Appendix B

Space of Probability Densities

There are many useful distances on a set of density functions, some of a statistical origin. A number of "divergences" lack symmetry, but similarly measure discrepancies between pairs of densities. In this chapter we review many distances, divergences and their relations, with particular attention for distances on probability densities.

Throughout the chapter $(\mathfrak{X}, \mathscr{X})$ is a measurable space equipped with σ-finite measure ν. A "density" is a measurable function $p\colon \mathfrak{X} \to \mathbb{R}$, typically nonnegative and integrable relative to ν. We denote a measurable function $p\colon \mathfrak{X} \to \mathbb{R}$ by a lowercase letter, and the induced measure $A \mapsto \int_A p\,d\nu$ by the corresponding upper case letter P.

B.1 Distances and Divergences

As probability densities are nonnegative and integrate to one, the γth power of their γth root is well defined and integrable (with integral one), and hence the distance $r_\gamma(p,q) = (\int |p^{1/\gamma} - q^{1/\gamma}|^\gamma\, d\nu)^{(1/\gamma)\wedge 1}$ is well defined for any $\gamma > 0$. The special cases $\gamma = 1$ and $\gamma = 2$ are the $\mathbb{L}_1$-*distance* (equivalent with the *total variation distance*) and *Hellinger distance*, given by

$$\|p-q\|_1 = \int |p-q|\,d\nu = 2 - 2\int p\wedge q\,d\nu = 2\|P-Q\|_{TV}, \tag{B.1}$$
$$d_H(p,q) = (\int (\sqrt{p} - \sqrt{q})^2\,d\nu)^{1/2}.$$

The total variation distance was defined in (A.3) as a supremum over sets; in terms of densities this supremum can be seen to be attained at the set $\{x\colon p(x) > q(x)\}$. All distances r_γ are invariant with respect to changing the dominating measure. For the $\mathbb{L}_1$-distance this is clear from its expression as $2\|P-Q\|_{TV}$. A similar, symbolic expression for the Hellinger distance, which also eliminates the dominating measure, is $(\int(\sqrt{dP} - \sqrt{dQ})^2)^{1/2}$. On the set of probability densities the two distances are maximally 2 and $\sqrt{2}$, respectively; the maximum values are attained by pairs of probability densities with disjoint regions of positivity.

For densities whose rth power is integrable, the $\mathbb{L}_r$-*distance* is defined as

$$\|p-q\|_r = \Bigl(\int |p-q|^r\,d\nu\Bigr)^{(1/r)\wedge 1}.$$

The case $r = 2$ is often used in classical density estimation, because the corresponding risks admit explicit expressions. A drawback of these distances is, that they are not invariant under the choice of a dominating measure.

The limiting case $r = \infty$ is the *essential supremum* of $|p - q|$ relative to ν. This is bounded above by the *supremum distance* (or *uniform distance*)

$$\|p - q\|_\infty = \sup_{x \in \mathfrak{X}} |p(x) - q(x)|.$$

Even if this is too strong in some applications, it can be useful as a technical device, especially when calculating the size of a space. Estimates of entropies (see Appendix C) relative to the supremum distance are readily available for several function spaces, and carry over to other distances, by domination (on bounded domains) or under other restrictions (e.g. for normal mixtures). All metrics r_γ for $\gamma \geq 1$ and $\mathbb{L}_r$ for $r \geq 1$ lead to locally convex topologies.

A fundamental quantity, both in non-Bayesian and Bayesian statistics, is the *Kullback-Leibler divergence* (or KL divergence)

$$K(p; q) = \int p \log(p/q)\, d\nu.$$

Here, $\log(p/q)$ is understood to be ∞ if $q = 0 < p$, meaning that $K(p; q) = \infty$ if $P(q = 0) > 0$. (Because $\log_-(p/q) \leq q/p - 1$, and $p(q/p - 1)$ is integrable, the integral $K(p; q)$ is always well defined.) The importance of the KL divergence stems from the fact that the likelihood ratio $\prod_{i=1}^n (q/p)(X_i)$ under i.i.d. sampling from p roughly behaves like $e^{-nK(p;q)}$. Like the total variation and Hellinger distances, the Kullback-Leibler divergence is not dependent on the choice of the dominating measure. Furthermore, it is nonnegative, and $K(p; q) = 0$ if and only if $P = Q$. However it is not a metric – it lacks both symmetry and transitivity. The asymmetry is sometimes rectified by considering the *symmetrized KL divergence* $K_S(p, q) = K(p; q) \wedge K(q; p)$, especially in model selection problems and in the context of reference priors.

Higher-order KL divergences and their centered versions are defined by, for $k > 1$,

$$\begin{aligned} V_k(p; q) &= \int p \big| \log(p/q) \big|^k \, d\nu, \\ V_{k,0}(p; q) &= \int p \big| \log(p/q) - K(p; q) \big|^k \, d\nu. \end{aligned} \tag{B.2}$$

The most important case is $k = 2$. We refer to them as *Kullback-Leibler variations*.

The following lemmas describe relationships between these distances and discrepancies. The $\mathbb{L}_1$- and Hellinger distances induces the same topology, which is stronger than the weak topology, and called the *strong topology* or *norm topology*. The norm and weak topologies agree on norm-compact classes of densities. Furthermore, they induce the same σ-field on any class of densities, by Proposition A.10.

Lemma B.1 *For any pair of probability densities p, q,*

(i) $\|p - q\|_1 \leq d_H(p, q)\sqrt{4 - d_H^2(p, q)} \leq 2d_H(p, q)$.

(ii) $d_H^2(p, q) \leq \|p - q\|_1$.

(iii) $\|p-q\|_1^2 \le 2K(p;q)$ (Kemperman's inequality *or* Pinsker's inequality).
(iv) $d_H^2(p,q) \le K(p;q)$.
(v) $d_H(p,q) \le \|(\sqrt{p}+\sqrt{q})^{-1}\|_\infty \|p-q\|_2$.
(vi) $\|p-q\|_2 \le \|\sqrt{p}+\sqrt{q}\|_\infty d_H(p,q)$.
(vii) $\|p-q\|_r \le \|p-q\|_\infty \nu(\mathfrak{X})^{1/r}$, *for* $r \ge 1$.

