In Lecture 2 we introduced the KL divergence that measures the dissimilarity between two distributions. This turns out to be a special case of the family of f-divergence between probability distributions, introduced by Csiszár [Csi67]. Like KL-divergence, f-divergences satisfy a number of useful properties:

- operational significance: KL divergence forms a basis of information theory by yielding fundamental answers to questions in channel coding and data compression. Simiarly, f-divergences such as χ^2, H^2 and TV have their foundational roles in parameter estimation, high-dimensional statistics and hypothesis testing, respectively.
- invariance to bijective transformations of the alphabet
- data-processing inequality
- variational representations (à la Donsker-Varadhan)
- local behavior given by χ^2 (in non-parametric cases) or Fisher information (in parametric cases).

The purpose of this Lecture is to establish these properties and prepare the ground for applications in subsequent chapters. The important highlight is a *joint range* Theorem of Harremoës and Vajda [HV11], which gives the sharpest possible comparison inequality between arbitrary f-divergences (and puts an end to a long sequence of results starting from Pinsker's inequality). This material can be skimmed on the first reading and referenced later upon need.

7.1 Definition and basic properties of f-divergences

Definition 7.1 (f-divergence)**.** Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Let P and Q be two probability distributions on a measurable space $(\mathcal{X}, \mathcal{F})$. If $P \ll Q$ then the f-divergence is defined as

$$D_f(P\|Q) \triangleq \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] \tag{7.1}$$

where $\frac{dP}{dQ}$ is a Radon-Nikodym derivative and $f(0) \triangleq f(0+)$. More generally, let $f'(\infty) \triangleq \lim_{x\downarrow 0} xf(1/x)$. Suppose that $Q(dx) = q(x)\mu(dx)$ and $P(dx) = p(x)\mu(dx)$ for some common dominating measure μ, then we have

$$D_f(P\|Q) = \int_{q>0} q(x) f\left(\frac{p(x)}{q(x)}\right) d\mu + f'(\infty)P[q = 0] \tag{7.2}$$

with the agreement that if $P[q = 0] = 0$ the last term is taken to be zero regardless of the value of $f'(\infty)$ (which could be infinite).

Remark 7.1. For the discrete case, with $Q(x)$ and $P(x)$ being the respective pmfs, we can also write

$$D_f(P\|Q) = \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right)$$

with the understanding that

- $f(0) = f(0+)$,
- $0f(\frac{0}{0}) = 0$, and
- $0f(\frac{a}{0}) = \lim_{x\downarrow 0} xf(\frac{a}{x}) = af'(\infty)$ for $a > 0$.

Remark 7.2. A nice property of $D_f(P\|Q)$ is that the definition is invariant to the choice of the dominating measure μ in (7.2). This is not the case for other dissimilarity measures, e.g., the squared L_2-distance between the densities $\|p - q\|^2_{L^2(d\mu)}$ which is a popular loss function for density estimation in statistics literature.

The following are common f-divergences:

- **Kullback-Leibler (KL) divergence:** We recover the usual $D(P\|Q)$ in Lecture 2 by taking $f(x) = x \log x$.
- **Total variation:** $f(x) = \frac{1}{2}|x - 1|$,

$$\mathrm{TV}(P, Q) \triangleq \frac{1}{2}\mathbb{E}_Q\left[\left|\frac{dP}{dQ} - 1\right|\right] = \frac{1}{2}\int |dP - dQ|.$$

  Moreover, $\mathrm{TV}(\cdot, \cdot)$ is a metric on the space of probability distributions.

- **χ^2-divergence**: $f(x) = (x - 1)^2$,

$$\chi^2(P\|Q) \triangleq \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^2\right] = \int \frac{(dP - dQ)^2}{dQ} = \int \frac{dP^2}{dQ} - 1. \tag{7.3}$$

  Note that we can also choose $f(x) = x^2 - 1$. Indeed, f's differing by a linear term lead to the same f-divergence, cf. Proposition 7.1.

- **Squared Hellinger distance:** $f(x) = (1 - \sqrt{x})^2$,

$$H^2(P, Q) \triangleq \mathbb{E}_Q\left[\left(1 - \sqrt{\frac{dP}{dQ}}\right)^2\right] = \int \left(\sqrt{dP} - \sqrt{dQ}\right)^2 = 2 - 2\int \sqrt{dP dQ}. \tag{7.4}$$

  Note that $H(P, Q) = \sqrt{H^2(P, Q)}$ defines a metric on the space of probability distributions (indeed, the triangle inequality follows from that of $L_2(\mu)$ for a common dominating measure).

- **Le Cam distance** [LC86, p. 47]: $f(x) = \frac{1-x}{2x+2}$,

$$\mathrm{LC}(P\|Q) = \frac{1}{2}\int \frac{(dP - dQ)^2}{dP + dQ}. \tag{7.5}$$

  Moreover, $\sqrt{\mathrm{LC}(P\|Q)}$ is a metric on the space of probability distributions [ES03].

- **Jensen-Shannon divergence**: $f(x) = x \log \frac{2x}{x+1} + \log \frac{2}{x+1}$,

$$\mathrm{JS}(P,Q) = D\Big(P\Big\|\frac{P+Q}{2}\Big) + D\Big(Q\Big\|\frac{P+Q}{2}\Big).$$

  Moreover, $\sqrt{\mathrm{JS}(P\|Q)}$ is a metric on the space of probability distributions [ES03].

Remark 7.3. If $D_f(P\|Q)$ is an f-divergence, then it is easy to verify that $D_f(\lambda P + \bar{\lambda}Q\|Q)$ and $D_f(P\|\lambda P + \bar{\lambda}Q)$ are f-divergences for all $\lambda \in [0,1]$. In particular, $D_f(Q\|P) = D_{\tilde{f}}(P\|Q)$ with $\tilde{f}(x) \triangleq xf(\frac{1}{x})$.

We start summarizing some formal observations about the f-divergences

Proposition 7.1 (Basic properties)**.** *The following hold:*

1. $D_{f_1+f_2}(P\|Q) = D_{f_1}(P\|Q) + D_{f_2}(P\|Q)$.
2. $D_f(P\|P) = 0$.
3. $D_f(P\|Q) = 0$ *for all* $P \neq Q$ *iff* $f(x) = c(x-1)$ *for some* c. *For any other* f *we have* $D_f(P\|Q) = f(0) + f'(\infty) > 0$ *for* $P \perp Q$.
4. *If* $P_{X,Y} = P_X P_{Y|X}$ *and* $Q_{X,Y} = P_X Q_{Y|X}$ *then the function* $D_f(P_{Y|X=x}\|Q_{Y|X=x})$ *is* $\mathcal{X}$*-measurable and*

$$D_f(P_{X,Y}\|Q_{X,Y}) = \int_{\mathcal{X}} dP_X(x) D_f(P_{Y|X=x}\|Q_{Y|X=x}) \triangleq D_f(P_{Y|X}\|Q_{Y|X}|P_X)\,. \tag{7.6}$$

5. *If* $P_{X,Y} = P_X P_{Y|X}$ *and* $Q_{X,Y} = Q_X P_{Y|X}$ *then*

$$D_f(P_{X,Y}\|Q_{X,Y}) = D_f(P_X\|Q_X)\,. \tag{7.7}$$

6. *Let* $f_1(x) = f(x) + c(x-1)$, *then*

$$D_{f_1}(P\|Q) = D_f(P\|Q) \qquad \forall P, Q\,.$$

   In particular, we can always assume that $f \geq 0$ *and (if* f *is differentiable at 1) that* $f'(1) = 0$.

Proof. The first and second are clear. For the third property, verify explicitly that $D_f(P\|Q) = 0$ for $f = c(x-1)$. Next consider general f and observe that for $P \perp Q$, by definition we have

$$D_f(P\|Q) = f(0) + f'(\infty), \tag{7.8}$$

which is well-defined (i.e., $\infty - \infty$ is not possible) since by convexity $f(0) > -\infty$ and $f'(\infty) > -\infty$. So all we need to verify is that $f(0) + f'(\infty) = 0$ if and only if $f = c(x-1)$ for some $c \in \mathbb{R}$. Indeed, since $f(1) = 0$, the convexity of f implies that $x \mapsto g(x) \triangleq \frac{f(x)}{x-1}$ is non-decreasing. By assumption, we have $g(0+) = g(\infty)$ and hence $g(x)$ is a constant on $x > 0$, as desired.

