inequality. First, if X is degenerate at 0, the result is trivial. Second, if $P(X = 0) = 1$, we are in the previous setting. So, consider the situation that $0 < P(X = 0) < 1$. Using the convexity of ϕ on $(0, \infty)$, conclude that whenever $X > 0$, $\phi(X) \geq \phi(EX) + \lambda\,(X - EX)$ for some finite λ. Whenever $X = 0$, the inequality is trivially satisfied. Thus $\phi(X)$ either has finite expectation, or $E(\phi(X)) = \infty$, as argued before. In this case, $E\phi(X) = \infty$, since $\phi(0) = \infty$ and $P(X = 0) > 0$. So $E\,\phi(X) \geq \phi(EX)$.

We are now in a position to show that $K(P, Q)$ exists and is non-negative. Consider $E_P(\log(p/q))$. Note that the random variable $\log(p/q)$ is well-defined, possibly as an extended random variable, on a set with P–probability 1, namely the set $\{p > 0\}$ and on this set $\log p/q = -\log(q/p)$. Now set $\phi = -\log$ in the above proposition and let $X = q/p$. Note that $E_P\,X$ is finite, and by the proposition, $E_P(-\log(q/p))$ exists, is either finite or ∞ and $K(P, Q) = E_P(\log(p/q)) = E_P(-\log(q/p)) \geq -\log(E_P(X)) = -\log\left(\int_{p>0} q\,d\,\mu\right) \geq 0$. There is one more issue that needs to be commented on. The KL distance is invariant to the choice of the dominating measure μ and the corresponding densities p and q and is an outcome of the fact that $p = (dP/d\mu_0)\,(d\mu_0/d\mu)$ and $q = (dQ/d\mu_0)\,(d\mu_0/d\mu)$ for $\mu_0 = P + Q$ (which is absolutely continuous with respect to μ).

It is clear from the above discussion that the Kulback-Leibler (KL) distance from P to Q, $K(P, Q)$, becomes infinitely large whenever $P(q = 0, p > 0) > 0$. However, this does not imply that $K(Q, P)$ is infinity as well. The KL distance is not symmetric in its arguments. In the discussion on ML estimation in Section 3, assume that all KL distances considered are finite, to avoid complication. We next discuss the connection of the Kullback-Leibler discrepancy to Hellinger distance. We have:

$$K(P, Q) \geq 2\,H^2(P, Q)\,.$$

It is easy to check that

$$2\,H^2(P, Q) = 2\left[\int \left(1 - \sqrt{\frac{q}{p}}\right) p\,d\,\mu\right].$$

Thus, we need to show that:

$$\int \left(-\log\,\frac{q}{p}\right) p\,d\,\mu \geq 2\int \left(1 - \sqrt{\frac{q}{p}}\right) p\,d\,\mu\,.$$

This is equivalent to showing that:

$$\int \left(-\log \sqrt{\frac{q}{p}}\right) p\, d\mu \geq \int \left(1 - \sqrt{\frac{q}{p}}\right) p\, d\mu\,.$$

But this follows immediately from the fact that for any positive x, $\log x \leq x - 1$ whence $-\log x \geq 1 - x$. We thus have:

$$d_{TV}^2\,(P, Q) \leq 2\, H^2(P, Q) \leq K(P, Q)\,.$$

It follows that if the Kullback-Leibler divergence between a sequence of probabilities $\{P_n\}$ and a fixed probability P goes to 0, then convergence of P_n to P must happen both in the Hellinger and the TV sense (the latter two modes of convergence being equivalent).

- Problem (d): Let P_θ and P_η denote the Uniform $(0, \theta)$ and Uniform $(0, \eta)$ distributions with $\theta \neq \eta$. Evaluate $K(P_\theta, P_\eta)$ and show that this is different from $K(P_\eta, P_\theta)$.

- Problem (e): Let $f_1, f_2, \ldots$ and $g_1, g_2, \ldots$ be two sequences of probability density functions defined on a measure space $(\Omega, \mathcal{A}, \mu)$, with $P_1, P_2, \ldots$ and $Q_1, Q_2, \ldots$ being the corresponding distributions. Let $\tilde{P}_n \equiv P_1 \times P_2 \times \ldots P_n$ and $\tilde{Q}_n = Q_1 \times Q_2 \times Q_n$ be the n'th stage product measures, defined on the appropriate product spaces. Calculate $H^2(\tilde{P}_n, \tilde{Q}_n)$ in terms of $H^2(P_i, Q_i)$'s.

- Problem (f): For probability measures P and Q, find the possible limit points of the sequence $H^2(P^n, Q^n)$.