Proof For (i) we factorize $|p-q| = |\sqrt{p}-\sqrt{q}|\,|\sqrt{p}+\sqrt{q}|$, and next use the Cauchy-Schwarz inequality to obtain the bound $\|p-q\|_1 \le d_H(p,q)\|\sqrt{p}+\sqrt{q}\|_2$. Finally we compute $\|\sqrt{p}+\sqrt{q}\|_2^2 = 2+2\int\sqrt{p}\sqrt{q}\,d\nu = 4-d_H^2(p,q)$.

Assertion (ii) follows from the inequality $|\sqrt{p}-\sqrt{q}|^2 \le |p-q|$, for all $p,q\ge 0$.

(iii). Apply the inequality $|x-1| \le ((4+2x)/3)^{1/2}(x\log x - x + 1)^{1/2}$, which is valid for $x \ge 0$, followed by the Cauchy-Schwarz inequality, to $\int |p/q-1|\,dQ$.

(iv). Since $\log x \le 2(\sqrt{x}-1)$, for $x\ge 0$, we have $K(p;q) = -2\int p\log(\sqrt{q}/\sqrt{p})\,d\nu \ge -2\int p(\sqrt{q/p}-1)\,d\nu$. This can be rewritten as $2-2\int\sqrt{pq}\,d\nu = d_H^2(p,q)$.

Assertions (v) and (vi) follow from the factorization as (i); (vii) is immediate from the definitions. □

Inequality (iv) cannot be reversed in general, as the Hellinger distance is bounded by $\sqrt{2}$ and the Kullback-Leibler divergence can be infinite. However, if the quotient p/q is bounded (or suitably integrable), then the square Hellinger distance and Kullback-Leibler convergence are (almost) comparable.

Lemma B.2 *For every $b>0$, there exists a constant $\epsilon_b > 0$ such that for all probability densities p and densities q with $0 < d_H^2(p,q) < \epsilon_b P(p/q)^b$,*

$$K(p;q) \lesssim d_H^2(p,q)\Big(1+\frac{1}{b}\log_- d_H(p,q) + \frac{1}{b}\log_+ P\Big(\frac{p}{q}\Big)^b\Big) + 1 - Q(\mathfrak{X}),$$

$$V_2(p;q) \lesssim d_H^2(p,q)\Big(1+\frac{1}{b}\log_- d_H(p,q) + \frac{1}{b}\log_+ P\Big(\frac{p}{q}\Big)^b\Big)^2.$$

Furthermore, for every pair of probability densities p and q and any $0<\epsilon<0.4$,

$$K(p;q) \le d_H^2(p,q)(1+2\log_-\epsilon) + 2P\Big[\Big(\log\frac{p}{q}\Big)\mathbb{1}\{q/p \le \epsilon\}\Big],$$

$$V_2(p;q) \le d_H^2(p,q)(12+2\log_-^2\epsilon) + 8P\Big[\Big(\log\frac{p}{q}\Big)^2\mathbb{1}\{q/p \le \epsilon\}\Big].$$

Consequently, for every pair of probability densities p and q,

$$K(p;q) \lesssim d_H^2(p,q)\Big(1+\log\Big\|\frac{p}{q}\Big\|_\infty\Big) \le 2d_H^2(p,q)\Big\|\frac{p}{q}\Big\|_\infty,$$

$$V_2(p;q) \lesssim d_H^2(p,q)\Big(1+\log\Big\|\frac{p}{q}\Big\|_\infty\Big)^2 \le 2d_H^2(p,q)\Big\|\frac{p}{q}\Big\|_\infty.$$

Proof The function $r\colon (0,\infty)\to\mathbb{R}$ defined implicitly by $\log x = 2(\sqrt{x}-1) - r(x)(\sqrt{x}-1)^2$ possesses the following properties:

(i) r is nonnegative and decreasing.
(ii) $r(x) \sim \log_- x$ as $x \downarrow 0$ and $r(x) \le 2\log_- x$ for $x \in [0, 0.4]$.

(iii) For every $b > 0$ there exists $\epsilon_b' > 0$ such that $x \mapsto x^b r(x)$ is increasing on $[0, \epsilon_b']$.

By these properties and the identity $d_H^2(p,q) = -2P(\sqrt{q/p} - 1) + 1 - Q(\mathfrak{X})$,

$$\begin{aligned} K(p;q) + Q(\mathfrak{X}) - 1 &= d_H^2(p,q) + P\Big[r\Big(\frac{q}{p}\Big)\Big(\sqrt{\frac{q}{p}} - 1\Big)^2\Big] \\ &\le d_H^2(p,q) + r(\epsilon)d_H^2(p,q) + P\Big[r\Big(\frac{q}{p}\Big)\mathbb{1}_{q/p\le\epsilon}\Big], \\ &\le d_H^2(p,q) + 2(\log_- \epsilon)d_H^2(p,q) + 2\epsilon^b \log_- \epsilon\, P\Big(\frac{p}{q}\Big)^b, \end{aligned}$$

for $\epsilon \le \epsilon_b' \wedge 0.4$. The first inequality now follows by choosing $\epsilon^b = d_H^2(p,q)/P(p/q)^b$ and $\epsilon_b = (\epsilon_b' \wedge 0.4)^b$. The first inequality of the second pair is the intermediate second result in the display.