The next two properties are easy to verify. By Assumption A1 in Section 3.4*, there exist jointly measurable functions p and q such that $dP_{Y|X=x} = p(y|x)d\mu_2$ and $Q_{Y|X} = q(y|x)d\mu_2$ for

some positive measure μ_2 on $\mathcal{Y}$. We can then take μ in (7.2) to be $\mu = P_X \times \mu_2$ which gives $dP_{X,Y} = p(y|x)d\mu$ and $dQ_{X,Y} = q(y|x)d\mu$ and thus

$$D_f(P_{X,Y}\|Q_{X,Y}) = \int_{\mathcal{X}} dP_X \int_{\{y:q(y|x)>0\}} d\mu_2\, q(y|x) f\left(\frac{p(y|x)}{q(y|x)}\right) + f'(\infty)\int_{\mathcal{X}} dP_X \int_{\{y:q(y|x)=0\}} d\mu_2\, p(y|x)$$

$$\overset{(7.2)}{=} \int_{\mathcal{X}} dP_X \underbrace{\left\{\int_{\{y:q(y|x)>0\}} d\mu_2\, q(y|x) f\left(\frac{p(y|x)}{q(y|x)}\right) + f'(\infty)\int_{\{y:q(y|x)=0\}} d\mu_2\, p(y|x)\right\}}_{D_f(P_{Y|X=x}\|Q_{Y|X=x})}$$

which is the desired (7.6).

The fifth property is verified similarly to the fourth. The sixth follows from the first and the third. Note also that reducing to $f \geq 0$ is done by taking $c = f'(1)$ (or any subdifferential at $x = 1$ if f is not differentiable). □

7.2 Data-processing inequality; approximation by finite partitions

Theorem 7.1 (Monotonicity)**.**

$$D_f(P_{X,Y}\|Q_{X,Y}) \geq D_f(P_X\|Q_X)\,. \tag{7.9}$$

Proof. Note that in the case $P_{X,Y} \ll Q_{X,Y}$ (and thus $P_X \ll Q_X$), the proof is a simple application of Jensen's inequality to definition (7.1):

$$\begin{aligned} D_f(P_{X,Y}\|Q_{X,Y}) &= \mathbb{E}_{X\sim Q_X}\,\mathbb{E}_{Y\sim Q_{Y|X}}\left[f\left(\frac{P_{Y|X}P_X}{Q_{Y|X}Q_X}\right)\right] \\ &\geq \mathbb{E}_{X\sim Q_X}\left[f\left(\mathbb{E}_{Y\sim Q_{Y|X}}\left[\frac{P_{Y|X}P_X}{Q_{Y|X}Q_X}\right]\right)\right] \\ &= \mathbb{E}_{X\sim Q_X}\left[f\left(\frac{P_X}{Q_X}\right)\right]\,. \end{aligned}$$

To prove the general case we need to be more careful. By Assumptions A1-A2 in Section 3.4* we may assume that there are functions p_1, p_2, q_1, q_2, positive measures μ_1, μ_2 on $\mathcal{X}$ and $\mathcal{Y}$, respectively, so that

$$dP_{XY} = p_1(x)p_2(y|x)d(\mu_1\times\mu_2), \quad dQ_{XY} = q_1(x)q_2(y|x)d(\mu_1\times\mu_2)$$

and $dP_{Y|X=x} = p_2(y|x)d\mu_2$, $dQ_{Y|X=x} = q_2(y|x)d\mu_2$. We also denote $p(x,y) = p_1(x)p_2(y|x)$, $q(x,y) = q_1(x)q_2(y|x)$ and $\mu = \mu_1\times\mu_2$.

Fix $t > 0$ and consider a supporting line to f at t with slope μ, so that

$$f(u) \geq f(t) + \mu(t-u)\,, \qquad \forall u \geq 0\,.$$

Thus, $f'(\infty) \geq \mu$ and taking $u = \lambda t$ for any $\lambda \in [0,1]$ we have shown:

$$f(\lambda t) + \bar{\lambda} t f'(\infty) \geq f(t)\,, \qquad \forall t > 0, \lambda \in [0,1]\,. \tag{7.10}$$

Note that we need to exclude the case of $t = 0$ since $f(0) = \infty$ is possible.

To rule out the latter possibility, suppose that indeed $f(0) = \infty$. If $Q_X[p_1(X) = 0] > 0$ then we also have $Q_{X,Y}[p_1(X)p_2(Y|X) = 0] > 0$. Consequently, both $D_f(P_X\|Q_X) = \infty$ and

$D_f(P_{X,Y}\|Q_{X,Y}) = \infty$. Thus, from now on we assume that either $f(0) < \infty$ (in which case (7.10) also holds with $t = 0$), or that $Q_X[p_1(X) = 0] = 0$.

Next, fix some x with $q_1(x) > 0$ and consider the chain

$$\begin{aligned}
&\int_{\{y:q_2(y|x)>0\}} d\mu_2\, q_2(y|x) f\left(\frac{p_1(x)p_2(y|x)}{q_1(x)q_2(y|x)}\right) + \frac{p_1(x)}{q_1(x)} P_{Y|X=x}[q_2(Y|x) = 0] f'(\infty) \\
&\overset{\text{(a)}}{\geq} f\left(\frac{p_1(x)}{q_1(x)} P_{Y|X=x}[q_2(Y|x) > 0]\right) + \frac{p_1(x)}{q_1(x)} P_{Y|X=x}[q_2(Y|x) = 0] f'(\infty) \\
&\overset{\text{(b)}}{\geq} f\left(\frac{p_1(x)}{q_1(x)}\right)
\end{aligned}$$

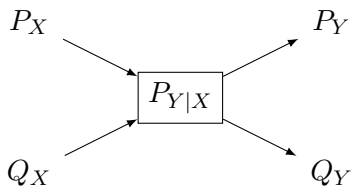
where (a) is by Jensen's inequality and the convexity of f, and (b) by taking $t = \frac{p_1(x)}{q_1(x)}$ and $\lambda = P_{Y|X=x}[q_2(Y|x) > 0]$ in (7.10). Now multiplying the obtained inequality by $q_1(x)$ and integrating over $\{x : q_1(x) > 0\}$ we get

$$\int_{\{q>0\}} d\mu\, q(x,y) f\left(\frac{p(x,y)}{q(x,y)}\right) + P_{X,Y}[q_1(X) > 0, q_2(Y|X) = 0] \geq \int_{\{q_1>0\}} d\mu_1\, q_1(x) f\left(\frac{p_1(x)}{q_1(x)}\right).$$

Adding $f'(\infty)P_X[q_1(X) = 0]$ to both sides we obtain (7.9) since both sides evaluate to definition (7.2). □

The following is the main result of this section.

Theorem 7.2 (Data processing)**.** *Consider a channel that produces Y given X based on the conditional law $P_{Y|X}$ (shown below).*



Let P_Y (resp. Q_Y) denote the distribution of Y when X is distributed as P_X (resp. Q_X). For any f-divergence $D_f(\cdot\|\cdot)$,

$$D_f(P_Y\|Q_Y) \leq D_f(P_X\|Q_X).$$

Proof. This follows from the monotonicity (7.9) and (7.7). □

Next we discuss some of the more useful properties of f-divergence that parallel those of KL divergence in Theorem 2.5:

Theorem 7.3 (Properties of f-divergences)**.**

- ***Non-negativity**: $D_f(P\|Q) \geq 0$. If f is strictly convex[1] at 1, then $D_f(P\|Q) = 0$ if and only if $P = Q$.*

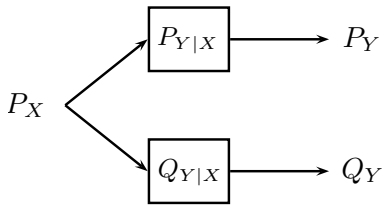
[1]By strict convexity at 1, we mean for all $s, t \in [0, \infty)$ and $\alpha \in (0, 1)$ such that $\alpha s + \bar{\alpha} t = 1$, we have $\alpha f(s) + (1 - \alpha) f(t) > f(1)$.