- Problem (g) Consider a one-parameter exponential family model in its natural form: $p(x, \theta) = \exp(\theta\, x - B(\theta))$, with $\theta \in \Theta$, an open subset of $\mathbb{R}$. Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from some $p(x, \theta_0)$.
  (i) Consider testing $H_0 : \theta \in [a, b]$ versus its complement and suppose that $\theta_0 \notin [a, b]$. Let λ_n denote the likelihood ratio statistic for testing H_0. Show that $\log \lambda_n / n$ converges to $K(P_{\theta_0}, P_\eta)$ where η is the point in the null hypothesis closest to θ_0 in KL distance. Compute the asymptotic distribution of $\log \lambda_n / n$.
  (ii) Now let $\theta_0 = 0$ (wlog) and consider testing $H_0 : \theta = 0$ versus $H_1 : \theta \geq 0$. Find the limit distribution of the likelihood ratio statistic for this problem.

3 Connections to Maximum Likelihood Estimation

The Kullback-Leibler divergence is intimately connected with maximum likelihood estimation as we demonstrate below. Consider a random variable/vector whose distribution comes from one of a class of densities $\{p(x, \theta) : \theta \in \Theta\}$. Here Θ is the parameter space (think of this as a subset of a metric space with metric τ). For the following discussion, we do not require to assume that Θ is finite-dimensional. Let θ_0 denote the data generating parameter. The MLE is defined as that value of θ that maximizes the log-likelihood function based on i.i.d. observations $X_1, X_2, \ldots, X_n$ from the underlying distribution; i.e.

$$\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} l(X_i, \theta)\,,$$

where $l(x,\theta) = \log\, p(x,\theta)$. Now

$$K(p_{\theta_0}, p_\theta) = E_{\theta_0}\left[\log\, \frac{p(X,\theta_0)}{p(X,\theta)}\right] \geq 0\,,$$

with equality if and only if $\theta = \theta_0$ (this requires the tacit assumption of identifiability), since:

$$E_{\theta_0}\left[\log\, \frac{p(X,\theta_0)}{p(X,\theta)}\right] = E_{\theta_0}\left[-\log\, \frac{p(X,\theta)}{p(X,\theta_0)}\right] \geq -\log\left[E_{\theta_0}\, \frac{p(X,\theta)}{p(X,\theta_0)}\right] = 0\,.$$

Equality happens if and only if the ratio $p(x,\theta)/p(x,\theta_0)$ is P_{θ_0} a.s. constant, in which case equality must hold a.s P_{θ_0} if P_θ and P_{θ_0} are mutually absolutely continuous. This would imply that the two distribution functions are identical. Now, define $B(\theta) = E_{\theta_0}\,(\log\, p(X,\theta))$. It is then clear that θ_0 is the unique maximizer of of $B(\theta)$. Based on the sample however, there is no way to compute the (theoretical) expectation that defines $B(\theta)$. A surrogate is the expectation of $l(X,\theta) \equiv \log\, p(X,\theta)$ based on the empirical measure $\mathbb{P}_n$ that assigns mass $1/n$ to every X_i. In other words, we stipulate $\hat{\theta}_n$ as an estimate of θ, where:

$$\hat{\theta}_n = \text{argmax}_{\theta\in\Theta}\, \mathbb{P}_n\,(l(X,\theta)) \equiv \text{argmax}_{\theta\in\Theta}\, n^{-1} \sum_{i=1}^{n} l(X_i,\theta) = \text{argmax}_{\theta\in\Theta} \sum_{i=1}^{n} l(X_i,\theta)\,.$$

Since $\mathbb{P}_n(l(X,\theta)) \to E_{\theta_0}\, l(X,\theta)$ a.s., one might expect $\hat{\theta}_n$, the empirical maximizer to converge to θ_0, the theoretical maximizer. This is the intuition behind ML estimation.

Define $g_\theta(X) = \log\,\{p(X,\theta_0)/p(X,\theta)\}$. Then $E_{\theta_0}\, g_\theta(X) = K(p_{\theta_0}, p_\theta)$. From the definition of $\hat{\theta}_n$ it follows that:

$$0 \geq \frac{1}{n}\sum_{i=1}^{n} g_{\hat{\theta}_n}(X_i) = \frac{1}{n}\sum_{i=1}^{n} (g_{\hat{\theta}_n}(X_i) - K(p_{\theta_0}, p_{\hat{\theta}_n})) + K(p_{\theta_0}, p_{\hat{\theta}_n})$$

which implies that:

$$0 \leq K(p_{\theta_0}, p_{\hat{\theta}_n}) \leq \left|\frac{1}{n}\sum_{i=1}^{n} g_{\hat{\theta}_n}(X_i) - K(p_{\theta_0}, p_{\hat{\theta}_n})\right|\,. \tag{3.1}$$