For the proof of the second inequality, we first use the inequality $\log x \le 2(\sqrt{x} - 1)$ to see that

$$P\Big[\Big(\log\frac{p}{q}\Big)^2 \mathbb{1}_{q/p\ge 1}\Big] \le 4P\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 = 4d_H^2(p,q).$$

Next, for $\epsilon \le \epsilon_{b/2}'$, in view of the third property of r,

$$\begin{aligned} P\Big[\Big(\log\frac{p}{q}\Big)^2 \mathbb{1}_{q/p\le 1}\Big] &\le 8P\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 + 2P\Big[r^2\Big(\frac{q}{p}\Big)\Big(\sqrt{\frac{q}{p}} - 1\Big)^4 \mathbb{1}_{q/p\le 1}\Big] \\ &\le 8d_H^2(p,q) + 2r^2(\epsilon)d_H^2(p,q) + 2\epsilon^b r^2(\epsilon)P\Big(\frac{p}{q}\Big)^b. \end{aligned}$$

With $\epsilon^b = d_H^2(p,q)/P(p/q)^b$ and $\epsilon_b \le (0.4 \wedge \epsilon_{b/2}')^b$, this can be bounded by the right side of the second inequality, by property (ii). The second inequality in the second pair is again the intermediate result.

For $b \ge 1$ we can choose $\epsilon_b' = 1$ in property (iii). Furthermore, if we replace the inequality in (ii) by $r(x) \le 2 + 2\log_- x$, which is valid for $x \in [0,1]$, then we can choose $\epsilon_b = 1$ later in the proof. This leads to a multiple of the bounds as obtained before, which are then valid for every $b \ge 2$ and any probability densities p and q with $d_H^2(p,q) \le P(p/q)^b$. Here $P(p/q)^b = Q(p/q)^{b+1} \ge (Q(p/q))^{b+1} \ge 1$ for $b > 1$, by Jensen's inequality. Thus the bounds are true for every sufficiently large b and every p and q with $d_H^2(p,q) \le 1$. Now as $b \uparrow \infty$, we have $b^{-1}\log_- d_H(p,q) \to 0$ and $(P(p/q)^b)^{1/b} \to \|p/q\|_\infty$. □

Lemma B.3 *For any pair of probability densities p and q, and $k \in \mathbb{N}$, $k \ge 2$,*

$$P\Big[e^{|\log(q/p)|} - 1 - |\log(q/p)|\Big] \le 2d_H^2(p,q)\Big\|\frac{p}{q}\Big\|_\infty, \tag{B.3}$$

$$V_k(p;q) \le k!2d_H^2(p,q)\Big\|\frac{p}{q}\Big\|_\infty. \tag{B.4}$$

Proof For every $c \ge 0$ and $x \ge -c$, we have the inequality $e^{|x|} - 1 - |x| \le 2e^c(e^{x/2} - 1)^2$. If $c = \log\|p/q\|_\infty$, then $\log(q/p) \ge -c$ and hence the integrand on the left side of (B.3) is bounded above by $2e^c(e^{\frac{1}{2}\log(q/p)} - 1)^2 = 2e^c(\sqrt{q/p} - 1)^2$. The result now follows by integration.

The second inequality follows by expanding the exponential in the left side of the first inequality. □

The Kullback-Leibler divergence may not be continuous in its arguments, but is always lower semicontinuous in its second argument.

Lemma B.4 *If p, q_n, q are probability densities such that $\|q_n - q\|_1 \to 0$ as $n \to \infty$, then* $\liminf_{n\to\infty} K(p; q_n) \geq K(p; q)$.

Proof If $X_n = q_n/p$ and $X = q/p$, then $X_n \to X$ in P-probability and in mean by assumption. We can write $K(p; q_n)$ as the sum of $\mathrm{E}(\log X_n)\mathbb{1}_{X_n>1}$ and $\mathrm{E}(\log X_n)\mathbb{1}_{X_n\leq 1}$. Because $0 \leq (\log x)\mathbb{1}_{x>1} \leq x$, the sequence $(\log X_n)\mathbb{1}_{X_n>1}$ is dominated by $|X_n|$, and hence is uniformly integrable. Because $x \mapsto (\log x)\mathbb{1}_{x>0}$ is continuous, $\mathrm{E}(\log X_n)\mathbb{1}_{X_n>1} \to \mathrm{E}(\log X)\mathbb{1}_{X>1}$. Because the variables $(\log X_n)\mathbb{1}_{X_n<1}$ are nonpositive, we can apply Fatou's lemma to see that $\limsup \mathrm{E}(\log X_n)\mathbb{1}_{X_n<1} \leq \mathrm{E}(\log X)\mathbb{1}_{X<1}$. □

B.2 Hellinger Transform, Affinity and Other Divergences

Other measures of "divergence" between densities are the *Hellinger transform*, *α-divergence* and *Renyi divergence*, given by

$$\rho_\alpha(p; q) = \int p^\alpha q^{1-\alpha}\, d\nu, \tag{B.5}$$

$$D_\alpha(p; q) = 1 - \rho_\alpha(p; q), \tag{B.6}$$

$$R_\alpha(p; q) = -\log \rho_\alpha(p; q). \tag{B.7}$$

The Hellinger transform at $\alpha = 1/2$ is also called *affinity*. For $0 \leq \alpha \leq 1$ and integrable functions p, q, the Hellinger transform is finite (and bounded by one for probability densities), by Hölder's inequality, but this is not guaranteed for $\alpha > 1$. The Hellinger transform for $\alpha = -1$ is equal to twice the χ^2*-divergence* $\int (p - q)^2/q\, d\nu$.

Lemma B.5 *For a probability density p and a density q, the function $\alpha \mapsto \rho_\alpha(p; q)$ is convex on* $[0, 1]$ *with limits $Q(p > 0)$ and $P(q > 0)$ as $\alpha \downarrow 0$ or $\alpha \uparrow 1$, respectively, and derivatives from the right and left equal to $-K(q\mathbb{1}_{p>0}; p)$ and $K(p\mathbb{1}_{q>0}; q)$ at $\alpha = 0$ and $\alpha = 1$, respectively. Furthermore, for $0 < \alpha < 1$ and probability densities p, q,*

(i) $2D_{1/2}(p; q) = d_H^2(p, q) = 2 - 2\rho_{1/2}(p, q)$.
(ii) $\|p - q\|_1^2 \leq 4(1 - \rho_{1/2}^2(p, q))$
(iii) $\rho_{1/2}(p, q) \geq 1 - \|p - q\|_1/2$.
(iv) $\min(\alpha, 1 - \alpha)d_H^2(p, q) \leq D_\alpha(p; q)$.
(v) $D_\alpha(p; q) \leq d_H^2(p, q)$.
(vi) $D_\alpha(p; q) \leq R_\alpha(p; q) \leq D_\alpha(p; q)/\rho_\alpha(p; q)$.