- ***Joint convexity***: *$(P, Q) \mapsto D_f(P\|Q)$ is a jointly convex function. Consequently, $P \mapsto D_f(P\|Q)$ and $Q \mapsto D_f(P\|Q)$ are also convex.*

- ***Conditioning increases f-divergence***: *Define the conditional f-divergence (similar to Definition 2.4):*

$$D_f\left(P_{Y|X}\|Q_{Y|X}|P_X\right) \triangleq \mathbb{E}_{X\sim P_X}\left[D_f\left(P_{Y|X}\|Q_{Y|X}\right)\right], \tag{7.11}$$

Let $P_X \xrightarrow{P_{Y|X}} P_Y$ and $P_X \xrightarrow{Q_{Y|X}} Q_Y$, i.e.,



Then

$$D_f\left(P_Y\|Q_Y\right) \le D_f\left(P_{Y|X}\|Q_{Y|X}|P_X\right).$$

Proof. • Non-negativity follows from monotonicity by taking X to be unary. To show strict positivity, suppose for the sake of contradiction that $D_f(P\|Q) = 0$ for some $P \neq Q$. Then there exists some measurable A such that $p = P(A) \neq q = Q(A) > 0$. Applying the data processing inequality (with $Y = \mathbf{1}_{\{X\in A\}}$), we obtain $D_f(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) = 0$. Consider two cases

1. $0 < q < 1$: Then $D_f(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) = qf(\frac{p}{q}) + \bar{q}f(\frac{\bar{p}}{\bar{q}}) = f(1)$;
2. $q = 1$: Then $p < 1$ and $D_f(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) = f(p) + \bar{p}f'(\infty) = 0$, i.e. $f'(\infty) = \frac{f(p)}{p-1}$. Since $x \mapsto \frac{f(x)}{x-1}$ is non-decreasing, we conclude that f is affine on $[p, \infty)$.

Both cases contradict the assumed strict convexity of f at 1.

- Convexity follows from the DPI as in the proof of Theorem 5.1.

- Recall that we defined conditional divergence by (7.11) and hence the inequality follows from the monotonicity. Another way to see the inequality is as result of applying Jensen's inequality to the jointly convex function $D_f(P\|Q)$.

□

Remark 7.4 (Strict convexity)**.** Note that even when f is strictly convex at 1, the map $(P, Q) \mapsto D_f(P\|Q)$ may not be strictly convex (e.g. $\mathrm{TV}(\mathrm{Bern}(p), \mathrm{Bern}(q)) = |p - q|$ is piecewise linear). However, if f is strictly convex everywhere on $\mathbb{R}_+$ then so is D_f. Indeed, if $P \neq Q$, then there exists E such that $P(E) \neq Q(E)$. By the DPI and the strict convexity of f, we have $D_f(P\|Q) \ge D_f(\mathrm{Bern}(P(E))\|\mathrm{Bern}(Q(E))) > 0$.

Remark 7.5. We note that, more generally, we may call functional $\mathcal{D}(P\|Q)$ a "g-divergence", or a generalized dissimilarity measure, if it satisfies the following properties: positivity, monotonicity, data processing inequality (DPI), conditioning increases divergence (CID) and convexity in the pair. As we have seen in the proof of Theorem 5.1 the latter two are exactly equivalent. Furthermore, our proof demonstrated that DPI and CID are both implied by monotonicity. If $\mathcal{D}(P\|P) = 0$ then monotonicity, as in (7.9), also implies positivity by taking X to be unary. Finally, notice

that DPI also implies monotonicity by applying it to the (deterministic) channel $(X, Y) \mapsto X$. Thus, requiring DPI (or monotonicity) for $\mathcal{D}$ automatically implies all the other main properties. We remark also that there exist g-divergences which are not monotone transformations of any f-divergence, cf. [PV10, Section V].

The following convenient property, a counterpart of Theorem 4.7, allows us to reduce any general problem about f-divergences to the problem on finite alphabets. The proof is in Section 7.14*.

Theorem 7.4. *Let P, Q be two probability measures on $\mathcal{X}$ with σ-algebra $\mathcal{F}$. Given a finite $\mathcal{F}$-measurable partitions $\mathcal{E} = \{E_1, \ldots, E_n\}$ define the distribution $P_\mathcal{E}$ on $[n]$ by $P_\mathcal{E}(i) = P[E_i]$ and $Q_\mathcal{E}(i) = Q[E_i]$. Then*

$$D_f(P\|Q) = \sup_\mathcal{E} D_f(P_\mathcal{E}\|Q_\mathcal{E}) \tag{7.12}$$

where the supremum is over all finite $\mathcal{F}$-measurable partitions $\mathcal{E}$.

7.3 Total variation and Hellinger distance in hypothesis testing

Different f-divergences have different operational significance. For example, for hypothesis testing the fundamental limit (minimum total probability of error) of binary hypothesis testing is given by the total variation; for estimation under quadratic loss the Le Cam divergence (7.5) is useful, etc. In this section, our goal is consider the problem of *binary hypothesis testing* and explain the special roles of the total variation and Hellinger distances for this problem. The problem is formulated as follows: given an observation (random variable) X, the goal is to decide whether X is drawn from the distribution P or Q, often phrased in terms of two two competing hypotheses: $H_0 : X \sim P$ versus $H_1 : X \sim Q$. We will undertake a systematic study of this problem in Part III and the extension to composite hypothesis testing in Part VI. For now, let us simply notice that one natural goal could be to find a (possibly randomized) decision function $\phi : \mathcal{X} \to \{0, 1\}$ such that the "total probability of error"

$$P[\phi(X) = 1] + Q[\phi(X) = 0] \tag{7.13}$$

is minimized.

Theorem 7.5. *1. We have the following sup-representations of total variation:*

$$\mathrm{TV}(P, Q) = \sup_E P(E) - Q(E) = \frac{1}{2} \sup_{f \in \mathcal{F}} \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)] \tag{7.14}$$

where the first supremum is over all measurable sets E, and the second is over $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}, \|f\|_\infty \le 1\}$. In particular, the minimal sum of error probabilities in (7.13) is given by

$$\min_\phi \{P[\phi(X) = 1] + Q[\phi(X) = 0]\} = 1 - \mathrm{TV}(P, Q), \tag{7.15}$$

where the minimum is over all decision rules $\phi : \mathcal{X} \to \{0, 1\}$.[2]

2. We have the following inf-representation of TV*. Provided that the diagonal $\{(x, x) : x \in \mathcal{X}\}$ is measurable,*

$$\mathrm{TV}(P, Q) = \inf_{P_{X,Y}} \{\mathbb{P}[X \ne Y] : P_X = P, P_Y = Q\}, \tag{7.16}$$

where the set of joint distribution $P_{X,Y}$ with the property $P_X = P$ and $P_Y = Q$ are called couplings of P and Q.

[2] The extension of (7.15) from from simple to composite hypothesis testing is in (34.1).

Proof. Let p, q, μ be as in Definition 7.1. Then for any $f \in \mathcal{F}$ we have

$$\int f(x)(p(x) - q(x))d\mu \le \int |p(x) - q(x)|d\mu = 2\mathrm{TV}(P, Q)\,,$$

which establishes that second supremum in (7.14) lower bounds TV, and hence (by taking $f(x) = 2 \cdot 1_E(x) - 1$) so does the first. For the other direction, let $E = \{x : p(x) > q(x)\}$ and notice

$$0 = \int (p(x) - q(x))d\mu = \int_E + \int_{E^c} (p(x) - q(x))d\mu\,,$$

implying that $\int_{E^c}(q(x) - p(x))d\mu = \int_E (p(x) - q(x))d\mu$. But the sum of these two integrals precisely equals 2TV, which implies that this choice of E attains equality in (7.14).

For the inf-representation [Str65], we notice that given a coupling $P_{X,Y}$ we have from (7.14) with E as above:

$$\mathrm{TV}(P, Q) = P[E] - Q[E] = P_{X,Y}[X \in E] - P_{X,Y}[Y \in E] \le P_{X,Y}[X \ne Y]\,,$$

showing that the inf-reprensentation is always an upper bound. To show that this bound is tight one constructs X, Y as follows: with probability $\pi \triangleq \int \min(p(x), q(x))d\mu$ we take $X = Y = c$ with c sampled from a distribution with density $r(x) = \frac{1}{\pi}\min(p(x), q(x))$, whereas with probability $1 - \pi$ we take X, Y sampled independently from distributions $p_1(x) = \frac{1}{1-\pi}(p(x) - \min(p(x), q(x)))$ and $q_1(x) = \frac{1}{1-\pi}(q(x) - \min(p(x), q(x)))$. The result follows upon verifying that this $P_{X,Y}$ indeed defines a coupling of P to Q and applying the identity

$$\mathrm{TV}(P, Q) = 1 - \int \min(p(x), q(x))d\mu\,.$$

□

Remark 7.6 (Variational representation)**.** The sup-representation (7.14) of the total variation will be extended to general f-divergences in Section 7.13. In turn, the inf-representation (7.16) has no analogs for other f-divergences, with the notable exception of Marton's d_2, see (**??**). Distances satisfying inf-representations are often called Wasserstein distances, and hence we may think of TV as the Wasserstein distance with respect to Hamming distance $d(x, x') = 1\{x \ne x'\}$ on $\mathcal{X}$. The benefit of variational representations is that choosing a particular coupling in (7.16) gives an upper bound on $\mathrm{TV}(P, Q)$, and choosing a particular f in (7.14) yields a lower bound.