By the strong law of large numbers, it is the case that for each fixed θ,

$$\frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i) - K(p_{\theta_0}, p_\theta) \to_{P_{\theta_0}\, a.s.}\, 0\,.$$

This however does not imply that:

$$\frac{1}{n}\sum_{i=1}^{n} g_{\hat{\theta}_n}(X_i) - K(p_{\theta_0}, p_{\hat{\theta}_n}) \to_{P_{\theta_0}\, a.s.}\, 0$$

since $\hat{\theta}_n$ is a *random argument*. But suppose that we could ensure:

$$\sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^{n} g_\theta(X_i) - K(p_{\theta_0}, p_\theta) \right| \to_{P_{\theta_0}\, a.s.} 0\,.$$

This is called a *Glivenko-Cantelli* condition: it ensures that the strong law of large numbers holds uniformly over a class of functions (which in this case is indexed by θ). Then, by (3.1) we can certainly conclude that $K(p_{\theta_0}, p_{\hat{\theta}_n})$ converges a.s. to 0. By the inequality relating the Hellinger distance to the Kullback-Leibler divergence, we conclude that $H^2(p_{\theta_0}, p_{\hat{\theta}_n})$ converges a.s. to 0. This is called *Hellinger consistency*. However, in many applications, we are really concerned with the consistency of $\hat{\theta}_n$ to θ_0 in the natural metric on Θ (which we denoted by τ). The following proposition, adapted from Van de Geer (Annals of Statistics, **21**, Hellinger consistency of certain nonparametric maximum likelihood estimators, pages 14 – 44) shows that consistency in the natural metric can be deduced from Hellinger consistency under some additional hypotheses.

Proposition: Say that θ_0 is identifiable for the metric τ on Θ if, for all $\theta \in \Theta$, $H(p_\theta, p_{\theta_0}) = 0$ implies that $\tau(\theta, \theta_0) = 0$. Suppose that (a) (Θ, τ) is a compact metric space, (b) $\theta \mapsto p(x, \theta)$ is μ–almost everywhere continuous (here, μ is the underlying dominating measure) in the τ metric, and (c) θ_0 is identifiable for τ. The $H(p_{\theta_n}, p_{\theta_0}) \to 0$ implies that $\tau(\theta_n, \theta_0) \to 0$.

Hence, under the conditions of the above proposition, Hellinger consistency a.s.(in probability) would imply a.s. consistency (in probability) of $\hat{\theta}_n$ for θ_0 in the τ–metric.

Proof: Suppose that $\tau(\theta_n, \theta_0) \nrightarrow 0$. Since $\{\theta_n\}$ lies in a compact set (assumption (a)) and does not converge to θ_0, there exists a subsequence $\{n'\}$ such that $\theta_{n'} \to \theta^\star \neq \theta_0$ in the τ-metric. Note that $h(p_{\theta_{n'}}, p_{\theta_0}) \to 0$. Now, by the triangle inequality:

$$h(p_{\theta_0}, p_{\theta^\star}) \leq h(p_{\theta_{n'}}, p_{\theta_0}) + h(p_{\theta_{n'}}, p_{\theta^\star})\,.$$

The first term on the right side of the above display converges to 0; as for the second term, this also goes to 0, since by Scheffe's theorem, the a.s. convergence of $p(x, \theta_{n'})$ to $p(x, \theta^\star)$ (see assumption (b)) guarantees that convergence of the densities $\{p_{\theta_{n'}}\}$ to $p_{\theta^\star}$ happens in total variation norm, and consequently in the Hellinger metric. Conclude that $h(p_{\theta_0}, p_{\theta^\star}) = 0$; by identifiability (assumption (c)) $\tau(\theta_0, \theta^\star) = 0$. This shows that $\theta_{n'}$ converges to θ_0 and provides a contradiction. □

Exercise: Consider the model $\{\text{Ber}(\theta) : 0 < a \leq \theta \leq b < 1\}$. Consider the M.L.E. of θ based on i.i.d. observations $\{X_i\}_{i=1}^n$ from the Bernoulli distribution, with the true parameter θ_0 lying in (a, b). Use the ideas developed above to show that the MLE converges to the truth, almost surely, in the Euclidean metric. This is admittedly akin to pointing to your nose by wrapping your arm around your head, but nevertheless, illustrates how these techniques work. For "real applications" of these ideas one has to work with high dimensional models, where the niceties of standard parametric inference fail.