Proof The function $\alpha \mapsto e^{\alpha y}$ is convex on $(0, 1)$ for all $y \in [-\infty, \infty)$, implying the convexity of $\alpha \mapsto \rho_\alpha(q; p) = P(q/p)^\alpha$ on $(0, 1)$. The function $\alpha \mapsto y^\alpha = e^{\alpha \log y}$ is continuous on [0, 1] for any $y > 0$, is decreasing for $y < 1$, increasing for $y > 1$ and constant for $y = 1$. By the monotone convergence theorem, as $\alpha \downarrow 0$,

$$Q\Big[\Big(\frac{p}{q}\Big)^{\alpha}\mathbb{1}_{0<p<q}\Big] \uparrow Q\Big[\Big(\frac{p}{q}\Big)^{0}\mathbb{1}_{0<p<q}\Big] = Q(0 < p < q). \tag{B.8}$$

By the dominated convergence theorem, with the functions $(p/q)^{\alpha}\mathbb{1}_{p\geq q}$ for $\alpha \leq \frac{1}{2}$ dominated by $(p/q)^{1/2}\mathbb{1}_{p\geq q}$, we have, as $\alpha \to 0$,

$$Q\Big[\Big(\frac{p}{q}\Big)^{\alpha}\mathbb{1}_{p\geq q}\Big] \to Q\Big[\Big(\frac{p}{q}\Big)^{0}\mathbb{1}_{p\geq q}\Big] = Q(p \geq q). \tag{B.9}$$

Together (B.8) and (B.9) show that $\rho_\alpha(q; p) = Q(p/q)^{\alpha} \to Q(p > 0)$, as $\alpha \downarrow 0$.

By the convexity of the function $\alpha \mapsto e^{\alpha y}$, the map $\alpha \mapsto f_\alpha(y) = (e^{\alpha y} - 1)/\alpha$ decreases as $\alpha \downarrow 0$, to $(d/d\alpha)|_{\alpha=0}e^{\alpha y} = y$, for every y. For $y \leq 0$ we have $f_\alpha(y) \leq 0$, while for $y \geq 0$, by Taylor's formula,

$$f_\alpha(y) \leq \sup_{0<\alpha'\leq\alpha} y e^{\alpha' y} \leq y e^{\alpha y} \leq \epsilon^{-1}e^{(\alpha+\epsilon)y}.$$

Hence we conclude that $f_\alpha(y) \leq 0 \vee \epsilon^{-1}e^{(\alpha+\epsilon)y}\mathbb{1}_{y\geq 0}$. Consequently $\alpha^{-1}(e^{\alpha\log(p/q)} - 1)$ decreases to $\log(p/q)$ as $\alpha \downarrow 0$ and is bounded above by $0 \vee \epsilon^{-1}(p/q)^{2\epsilon}\mathbb{1}_{p\geq q}$ for small $\alpha > 0$, which is Q-integrable for $\epsilon < \frac{1}{2}$. We conclude that

$$\frac{1}{\alpha}(\rho_\alpha(p; q) - \rho_0(p; q)) = \frac{1}{\alpha}Q\Big[((p/q)^{\alpha} - 1)\mathbb{1}_{p>0}\Big] \downarrow Q\Big[\log(p/q)\mathbb{1}_{p>0}\Big],$$

as $\alpha \downarrow 0$, by the monotone convergence theorem.

Assertion (i) is clear by expanding the square in the integrand of d_H^2. Next (ii) and (iii) are rewrites of the first (not the second!) inequality in (i) and of (ii) of Lemma B.1, respectively.

(iv). By convexity $\rho_\alpha(p; q) \leq (1 - 2\alpha)\rho_0(p; q) + 2\alpha\rho_{1/2}(p; q)$, for $\alpha \leq 1/2$, which can be further bounded by $1 - 2\alpha + 2\alpha\rho_{1/2}(p; q)$. This can be rearranged to obtain (iv) for $\alpha \leq 1/2$. The case $\alpha > 1/2$ follows similarly or by symmetry.

(v). By convexity $\rho_{1/2}(p; q) \leq (1/2)\rho_\alpha(p; q) + (1/2)\rho_{1-\alpha}(p; q)$, for $\alpha \leq 1/2$, which can be further bounded by $(1/2)\rho_\alpha(p; q) + 1/2$.

(vi). This follows by applying the inequalities $1-x \leq -\log x \leq x^{-1}-1$, valid for $x > 0$, to $x = \rho_\alpha(p; q)$. □

If $Q \ll P$, then $K(q\mathbb{1}_{p>0}; p) = K(q; p)$, and hence the lemma shows that the right derivative at $\alpha = 0$ of the Hellinger transform is equal to $-K(q; p)$. Similarly if $P \ll Q$, then the left derivative at $\alpha = 1$ is $K(p; q)$. Without absolute continuity minus/plus these derivatives are not equal to the Kullback-Leibler divergences, and may also be negative, even when both P and Q are probability measures.

The following two lemmas quantify the remainder if $\rho_\alpha(p; q)$ is approximated by its Taylor expansion at $\alpha = 0$, in a "misspecified setting", when the "true measure" P_0 can be different from P.