Of particular relevance is the special case of multiple-sample testing, where the data $X = (X_1, \ldots, X_n)$ are i.i.d. drawn from either P or Q. In other words, the goal is to test

$$H_0 : X \sim P^{\otimes n} \quad \text{vs} \quad H_1 : X \sim Q^{\otimes n}.$$

By Theorem 7.5, the optimal total probability of error is given by $1 - \mathrm{TV}(P^{\otimes n}, Q^{\otimes n})$. By the data processing inequality, $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n})$ is a non-decreasing sequence in n (and bounded by 1 by definition) and hence converges. One would expect that as $n \to \infty$, $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n})$ converges to 1 and consequently, the probability of error in the hypothesis test vanishes. It turns out that for fixed distributions $P \ne Q$, large deviation theory (see Lecture 15) shows that $\mathrm{TV}(P^{\otimes n}, Q^{\otimes n})$ indeed converges to one as $n \to \infty$ and, in fact, exponentially fast:

$$\mathrm{TV}(P^{\otimes n}, Q^{\otimes n}) = 1 - \exp(-nC(P, Q) + o(n)), \tag{7.17}$$

where the exponent $C(P,Q) > 0$ is known as the *Chernoff Information* of P and Q. However, as frequently encountered in high-dimensional statistical problems, if the distributions $P = P_n$ and $Q = Q_n$ depend on n, then the large-deviation asymptotics in (7.17) can no longer be directly applied. Since computing the total variation between two n-fold product distributions is typically difficult, understanding how a more computationally tractable f-divergence is related to the total variation may give insight on its behavior. It turns out Hellinger distance is precisely suited for this task.

Shortly, we will show the following relation between TV and the Hellinger divergence:

$$\frac{1}{2}H^2(P,Q) \le \mathrm{TV}(P,Q) \le H(P,Q)\sqrt{1-\frac{H^2(P,Q)}{4}} \le 1. \tag{7.18}$$

Direct consequences of the bound (7.18) are:

- $H^2(P,Q) = 2$, if and only if $\mathrm{TV}(P,Q) = 1$. In this case, the probability of error is zero since essentially P and Q have disjoint supports.
- $H^2(P,Q) = 0$ if and only if $\mathrm{TV}(P,Q) = 0$. In this case, the smallest total probability of error is one, meaning the best thing to do is to flip a coin.
- Hellinger consistency is equivalent to TV consistency: we have

$$H^2(P_n,Q_n) \to 0 \iff \mathrm{TV}(P_n,Q_n) \to 0 \tag{7.19}$$
$$H^2(P_n,Q_n) \to 2 \iff \mathrm{TV}(P_n,Q_n) \to 1; \tag{7.20}$$

  however, the speed of convergence need not be the same.

Theorem 7.6. *For any sequence of distributions P_n and Q_n, as $n \to \infty$,*

$$\mathrm{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0 \iff H^2(P_n,Q_n) = o\left(\frac{1}{n}\right)$$
$$\mathrm{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1 \iff H^2(P_n,Q_n) = \omega\left(\frac{1}{n}\right)$$

Proof. For convenience, let $X_1, X_2, \ldots X_n \overset{\text{i.i.d.}}{\sim} Q_n$. Then

$$\begin{aligned} H^2(P_n^{\otimes n}, Q_n^{\otimes n}) &= 2 - 2\mathbb{E}\left[\sqrt{\prod_{i=1}^n \frac{P_n}{Q_n}(X_i)}\right] \\ &= 2 - 2\prod_{i=1}^n \mathbb{E}\left[\sqrt{\frac{P_n}{Q_n}(X_i)}\right] = 2 - 2\left(\mathbb{E}\left[\sqrt{\frac{P_n}{Q_n}}\right]\right)^n \\ &= 2 - 2\left(1 - \frac{1}{2}H^2(P_n,Q_n)\right)^n. \end{aligned} \tag{7.21}$$

We now use (7.21) to conclude the proof. Recall from (7.19) that $\mathrm{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \to 0$, which happens precisely when $H^2(P_n,Q_n) = o(\frac{1}{n})$. Similarly, by (7.20), $\mathrm{TV}(P_n^{\otimes n}, Q_n^{\otimes n}) \to 1$ if and only if $H^2(P_n^{\otimes n}, Q_n^{\otimes n}) \to 2$, which is further equivalent to $H^2(P_n,Q_n) = \omega(\frac{1}{n})$. □

Remark 7.7. Property (7.21) is known as *tensorization*. While other f-divergences also satisfy tensorization, see Section 7.12, the H^2 has the advantage of a sandwich bound (7.18) making it the most convenient tool for checking asymptotic testability of hypotheses.

7.4 Inequalities between f-divergences and joint range

In this section we study the relationship, in particular, inequalities, between f-divergences. To gain some intuition, we start with the ad hoc approach by proving the *Pinsker's inequality*, which bounds total variation from above in terms of the KL divergence.

Theorem 7.7 (Pinsker's inequality)**.**

$$D(P\|Q) \geq (2\log e)\mathrm{TV}^2(P,Q). \tag{7.22}$$

Proof. It suffices to consider the natural logarithm for the KL divergence. First we show that, by the data processing inequality, it suffices to prove the result for Bernoulli distributions. For any event E, let $Y = \mathbf{1}_{\{X\in E\}}$ which is Bernoulli distributed with parameter $P(E)$ or $Q(E)$. By the DPI, $D(P\|Q) \geq d(P(E)\|Q(E))$. If Pinsker's inequality holds for all Bernoulli distributions, we have

$$\sqrt{\frac{1}{2}D(P\|Q)} \geq \mathrm{TV}(\mathrm{Bern}(P(E)), \mathrm{Bern}(Q(E)) = |P(E) - Q(E)|$$

Taking the supremum over E gives $\sqrt{\frac{1}{2}D(P\|Q)} \geq \sup_E |P(E) - Q(E)| = \mathrm{TV}(P,Q)$, in view of Theorem 7.5.

The binary case follows easily from a second-order Taylor expansion (with integral remainder form) of $p \mapsto d(p\|q)$:

$$d(p\|q) = \int_q^p \frac{p-t}{t(1-t)}dt \geq 4\int_q^p (p-t)dt = 2(p-q)^2$$

and $\mathrm{TV}(\mathrm{Bern}(p), \mathrm{Bern}(q)) = |p-q|$. □

Pinsker's inequality is sharp in the sense that the constant $(2\log e)$ in (7.22) is not improvable, i.e., there exist $\{P_n, Q_n\}$ such that $\frac{\mathrm{LHS}}{\mathrm{RHS}} \to 2$ as $n \to \infty$. (This is best seen by inspecting the local quadratic behavior in Proposition 5.2.) Nevertheless, this does not mean that the inequality (7.22) is not improvable, as the RHS can be replaced by some other function of $\mathrm{TV}(P,Q)$. Indeed, several such improvements of Pinsker's inequality are known. But what is the best inequality? In addition, another natural question is the reverse inequality: can we upper-bound $D(P\|Q)$ in terms of $\mathrm{TV}(P,Q)$? Settling these questions rests on characterizing the *joint range* (the set of possible values) of a given pair f-divergences. This systematic approach to comparing f-divergences (as opposed to the ad hoc proof of Theorem 7.7 we presented above) is the subject of this section.