Lemma B.6 *There exists a universal constant C such that for any probability measure P_0 and all finite measures $P, Q, Q_1, \ldots, Q_m$ and constants $0 < \alpha \leq 1$, $\lambda_i \geq 0$ with $\sum_{i=1}^m \lambda_i = 1$,*

$$\Big|1 - P_0\Big(\frac{q}{p}\Big)^{\alpha} - \alpha P_0 \log\frac{p}{q}\Big| \le \alpha^2 C P_0\Big[\log^2\frac{p}{q}\Big(\Big(\frac{q}{p}\Big)^{\alpha} \mathbb{1}\{q > p\} + \mathbb{1}\{q \le p\}\Big)\Big],$$

$$\Big|1 - P_0\Big(\frac{\sum_{i=1}^m \lambda_i q_i}{p}\Big)^{\alpha} - \alpha P_0 \log\frac{p}{\sum_{i=1}^m \lambda_i q_i}\Big| \le 2\alpha^2 C \sum_{i=1}^m \lambda_i P_0\Big[\log^2\frac{q_i}{p}\Big\{\Big(\frac{q_i}{p}\Big)^2 + 1\Big\}\Big].$$

Proof The function R defined by $R(x) = (e^x - 1 - x)/(x^2 e^x)$ if $x > 0$, $(e^x - 1 - x)/x^2$ if $x < 0$ and $R(0) = \frac{1}{2}$ is uniformly bounded on $\mathbb{R}$ by a constant $C < \infty$. The first inequality follows by applying this with $x = \alpha\log(q/p)$.

Taking $q = \sum_{i=1}^m \lambda_i q_i$ in the first inequality, we see that for the proof of the second it suffices to bound

$$P_0\Big[\Big(\log\frac{\sum_{i=1}^m \lambda_i q_i}{p}\Big)^2\Big\{\Big(\frac{\sum_{i=1}^m \lambda_i q_i}{p}\Big)^{\alpha} \mathbb{1}\{\sum_{i=1}^m \lambda_i q_i > p\} + \mathbb{1}\{\sum_{i=1}^m \lambda_i q_i \le p\}\Big\}\Big].$$

Replacing α by 2 makes the expression larger. Next we bound the two terms corresponding to the decomposition by indicators separately.

If $\sum_{i=1}^m \lambda_i q_i > p$, then we first bound, using convexity of $x \mapsto x\log x$,

$$\Big(\log\frac{\sum_{i=1}^m \lambda_i q_i}{p}\Big)\Big(\frac{\sum_{i=1}^m \lambda_i q_i}{p}\Big) \le \sum_{i=1}^m \lambda_i \Big(\log\frac{q_i}{p}\Big)\Big(\frac{q_i}{p}\Big). \tag{B.10}$$

Since the left side is positive, squaring preserves the inequality, where on the right side the square can be lowered within the sum, by Jensen's inequality.

If $\sum_{i=1}^m \lambda_i q_i < p$, then concavity of the logarithm gives

$$0 < -\log\big(\sum_{i=1}^m \lambda_i q_i/p\big) \le -\sum_{i=1}^m \lambda_i \log(q_i/p).$$

As before, we square this and next lower the square into the sum by Jensen's inequality. □

Lemma B.7 *There exists a universal constant C such for any probability measure P_0 and any finite measures $P, Q, Q_1, \ldots, Q_m$ and any $0 < \alpha \le 1$, $\lambda_1, \ldots, \lambda_m \ge 0$ with $\sum_{i=1}^m \lambda_i = 1$ and $0 < \alpha \le 1$,*

$$\Big|1 - P_0\Big(\frac{q}{p}\Big)^{\alpha} - \alpha P_0 \log\frac{p}{q}\Big| \le \alpha^2 C P_0\Big[\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 \mathbb{1}\{q > p\} + \log^2\frac{p}{q}\mathbb{1}\{q \le p\}\Big],$$

$$\Big|1 - P_0\Big(\frac{\sum_{i=1}^m \lambda_i q_i}{p}\Big)^{\alpha} - \alpha P_0 \log\frac{p}{\sum_{i=1}^m \lambda_i q_i}\Big| \le 2\alpha^2 C \sum_{i=1}^m \lambda_i P_0\Big[\Big(\sqrt{\frac{q_i}{p}} - 1\Big)^2 + \log^2\frac{q_i}{p}\Big].$$

Proof The function R defined by $R(x) = (e^x - 1 - x)/\alpha^2(e^{x/2\alpha} - 1)^2$ for $x \ge 0$ and $R(x) = (e^x - 1 - x)/x^2$ for $x \le 0$ is uniformly bounded on $\mathbb{R}$ by a constant C, independent of $\alpha \in (0, 1]$. Now the proof follows by proceeding as in the proof of Lemma B.6 and using the convexity of $x \mapsto |\sqrt{x} - 1|^2$ on $[0, \infty)$. □

B.3 Product Densities

Given densities p_i relative to σ-finite measures ν_i on measurable spaces $(\mathfrak{X}_i, \mathscr{X}_i)$, let $\otimes_{i=1}^n p_i$ denote the density of the product measure. Some distances are friendlier than other with product densities. For instance, the following lemma bounds the rate of growth of the $\mathbb{L}_1$-distance between products of n identical densities as n, whereas for the Hellinger distance the rate is $\sqrt{n}$.

Lemma B.8 *For any probability densities $p_1, \dots, p_n$ and $q_1, \dots, q_n$,*

(i) $\| \otimes_{i=1}^n p_i - \otimes_{i=1}^n q_i \|_1 \le \sum_{i=1}^n \|p_i - q_i\|_1$.
(ii) $\rho_\alpha(\otimes_{i=1}^n p_i; \otimes_{i=1}^n q_i) = \prod_{i=1}^n \rho_\alpha(p_i; q_i)$.
(iii) $d_H^2(\otimes_{i=1}^n p_i, \otimes_{i=1}^n q_i) \le \sum_{i=1}^n d_H^2(p_i, q_i)$.
(iv) $\rho_\alpha(\otimes_{i=1}^n p_i; \otimes_{i=1}^n q_i) \le \prod_{i=1}^n [1 - \beta d_H^2(p_i, q_i)]$, *for* $\beta = \alpha \wedge (1-\alpha)$.
(v) $K(\otimes_{i=1}^n p_i; \otimes_{i=1}^n q_i) = \sum_{i=1}^n K(p_i; q_i)$.
(vi) $V_{2,0}\left(\otimes_{i=1}^n p_i; \otimes_{i=1}^n q_i\right) = \sum_{i=1}^n V_{2,0}(p_i; q_i)$.
(vii) $V_{k,0}\left(\otimes_{i=1}^n p_i; \otimes_{i=1}^n q_i\right) \le d_k n^{k/2-1} \sum_{i=1}^n V_{k,0}(p_i; q_i)$, *for a constant d_k that depends on $k \ge 2$ only.*