Definition 7.2 (Joint range)**.** Consider two f-divergences $D_f(P\|Q)$ and $D_g(P\|Q)$. Their joint range is a subset of $[0,\infty]^2$ defined by

$$\mathcal{R} \triangleq \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ are probability measures on some measurable space}\}.$$

In addition, the joint range over all k-ary distributions is defined as

$$\mathcal{R}_k \triangleq \{(D_f(P\|Q), D_g(P\|Q)) : P, Q \text{ are probability measures on } [k]\}.$$

As an example, Fig. 7.1 gives the joint range $\mathcal{R}$ between the KL divergence and the total variation. By definition, the lower boundary of the region $\mathcal{R}$ gives the optimal refinement of Pinsker's inequality:

$$D(P\|Q) \geq F(\mathrm{TV}(P,Q)), \quad F(\epsilon) \triangleq \inf_{(P,Q):\mathrm{TV}(P,Q)=\epsilon} D(P\|Q) = \inf\{s : (\epsilon, s) \in \mathcal{R}\}.$$

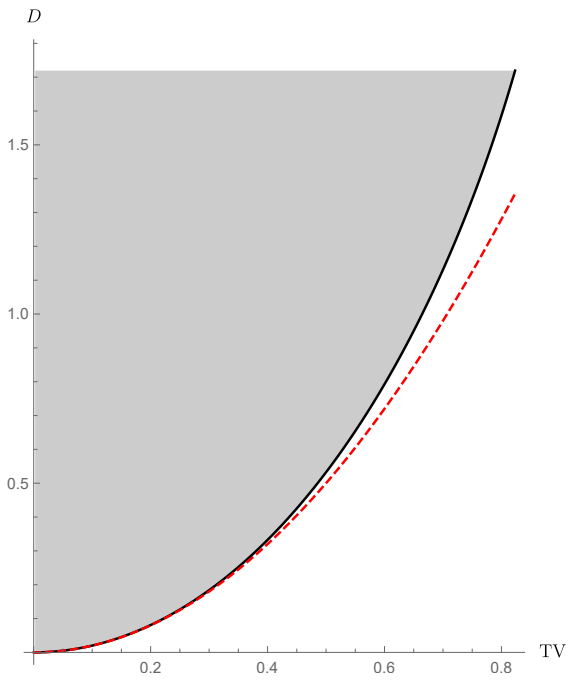


Figure 7.1: Joint range of TV and KL divergence. The dashed line is the quadratic lower bound given by Pinsker's inequality (7.22).

Also from Fig. 7.1 we see that it is impossible to bound $D(P\|Q)$ from above in terms of $\mathrm{TV}(P,Q)$ due to the lack of upper boundary.

The joint range $\mathcal{R}$ may appear difficult to characterize since we need to consider P, Q over all measurable spaces; on the other hand, the region $\mathcal{R}_k$ for small k is easy to obtain (at least numerically). Revisiting the proof of Pinkser's inequality in Theorem 7.7, we see that the key step is the reduction to Bernoulli distributions. It is natural to ask: to obtain full joint range is it possible to reduce to the binary case? It turns out that it is always sufficient to consider quaternary distributions, or the convex hull of that of binary distributions.

Theorem 7.8 (Harremoës-Vajda [HV11])**.**

$$\mathcal{R} = \mathrm{co}(\mathcal{R}_2) = \mathcal{R}_4.$$

where co *denotes the convex hull with a natural extension of convex operations to* $[0,\infty]^2$*.*

We will rely on the following famous result from convex analysis (cf. e.g. [Egg58, Chapter 2, Theorem 18]).

Lemma 7.1 (Fenchel-Eggleston-Carathéodory theorem)**.** *Let* $S \subseteq \mathbb{R}^d$ *and* $x \in \mathrm{co}(S)$*. Then there exists a set of* $d+1$ *points* $S' = \{x_1, x_2, \ldots, x_{d+1}\} \in S$ *such that* $x \in \mathrm{co}(S')$*. If* S *has at most* d *connected components, then* d *points are enough.*

Proof. Our proof will consist of three claims:

- *Claim 1:* $\mathrm{co}(\mathcal{R}_2) \subset \mathcal{R}_4$
- *Claim 2:* $\mathcal{R}_k \subset \mathrm{co}(\mathcal{R}_2)$

- *Claim 3:* $\mathcal{R} = \mathcal{R}_4$

We can see that Claims 1-2 prove the most interesting part: $\bigcup_{k=1}^{\infty} \mathcal{R}_k = \mathrm{co}(\mathcal{R}_2)$. Claim 3 is more technical and its proof can be found in [HV11]. However, we notice that approximation theorem 7.4 allows us to conclude that any point $(d_f, d_g) \in \mathcal{R}$ is a limit of points in $\bigcup_{k=1}^{\infty} \mathcal{R}_k$. For inequalities between D_f and D_g we are only interested in the closure of $\mathcal{R}$, and thus Claims 1-2 are sufficient.

We start with Claim 1. Given any two pairs of distributions (P_0, Q_0) and (P_1, Q_1) on some space $\mathcal{X}$ and given any $\alpha \in [0, 1]$, we construct a random variable $Z = (X, B)$ with $B \sim \mathrm{Bern}(\alpha)$, where $P_{X|B=i} = P_i$ and $Q_{X|B=i} = Q_i$ for $i = 0, 1$. Then by (7.6) we get

$$D_f(P_{X,B} \| Q_{X,B}) = \bar{\alpha} D_f(P_0 \| Q_0) + \alpha D_f(P_1 \| Q_1)\,,$$

and similarly for the D_g. Thus, $\mathcal{R}$ is convex. Next, notice that $\mathcal{R}_2$ is the image of $[0, 1]^2$ and

$$\mathcal{R}_2 = \tilde{\mathcal{R}}_2 \cup \{(pf'(\infty), pg'(\infty)) : p \in (0, 1]\} \cup \{(qf(0), qg(0)) : q \in (0, 1]\}\,,$$

where $\tilde{\mathcal{R}}_2$ is the image of $(0, 1)^2$ of the continuous map

$$(p, q) \mapsto \Big(D_f(\mathrm{Bern}(p) \| \mathrm{Bern}(q)), D_g(\mathrm{Bern}(p) \| \mathrm{Bern}(q)) \Big)\,.$$

Since $(0, 0) \in \tilde{\mathcal{R}}_2$, we can see that regardless of which $f(0), f'(\infty), g(0), g'(\infty)$ are infinite, the set $\mathcal{R}_2 \cap \mathbb{R}^2$ is connected. Thus, by Lemma 7.1 any point in $\mathrm{co}(\mathcal{R}_2)$ is a combination of two points in $\mathcal{R}_2$. By the argument above, then, $\mathcal{R}_2 \subset \mathcal{R}_4$.

Next, we prove Claim 2. Fix P, Q on $[k]$ and denote their PMFs (p_j) and (q_j), respectively. Note that without changing either $D_f(P\|Q)$ or $D_g(P\|Q)$ (but perhaps, by increasing k by 1), we can make $q_j > 0$ for $j > 1$ and $q_1 = 0$, which we thus assume. Denote $\phi_j = \frac{p_j}{q_j}$ for $j > 1$ and consider the set

$$\mathcal{S} = \{\tilde{Q} = (\tilde{q}_j)_{j \in [k]} : \tilde{q}_j \geq 0, \sum \tilde{q}_j = 1, \tilde{q}_1 = 0, \sum_{j=2}^{k} \tilde{q}_j \phi_j \leq 1\}\,.$$

We also define a subset $\mathcal{S}_e \subset \mathcal{S}$ consisting of points $\tilde{Q}$ of two types:

1. $\tilde{q}_j = 1$ for some $j \geq 2$ and $\phi_j \leq 1$.
2. $\tilde{q}_{j_1} + \tilde{q}_{j_2} = 1$ for some $j_1, j_2 \geq 2$ and $\tilde{q}_{j_1}\phi_{j_1} + \tilde{q}_{j_2}\phi_{j_2} = 1$.

It can be seen that $\mathcal{S}_e$ are precisely all the extreme points of $\mathcal{S}$. Indeed, any $\tilde{Q} \in \mathcal{S}$ with $\sum_{j \geq 2} \tilde{q}_j \phi_j < 1$ with more than one non-zero atom cannot be extremal (since there is only one active linear constraint $\sum_j \tilde{q}_j = 1$). Similarly, $\tilde{Q}$ with $\sum_{j \geq 2} \tilde{q}_j \phi_j = 1$ can only be extremal if it has one or two non-zero atoms.

We next claim that any point in $\mathcal{S}$ can be written as a convex combination of finitely many points in $\mathcal{S}_e$. This can be seen as follows. First, we can view $\mathcal{S}$ and $\mathcal{S}_e$ as subsets of $\mathbb{R}^{k-1}$. $\mathcal{S}$ is clearly closed and convex. By a theorem of Krein-Milman $\mathcal{S}$ coincides with the closure of the convex hull of its extreme points. However, $\mathrm{co}(\mathcal{S}_e)$ is compact (hence closed) since by Lemma 7.1 it is a continuous image of a product of k copies of $\mathcal{S}_e$ and a probability simplex on $[k]$. Thus $\mathcal{S} = \mathrm{co}(\mathcal{S}_e)$ and, in particular, there are probability weights $\{\alpha_i, i \in [m]\}$ and extreme points $\tilde{Q}_i \in \mathcal{S}_e$ so that

$$Q = \sum_{i=1}^{m} \alpha_i \tilde{Q}_i\,. \tag{7.23}$$

Next, to each $\tilde{Q}$ we associate $\tilde{P} = (\tilde{p}_j)_{j\in[k]}$ as follows:

$$\tilde{p}_j = \begin{cases} \phi_j \tilde{q}_j, & j \in \{2, \dots, k\}, \\ 1 - \sum_{j=2}^{k} \phi_j \tilde{q}_j, & j = 1 \end{cases}$$

We then have that

$$\tilde{Q} \mapsto D_f(\tilde{P}\|\tilde{Q}) = \sum_{j\geq 2} \tilde{q}_j f(\phi_j) + f'(\infty)\tilde{p}_1$$

affinely maps $\mathcal{S}$ to $[0, \infty]$ (note that $f(0)$ or $f'(\infty)$ can equal ∞). In particular, if we denote $\tilde{P}_i = \tilde{P}(\tilde{Q}_i)$ corresponding to $\tilde{Q}_i$ in decomposition (7.23), we get

$$D_f(P\|Q) = \sum_{i=1}^{m} \alpha_i D_f(\tilde{P}_i\|\tilde{Q}_i),$$

and similarly for $D_g(P\|Q)$. We are left to show that $(\tilde{P}_i, \tilde{Q}_i)$ are supported on at most two points. Indeed, for $\tilde{Q} \in \mathcal{S}_e$ the set $\{j \in [k] : \tilde{q}_j > 0 \text{ or } \tilde{p}_j > 0\}$ has cardinality at most two (for the second type extremal points we notice $\tilde{p}_{j_1} + \tilde{p}_{j_2} = 1$ implying $\tilde{p}_1 = 0$).

From (7.23) we were able to represent an element of $\tilde{\mathcal{R}}_k$ as convex combination of k elements of $\tilde{\mathcal{R}}_2$, concluding the proof of Claim 2. □

7.5 Examples of computing joint range

7.5.1 Hellinger distance versus total variation

The joint range $\mathcal{R}_2$ of H^2 and TV over binary distributions is simply:

$$\mathcal{R}_2 = \left\{(2(1 - \sqrt{pq} - \sqrt{\bar{p}\bar{q}}), |p - q|) : 0 \leq p \leq 1, 0 \leq q \leq 1\right\}.$$

shown as non-convex grey region in Fig. 7.2. By Theorem 7.8, their full joint range $\mathcal{R}$ is the convex hull of $\mathcal{R}_2$, which turns out to be exactly described by the sandwich bound (7.18) shown earlier in Section 7.3. This means that (7.18) is not improvable. Indeed, with t ranging from 0 to 1,

- the upper boundary is achieved by $P = \text{Bern}(\frac{1+t}{2}), Q = \text{Bern}(\frac{1-t}{2})$,
- the lower boundary is achieved by $P = (1 - t, t, 0), Q = (1 - t, 0, t)$.

7.5.2 KL divergence versus total variation

The joint range between KL and TV was previously shown in Fig. 7.1. Although there is no known close-form expression, the following parametric formula of the lower boundary (see Fig. 7.1) is known [FHT03, Theorem 1]:

$$\begin{cases} \text{TV}_t = \frac{1}{2} t \left(1 - \left(\coth(t) - \frac{1}{t}\right)^2\right) \\ D_t = -t^2 \text{csch}^2(t) + t \coth(t) + \log(t\text{csch}(t)) \end{cases}, \quad t \geq 0. \tag{7.24}$$

where we take the natural logarithm. Here is a corollary (weaker bound) due to [Vaj70]:

$$D(P\|Q) \geq \log \frac{1 + \text{TV}(P, Q)}{1 - \text{TV}(P, Q)} - \frac{2\text{TV}(P, Q)}{1 + \text{TV}(P, Q)}. \tag{7.25}$$

Both bounds are stronger than Pinsker's inequality (7.22). Note the following consequences:

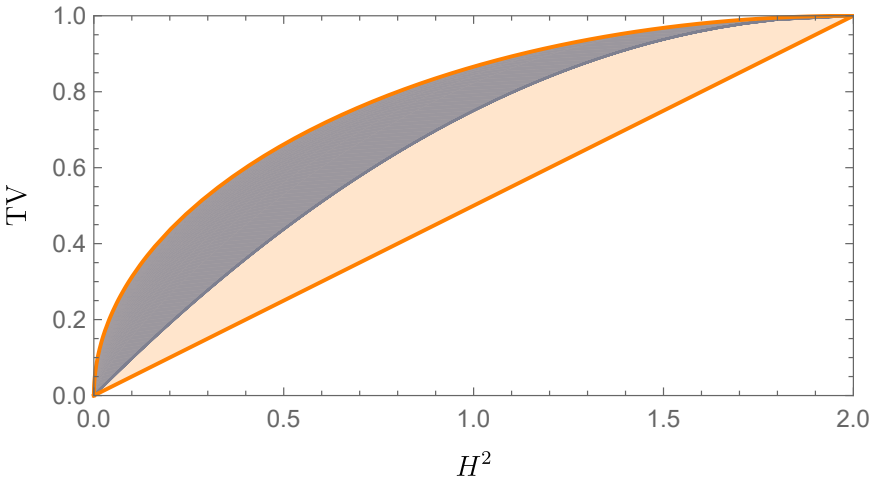


Figure 7.2: The joint range $\mathcal{R}$ of TV and H^2 is characterized by (7.18), which is the convex hull of the grey region $\mathcal{R}_2$.

- $D \to 0 \Rightarrow \mathrm{TV} \to 0$, which can be deduced from Pinsker's inequality;
- $\mathrm{TV} \to 1 \Rightarrow D \to \infty$ and hence $D = O(1)$ implies that TV is bounded away from one. This can be obtained from (7.24) or (7.25), but not Pinsker's inequality.

7.5.3 χ^2 and total variation

Proposition 7.2. *We have the following bound*

$$\chi^2(P\|Q) \geq f(\mathrm{TV}(P,Q)), \qquad f(t) = \begin{cases} 4t^2\,. & t \leq \frac{1}{2} \\ \frac{t}{1-t} & t \geq \frac{1}{2}. \end{cases}, \tag{7.26}$$

where function f is a a convex increasing bijection of $[0,1)$ onto $[0,\infty)$. Furthermore, for every $s \geq f(t)$ there exists a pair of distributions such that $\chi^2(P\|Q) = s$ and $\mathrm{TV}(P,Q) = t$.

Proof. We claim that the binary joint range is convex. Indeed,

$$\mathrm{TV}(\mathrm{Bern}(p), \mathrm{Bern}(q)) = |p-q| \triangleq t, \quad \chi^2(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) = \frac{(p-q)^2}{q(1-q)} = \frac{t^2}{q(1-q)}.$$

Given $|p-q| = t$, let us determine the possible range of $q(1-q)$. The smalles value of $q(1-q)$ is always 0 by choosing $p = t, q = 0$. The largest value will be $1/4$ if $t \leq 1/2$ (by choosing $p = 1/2 - t$, $q = 1/2$). If $t > 1/2$ then we can at most get $t(1-t)$ (by setting $p = 0$ and $q = t$). Thus we get $\chi^2(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) \geq f(|p-q|)$ as claimed. Convexity follows since derivative of f is monotoinically increasing. □

7.6 A selection of inequalities between various divergences

This section presents a collection of useful inequalities. For a more complete treatment, consider [SV16] and [Tsy09, Sec. 2.4]. Most of these inequalities are joint ranges, which means they are tight.