Proof For (iii), we rewrite the Hellinger distance in terms of the affinity, and use the inequality $\prod_i (1 - x_i) \ge 1 - \sum_i x_i$, which is valid for any numbers $x_i \in [0, 1]$, with $1 - x_i$ the affinities. For (iv) we employ Lemma B.5(iv).

The left side of (vii) is $\mathrm{E}|\sum_{i=1}^n (Y_i - \mathrm{E}(Y_i))|^k$, for $Y_i = \log(p_i/q_i)$, where the expectation is taken with respect to the density $\otimes_{i=1}^n p_i$. The Marcinkiewicz-Zygmund inequality (see Lemma K.4) implies that this is bounded by $d_k n^{k/2-1} \sum_{i=1}^n \mathrm{E}|Y_i - \mathrm{E}(Y_i)|^k$, which is the right side of (vii). The proofs of the other assertions are easier. □

Corollary B.9 (Kakutani's criterion) *Two infinite product probability measures $\prod_{i=1}^\infty P_i$ and $\prod_{i=1}^\infty Q_i$ are either mutually absolutely continuous or mutually singular, depending on whether $\sum_{i=1}^\infty d_H^2(P_i, Q_i)$ converges or diverges; equivalently on whether $\prod_{i=1}^\infty \rho_{1/2}(P_i; Q_i)$ is positive or zero. In particular, P^∞ and Q^∞ are mutually singular if $P \ne Q$.*

B.4 Discretization

Given a measurable partition $\{\mathfrak{X}_1, \dots, \mathfrak{X}_k\}$ of $\mathfrak{X}$ in sets of finite ν-measure, and a density p, consider the density p^* defined by

$$p^*(x) = \sum_{j=1}^k \frac{P(\mathfrak{X}_j)}{\nu(\mathfrak{X}_j)} \mathbb{1}\{x \in \mathfrak{X}_j\}. \tag{B.11}$$

The density p^* is constant on each of the partitioning sets, with level equal to the average of p over the set. The following lemma measures the "distance" between p and its "locally uniform discretization" p^* in terms of the *variation* of p or $\log p$ over the partition. The variation of a function $p\colon \mathfrak{X} \to \mathbb{R}$ is defined by

$$\Delta p = \max_{j=1,\ldots,k} \sup_{x,y\in\mathfrak{X}_j} |p(x)-p(y)|. \tag{B.12}$$

Lemma B.10 *For any nonnegative density p and finite, measurable partition* $\{\mathfrak{X}_1,\ldots,\mathfrak{X}_k\}$ *in sets of finite measure,*

(i) $\|p-p^*\|_1 \le \nu(\mathfrak{X})\|p-p^*\|_\infty \le \nu(\mathfrak{X})\,\Delta p$,
(ii) $K(p;p^*) \le \Delta \log p$,
(iii) $\|p\|_\infty^{-1}\Delta p \le \Delta \log p \le \|p^{-1}\|_\infty \Delta p$ *if p is bounded away from zero and infinity.*

Furthermore, if ν *is a probability measure and* $\int p\log p\,d\nu < \infty$, *then* $\int p\log p^* d\nu \to K(p;\nu)$ *as* $k\to\infty$.

Proof The proofs of (i) and (iii) are straightforward. For the proof of (ii), by concavity of the logarithm,

$$K(p;p^*) = \sum_{j=1}^k \int_{\mathfrak{X}_j} p\log\frac{p}{\int_{\mathfrak{X}_j} p\,d\nu/\nu(\mathfrak{X}_j)}\,d\nu \le \sum_{j=1}^k \int_{\mathfrak{X}_j}\int_{\mathfrak{X}_j} p(x)\log\frac{p(x)}{p(y)}\,\frac{d\nu(y)}{\nu(\mathfrak{X}_j)}\,d\nu(x).$$

Here $\log(p(x)/p(y))$ can be bounded uniformly by $\Delta\log p$, after which the remaining integral can be evaluated as one.

To prove the final assertion, write $\int \log(p/p^*)p_0\,d\nu = K(p;\nu) - \int(\log p^*)\,p^*\,d\nu$. Let $\mathscr{T}_k = \sigma\langle\mathfrak{X}_1,\ldots,\mathfrak{X}_k\rangle$. By the convexity of the function $x\mapsto x\log x$ and Jensen's inequality, $p^*\log p^* \le \mathrm{E}_\nu(p\log p|\,\mathscr{T}_k)$, which is ν-uniformly integrable, as $p\log p$ is integrable by assumption. Because also $p^*\log p^* \ge -e^{-1}$ it follows that the sequence $p^*\log p^*$, as k varies, is ν-uniformly integrable, whence $\int p\log p^*\,d\nu = \int p^*\log p^*\,d\nu \to K(p;\nu)$. □

B.5 Information Loss

Kullback-Leibler information and negative Hellinger affinities are measures of statistical separation. They decrease if statistical information is lost by mapping or randomization.