- KL vs TV: see (7.24). There is partial comparison in the other direction ("reverse Pinsker", cf. [SV16, Section VI]):

$$D(P\|Q) \le \log\left(1 + \frac{2}{Q_{min}}\mathrm{TV}(P,Q)^2\right) \le \frac{2\log e}{Q_{min}}\mathrm{TV}(P,Q)^2\,, \qquad Q_{min} = \min_x Q(x)$$

- KL vs Hellinger:

$$D(P||Q) \ge 2\log\frac{2}{2 - H^2(P,Q)}. \tag{7.27}$$

  There is a partial result in the opposite direction (log-Sobolev inequality for Bonami-Beckner semigroup, cf. [DSC, Theorem A.1]):

$$D(P\|Q) \le \frac{\log(\frac{1}{Q_{min}} - 1)}{1 - 2Q_{min}}\left(1 - (1 - H^2(P,Q))^2\right)\,, \qquad Q_{min} = \min_x Q(x)$$

- KL vs χ^2:

$$0 \le D(P||Q) \le \log(1 + \chi^2(P||Q)) \le \log e \cdot \chi^2(P\|Q)\,. \tag{7.28}$$

  (i.e. no lower-bound on KL in terms of χ^2 is possible).

- TV and Hellinger: see (7.18). Another bound [Gil10]:

$$\mathrm{TV}(P,Q) \le \sqrt{-2\ln\left(1 - \frac{H^2(P,Q)}{2}\right)}$$

- Le Cam and Hellinger [LC86, p. 48]:

$$\frac{1}{2}H^2(P,Q) \le \mathrm{LC}(P,Q) \le H^2(P,Q). \tag{7.29}$$

- Le Cam and Jensen-Shannon [Top00]:

$$\mathrm{LC}(P||Q)\log e \le \mathrm{JS}(P,Q) \le \mathrm{LC}(P||Q)\cdot 2\log 2 \tag{7.30}$$

- χ^2 and TV: The full joint range is given by (7.26). Two simple consequences are:

$$\mathrm{TV}(P,Q) \le \frac{1}{2}\sqrt{\chi^2(P\|Q)}$$
$$\mathrm{TV}(P,Q) \le \max\left\{\frac{1}{2}, \frac{\chi^2(P\|Q)}{1 + \chi^2(P\|Q)}\right\}$$

  where the second is useful for bounding TV away from one.

- JS and TV: the full joint region is given by

$$2d\left(\frac{1 - \mathrm{TV}(P,Q)}{2}\middle\|\frac{1}{2}\right) \le \mathrm{JS}(P,Q) \le \mathrm{TV}(P,Q)\cdot 2\log 2\,. \tag{7.31}$$

  The lower bound is a consequence of Fano's inequality. For the upper bound notice that for $p, q \in [0,1]$ and $|p-q| = \tau$ the maximum of $d(p\|\frac{p+q}{2})$ is attained at $p = 0, q = \tau$ (from the convexity of $d(\cdot\|\cdot)$) and, thus, the binary joint-range is given by $\tau \mapsto d(\tau\|\tau/2) + d(1-\tau\|1-\tau/2)$. Since the latter is convex, its concave envelope is a straightline connecting endpoints at $\tau = 0$ and $\tau = 1$.

7.7 Example: divergences between Gaussians

1. Total variation:

$$\begin{aligned} \mathrm{TV}(\mathcal{N}(0,\sigma^2),\mathcal{N}(\mu,\sigma^2)) &= 2\Phi\left(\frac{|\mu|}{2\sigma}\right)-1 && (7.32)\\ &= \int_{-\frac{|\mu|}{2\sigma}}^{\frac{|\mu|}{2\sigma}} \varphi(x)\mathrm{d}x && (7.33)\\ &= \frac{|\mu|}{\sqrt{2\pi}\sigma}+O(\mu^2), \quad \mu\to 0. && (7.34)\end{aligned}$$

2. Hellinger:

$$H^2(\mathcal{N}(0,\sigma^2)||\mathcal{N}(\mu,\sigma^2)) = 2-2e^{-\frac{\mu^2}{8\sigma^2}} = \frac{\mu^2}{4\sigma^2}+O(\mu^3), \quad \mu\to 0.$$

   More generally,

$$H^2(\mathcal{N}(\mu_1,\Sigma_1)||\mathcal{N}(\mu_2,\Sigma_2)) = 2-2\frac{|\Sigma_1|^{\frac{1}{4}}|\Sigma_2|^{\frac{1}{4}}}{|\bar{\Sigma}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{8}(\mu_1-\mu_2)'\bar{\Sigma}^{-1}(\mu_1-\mu_2)\right\},$$

   where $\bar{\Sigma}=\frac{\Sigma_1+\Sigma_2}{2}$.

3. KL divergence:

$$D(\mathcal{N}(\mu_1,\sigma_1^2)||\mathcal{N}(\mu_2,\sigma_2^2)) = \frac{1}{2}\log\frac{\sigma_2^2}{\sigma_1^2}+\frac{1}{2}\left(\frac{(\mu_1-\mu_2)^2}{\sigma_2^2}+\frac{\sigma_1^2}{\sigma_2^2}-1\right)\log e.$$

   For a more general result see (2.4).

4. χ^2-divergence:

$$\chi^2(\mathcal{N}(\mu,\sigma^2)||\mathcal{N}(0,\sigma^2)) = e^{\frac{\mu^2}{\sigma^2}}-1 = \frac{\mu^2}{\sigma^2}+O(\mu^3), \quad \mu\to 0$$

$$\chi^2(\mathcal{N}(\mu,\sigma^2)||\mathcal{N}(0,1)) = \begin{cases} \frac{e^{\mu^2/(2-\sigma^2)}}{\sigma\sqrt{2-\sigma^2}}-1 & \sigma^2<2 \\ \infty & \sigma^2\geq 2 \end{cases}$$

5. χ^2-divergence for Gaussian mixtures [IS03]:

$$\chi^2(P*\mathcal{N}(0,\Sigma)||\mathcal{N}(0,\Sigma)) = \mathbb{E}[e^{\langle\Sigma^{-1}X,X'\rangle}]-1\,, \qquad X\perp\!\!\!\perp X'\sim P\,.$$

7.8 Mutual information based on f-divergence

Given an f-divergence D_f, we can define the a version of mutual information

$$I_f(X;Y)\triangleq D_f(P_{X,Y}\|P_XP_Y)\,.$$

Theorem 7.9 (Data processing)**.** *For $U\to X\to Y$, we have $I_f(U;Y)\leq I_f(U;X)$.*

Proof. Note that $I_f(U;X,Y)=I_f(U;X)=D_f(P_{U,X}\|P_UP_X)\geq D_f(P_{U,Y}\|P_UP_Y)$, where we applied the data-processing Theorem 7.2 to the (possibly stochastic) map $(U,X)\mapsto(U,Y)$. See also Remark 3.2. □

One often used property of the standard mutual information is the *subadditivity*: If $P_{A,B|X} = P_{A|X}P_{B|X}$ (i.e. A and B are conditionally independent given X), then

$$I(X;A,B) \le I(X;A) + I(X;B). \tag{7.35}$$

However, other notions of f-information have complicated relationship with subadditivity:

1. The f-information corresponding to the χ^2-divergence, $I_{\chi^2}(X;Y) = \chi^2(P_{X,Y}\|P_XP_Y)$ is not subadditive.

2. The f-information corresponding to total-variation $I_{\mathrm{TV}}(X;Y) = \mathrm{TV}(P_{X,Y}, P_XP_Y)$ is not subadditive. Even worse, it can get stuck. For example, take $X \sim \mathrm{Bern}(1/2)$ and $A = \mathrm{BSC}_\delta(X)$, $B = \mathrm{BSC}_\delta(X)$ – two independent observations of X across the BSC. A simple computation shows:
$$I_{\mathrm{TV}}(X;A,B) = I_{\mathrm{TV}}(X;A) = I_{\mathrm{TV}}(X;B)\,.$$
In other words, an additional observation does not improve TV-information at all. This is the main reason for the famous herding effect in economics [Ban92].