Lemma B.11 *If* $\tilde p$ *and* $\tilde q$ *are the densities of a measurable function* $T(X,U)$ *of a variable* $U\sim \mathrm{Unif}[0,1]$ *and an independent variable X with densities p or q, respectively, then*

$$\begin{aligned} K(\tilde p;\tilde q) &\le K(p;q), \\ \rho_\alpha(\tilde p;\tilde q) &\ge \rho_\alpha(p;q), \qquad 0<\alpha<1, \\ d_H(\tilde p,\tilde q) &\le d_H(p,q), \\ \|\tilde p-\tilde q\|_1 &\le \|p-q\|_1. \end{aligned}$$

Proof Because all four quantities are independent of the dominating measure ν, we may without loss of generality assume that ν is a probability measure. Then $\mathrm{P}_p(T\in A) = \mathrm{E}_\nu \mathbb{1}\{T\in A\}p(X) = \mathrm{E}_\nu\mathbb{1}\{T\in A\}\mathrm{E}_\nu(p(X)|\,T)$, and hence $\tilde p(T) = \mathrm{E}_\nu(p(X)|\,T)$ is a density of $\tilde p$ relative to $\nu\circ T^{-1}$. The analogous formula for $\tilde q$, convexity of the map $(u,v)\mapsto$

$u\log(u/v)$ on $[0,\infty)$ and Jensen's inequality give

$$K(\tilde{p};\tilde{q}) = \mathrm{E}_{\nu\circ T^{-1}}\tilde{p}(T)\log\frac{\tilde{p}}{\tilde{q}}(T) \le \mathrm{E}_{\nu\circ T^{-1}}\mathrm{E}_{\nu}\Big[p(X)\log\frac{p}{q}(X)|\,T\Big] = K(p;q).$$

The assertion for the affinity, Hellinger distance and total variation follow similarly, now using the concavity of $(u,v)\mapsto u^{\alpha}v^{1-\alpha}$, or the convexity of the maps $(u,v)\mapsto|\sqrt{u}-\sqrt{v}|^2$ or $(u,v)\mapsto|u-v|$. (Alternatively, the Hellinger distance can be treated by writing its square as $(1-\rho_{1/2})/2$.) □

Example B.12 The mixture densities $\int p_\theta\,dG$ and $\int q_\theta\,dG$ obtained from two families of (jointly measurable) densities p_θ and q_θ on a measurable space indexed by a common parameter θ and a given probability distribution G satisfy

$$d_H^2\Big(\int p_\theta\,dG(\theta), \int q_\theta\,dG(\theta)\Big) \le \int d_H^2(p_\theta,q_\theta)\,dG(\theta). \tag{B.13}$$

In particular, for any three Lebesgue densities p, q and ϕ on $\mathbb{R}^d$,

$$d_H(\phi * p, \phi * q) \le d_H(p,q). \tag{B.14}$$

Similar inequalities are valid for the Kullback-Leibler divergence, the total variation distance and (the reverse for) the Hellinger transform.

To see this, consider the model in which $Z|\,\theta$ has density p_θ or q_θ, and $\theta\sim G$. Then Z is distributed according to the mixture density and (Z,θ) is distributed according to $p = G\otimes p_\theta$ or $q = G\otimes q_\theta$, respectively. Hence by Lemma B.11 with $T(Z,\theta,U) = Z$, the Hellinger distance between the mixtures is bounded by the Hellinger distance between the measures p and q, which has square $\int\int(\sqrt{p}_\theta-\sqrt{q}_\theta)^2\,d\nu\,dG(\theta) = \int d_H^2(p_\theta,q_\theta)\,dG(\theta)$.

B.6 Signed Kullback-Leibler and Comparisons

Define the *positive part Kullback-Leibler divergence* and *negative part Kullback-Leibler divergence* respectively by $K^+(p;q) = P\log_+(p/q)$ and $K^-(p;q) = P\log_-(p/q)$, and define positive and negative Kullback-Leibler variations $V^{\pm}(p;q) = P(\log_{\pm}(p/q))^2$, $V_0^{\pm}(p;q) = P(\log(p/q)-K(p;q))_{\pm}^2$, higher-order variations $V_k^{\pm}(p;q) = P(\log_{\pm}(p/q))^k$, $V_{k,0}^{\pm}(p;q) = P(\log(p/q)-K(p;q))_{\pm}^k$, similarly. Note that $K = K^+ - K^-$ and $V = V^+ + V^-$.

Lemma B.13 *For any two densities p and q,*

$$K^-(p;q) \le \tfrac{1}{2}\|p-q\|_1 \le \sqrt{\tfrac{1}{2}K(p;q)}, \tag{B.15}$$

$$K^+(p;q) \le \tfrac{1}{2}\|p-q\|_1 + K(p;q) \le K(p;q) + \sqrt{\tfrac{1}{2}K(p;q)}, \tag{B.16}$$

$$V^-(p;q) \le 4d_H^2(p,q) \le 4K(p;q). \tag{B.17}$$

Proof Using $\log x \le x-1$ and (B.1), we obtain

$$K^-(p;q) = \int_{q>p} p\log(q/p)\,d\nu \le \int_{q>p}(q-p)\,d\nu = \|p-q\|_1/2.$$

The first part of relation (B.16) follows because $K^+ = K + K^-$. The second parts of (B.15) and (B.16) follow from Kemperman's inequality connecting the KL divergence and the $\mathbb{L}_1$-distance. The first inequality on V^- follows using the inequality $(\log_- x)^2 \leq 4(\sqrt{x} - 1)^2$; the second is part (iv) of Lemma B.7. □

The following results bound the Kullback-Leibler divergence and variation in terms of the same objects evaluated at a third density.

Lemma B.14 *For any probability densities p, q, r,*

$$K(p; q) \leq \log\Big\|\frac{p}{r}\Big\|_\infty + \Big\|\frac{p}{r}\Big\|_\infty\Big(K(r; q) + \sqrt{\tfrac{1}{2}K(r; q)}\Big),$$
$$V^+(p; q) \leq 2\Big\|\frac{p}{r}\Big\|_\infty \log^2\Big\|\frac{p}{r}\Big\|_\infty + 2\Big\|\frac{p}{r}\Big\|_\infty V^+(r; q).$$

Proof If $p \leq Cr$, then $K(p; q)$ is bounded by $P\log(Cr/q) = \log C + P\log(r/q)$. We bound $\log(r/q)$ by its positive part, and next P by CR, and finally apply (B.16). The inequality for V^+ follows similarly. □

Lemma B.15 *For any probability densities p, q and r,*

$$K(p; q) \leq K(p; r) + 2d_H(r, q)\Big\|\frac{p}{q}\Big\|_\infty^{1/2},$$
$$V_2(p; q) \leq 4V_2(p; r) + 16d_H^2(p, r) + 16d_H^2(r, q)\Big\|\frac{p}{q}\Big\|_\infty + 16d_H^2(p, q).$$

Here p/q is read as 0 if $p = 0$ and as ∞ if $q = 0 < p$.

Proof Writing $P\log(p/q) = P\log(p/r) + P\log(r/q)$ and using $\log x \leq 2(\sqrt{x} - 1)$ and the Cauchy-Schwarz inequality, we have

$$P\log\frac{r}{q} \leq 2\int\frac{p}{\sqrt{q}}(\sqrt{r} - \sqrt{q}) \leq 2\Big\|\frac{p}{q}\Big\|_\infty^{1/2}\int\sqrt{p}(\sqrt{r} - \sqrt{q})\,d\lambda \leq 2\Big\|\frac{p}{q}\Big\|_\infty^{1/2} d_H(r, q).$$

By the relations $\log_+ x \leq 2|\sqrt{x} - 1|$ and $\log_- x = \log_+(1/x) \leq 2|\sqrt{1/x} - 1|$, we have for any probability densities p, q, r,

$$P\log_+^2\frac{r}{q} \leq 4P\Big(\sqrt{\frac{r}{q}} - 1\Big)^2 \leq 4\Big\|\frac{p}{q}\Big\|_\infty d_H^2(r, q),$$
$$P\log_-^2\frac{p}{q} \leq 4P\Big(\sqrt{\frac{q}{p}} - 1\Big)^2 = 4d_H^2(p, q).$$

Since $|\log(p/q)| \leq \log(p/r) + \log_-(p/r) + \log_+(r/q) + \log_-(p/q)$ the second relation follows from the triangle inequality for the $\mathbb{L}_2(P)$-norm. □

Problems

B.1 (Normal distribution) Let f_μ be the density of the Nor(μ, σ^2)-distribution, for given $\sigma > 0$. Show that

(i) $\|f_\mu - f_\nu\|_1 \le \sqrt{2/\pi}|\mu-\nu|/\sigma$,
(ii) $\rho_{1/2}(f_\mu, f_\nu) = e^{-(\mu-\nu)^2/(8\sigma^2)}$,
(iii) $d_H(f_\mu, f_\nu) \le |\mu-\nu|/(2\sigma)$,
(iv) $K(f_\mu; f_\nu) = (\mu-\nu)^2/(2\sigma^2)$,
(v) $V_{2,0}(f_\mu; f_\nu) = (\mu-\nu)^2/\sigma^2$.

Find the corresponding expressions for multivariate normal densities.

B.2 (Bernoulli distribution) Let p_θ stand for the density of the $\text{Bin}(1,\theta)$-distribution. Show that

(i) $\|p_\mu - p_\nu\|_1 = 2|\mu-\nu|$,
(ii) $\rho_{1/2}(p_\mu, p_\nu) = \sqrt{\mu\nu} + \sqrt{1-\mu}\sqrt{1-\nu}$,
(iii) $d_H^2(p_\mu, p_\nu) = (\sqrt{\mu}-\sqrt{\nu})^2 + (\sqrt{1-\mu}-\sqrt{1-\nu})^2$,
(iv) $K(p_\mu; p_\nu) = \mu\log(\mu/\nu) + (1-\mu)\log((1-\mu)/(1-\nu))$,
(v) $V_2(p_\mu; p_\nu) = \mu\log^2(\mu/\nu) + (1-\mu)\log^2((1-\mu)/(1-\nu))$.

Show that the latter three quantities are bounded above by a multiple of $|\mu-\nu|^2$ if μ and ν are bounded away from 0 and 1.

B.3 The condition of boundedness of likelihood ratio in Lemma B.3 can be relaxed to an integrability condition, but then the bound will not be equally sharp. Show that for any pair of probability densities p and p_0 such that $P_0(p_0/p) < \infty$,

$$P_0\Big(e^{|\log(p/p_0)|} - 1 - |\log(p/p_0)|\Big) \le 4d_H^2(p,p_0)\Big(1 + \Phi^{-1}(d_H^2(p,p_0))\Big),$$

for $\Phi^{-1}(\epsilon) = \sup\{M\colon \Phi(M) \ge \epsilon\}$ inverse of $\Phi(M) = M^{-1}P_0(p_0/p)\mathbb{1}\{p_0/p \ge M\}$.

B.4 Let P and Q be probability measures and define $K^*(P;Q) = P[\log(p/q)\mathbb{1}\{q>0\}]$. Show that $K^*(P^n;Q^n) = K^*(P;Q)(P\{q>0\})^{n-1}$.

B.5 Let $\mathfrak{F}$ be a class of densities and $f_0 \in \mathfrak{F}$ such that for every x the map $f \mapsto f(x)$ is continuous on $\mathfrak{F}$. Show that $f \mapsto K(f_0; f)$ is a measurable (possibly extended) real-valued functional. In particular, show that Kullback-Leibler neighborhood $\{f\colon K(f_0;f) < \epsilon\}$ is a Borel measurable subset of $\mathfrak{F}$.

B.6 By developing the exponentials in their power series, show that $(e^x-1-x)/(e^{x/2}-1)^2$ and $\alpha^{-1}(e^x-1)/(e^{\alpha x}-1)$ are bounded independently of α.

B.7 Show that $K^-(p;q) \le \|p-q\|_1 - d_H^2(p;q)$.

B.8 Using the convexity of $x \mapsto x(\log_+ x)^2$, show that the positive part Kullback-Leibler variation satisfies a information loss inequality analogous to Lemma B.11.