3. The symmetric KL-divergence[3] $I_{\mathrm{SKL}}(X;Y) = D(P_{X,Y}\|P_XP_Y) + D(P_XP_Y\|P_{X,Y})$ satisfies, quite amazingly [KF+09], the *additivity property*:
$$I_{\mathrm{SKL}}(X;A,B) = I_{\mathrm{SKL}}(X;A) + I_{\mathrm{SKL}}(X;B) \tag{7.36}$$
To prove this, we first notice the following equivalent expression for I_{SKL}:
$$I_{\mathrm{SKL}}(X;Y) = \sum_{x,x'} P_X(x)P_X(x')D(P_{Y|X=x}\|P_{Y|X=x'})\,. \tag{7.37}$$
From (7.37) we get (7.36) by additivity of $D(P_{A,B|X=x}\|P_{A,B|X=x'})$. To prove (7.37) first consider the obvious identity:
$$\sum_{x,x'} P_X(x)P_X(x')[D(P_Y\|P_{Y|X=x'}) - D(P_Y\|P_{Y|X=x})] = 0$$
which is rewritten as
$$\sum_{x,x'} P_X(x)P_X(x') \sum_y P_Y(y) \log \frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')} = 0\,. \tag{7.38}$$
Next, notice that
$$I_{\mathrm{SKL}}(X;Y) = \sum_{x,y} [P_{X,Y}(x,y) - P_X(x)P_Y(y)] \log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\,.$$
Since the marginals of $P_{X,Y}$ and P_XP_Y coincide, we can replace $\log \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}$ by any $\log \frac{P_{Y|X}(y|x)}{f(y)}$ for any f. We choose $f(y) = P_{Y|X}(y|x')$ to get
$$I_{\mathrm{SKL}}(X;Y) = \sum_{x,y} [P_{X,Y}(x,y) - P_X(x)P_Y(y)] \log \frac{P_{Y|X}(y|x)}{P_{Y|X}(y|x')}\,.$$
Now averaging this over $P_X(x')$ and applying (7.38) to get rid of the second term in $[\cdots]$, we obtain (7.37).

[3] This is the f-information corresponding to the Jeffery divergence $D(P\|Q) + D(Q\|P)$.

7.9 Empirical distribution and χ^2-information

Consider an arbitrary channel $P_{Y|X}$ and some input distribution P_X. Suppose that we have $X_i \overset{\text{i.i.d.}}{\sim} P_X$ for $i = 1, \ldots, n$. Let

$$\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

denote the empirical distribution corresponding to this sample. Let $P_Y = P_{Y|X} \circ P_X$ be the output distribution corresponding to P_X and $P_{Y|X} \circ \hat{P}_n$ be the output distribution corresponding to $\hat{P}_n$ (a random distribution). Note that when $P_{Y|X=x}(\cdot) = \phi(\cdot - x)$, where ϕ is a fixed density, we can think of $P_{Y|X} \circ \hat{P}_n$ as a *kernel density estimator (KDE)*, whose density is $\hat{p}_n(x) = (\phi * \hat{P}_n)(x) \frac{1}{n}\sum_{i=1}^{n} \phi(X_i - x)$. Furthermore, using the fact that $\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n] = P_Y$, we have

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_X)] = D(P_Y \| P_X) + \mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)]\,,$$

where the first term represents the bias of the KDE due to convolution and increases with bandwidth of ϕ, while the second term represents the variability of the KDE and decreases with the bandwidth of ϕ. Surprisingly, the second term is is sharply (within a factor of two) given by the I_{χ^2} information. More exactly, we prove the following result.

Proposition 7.3. *We have*

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] \le \log\left(1 + \frac{1}{n} I_{\chi^2}(X;Y)\right)\,. \tag{7.39}$$

Furthermore,

$$\liminf_{n\to\infty} n\,\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] \ge \frac{\log e}{2} I_{\chi^2}(X;Y)\,. \tag{7.40}$$

In particular, $\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] = O(1/n)$ *if* $I_{\chi^2}(X;Y) < \infty$ *and* $\omega(1/n)$ *otherwise.*

Proof. First, a simple calculation shows that

$$\mathbb{E}[\chi^2(P_{Y|X} \circ \hat{P}_n \| P_Y)] = \frac{1}{n} I_{\chi^2}(X;Y)\,.$$

Then from (7.28) and Jensen's inequality we get (7.39).

To get the lower bound in (7.40), let $\bar{X}$ be drawn uniformly at random from the sample $\{X_1, \ldots, X_n\}$ and let $\bar{Y}$ be the output of the $P_{Y|X}$ channel with input $\bar{X}$. With this definition we have:

$$\mathbb{E}[D(P_{Y|X} \circ \hat{P}_n \| P_Y)] = I(X^n; \bar{Y})\,.$$

Next, apply (6.2) to get

$$I(X^n; \bar{Y}) \ge \sum_{i=1}^{n} I(X_i; \bar{Y}) = n I(X_1; \bar{Y})\,.$$

Finally, notice that

$$I(X_1; \bar{Y}) = D\left(\frac{n-1}{n} P_X P_Y + \frac{1}{n} P_{XY} \middle\| P_X P_Y\right)$$

and apply the local expansion of KL divergence (Proposition 5.2) to get (7.40). □

In the discrete case, by taking $P_{Y|X}$ to be the identity ($Y = X$) we obtain the following guarantee on the closeness between the empirical and the population distribution. This fact can be used to test whether the sample was truly generated by the distribution P_X.

Corollary 7.1. *Suppose P_X is discrete with support $\mathcal{X}$, If $\mathcal{X}$ is infinite, then*

$$\lim_{n\to\infty} n\,\mathbb{E}[D(\hat{P}_n\|P_X)] = \infty\,. \tag{7.41}$$

Otherwise, we have

$$\mathbb{E}[D(\hat{P}_n\|P_X)] \le \frac{\log e}{n}(|\mathcal{X}| - 1)\,. \tag{7.42}$$

Proof. Simply notice that $I_{\chi^2}(X;X) = |\mathcal{X}| - 1$. □

Remark 7.8. For fixed P_X, the tight asymptotic result is

$$\lim_{n\to\infty} n\,\mathbb{E}[D(\hat{P}_n\|P_X)] = \frac{\log e}{2}(|\mathrm{supp}(P_X)| - 1)\,. \tag{7.43}$$

See Lemma 11.1 below.

Corollary 7.1 is also useful for the statistical application of *entropy estimation*. Given n iid samples, a natural estimator of the entropy of P_X is the empirical entropy $\hat{H}_{\mathrm{emp}} = H(\hat{P}_n)$ (plug-in estimator). It is clear that empirical entropy is an *underestimate*, in the sense that the bias

$$\mathbb{E}[\hat{H}_{\mathrm{emp}}] - H(P_X) = -\,\mathbb{E}[D(\hat{P}_n\|P_X)]$$

is always non-negative. For fixed P_X, $\hat{H}_{\mathrm{emp}}$ is known to be consistent even on countably infinite alphabets [AK01], although the convergence rate can be arbitrarily slow, which aligns with the conclusion of (7.41). However, for large alphabet of size $\Theta(n)$, the upper bound (7.42) does not vanish (this is tight for, e.g., uniform distribution). In this case, one need to de-bias the empirical entropy (e.g. on the basis of (7.43)) or employ different techniques in order to achieve consistent estimation.

7.10 Most f-divergences are locally χ^2-like

In this section we prove analogs of Proposition 5.1 and Proposition 5.2 for the general f-divergences.

Theorem 7.10. *Suppose that $D_f(P\|Q) < \infty$ and derivative of $f(x)$ at $x = 1$ exist. Then,*

$$\lim_{\lambda\to 0}\frac{1}{\lambda}D_f(\lambda P + \bar{\lambda}Q\|Q) = (1 - P[\mathrm{supp}Q])f'(\infty)\,,$$

where as usual we take $0\cdot\infty = 0$ in the left-hand side.

Remark 7.9. Note that we do not need a separate theorem for $D_f(Q\|\lambda P + \bar{\lambda}Q)$ since the exchange of arguments leads to another f-divergence with $f(x)$ replaced by $xf(1/x)$.

Proof. Without loss of generality we may assume $f(1) = f'(1) = 0$ and $f \ge 0$. Then, decomposing $P = \mu P_1 + \bar{\mu}P_0$ with $P_0 \perp Q$ and $P_1 \ll Q$ we have

$$\frac{1}{\lambda}D_f(\lambda P + \bar{\lambda}Q\|Q) = \bar{\mu}f'(\infty) + \int dQ\frac{1}{\lambda}f\left(1 + \lambda\frac{\mu P_1 - Q}{Q}\right)\,.$$

Note that $g(\lambda) = f\,(1 + \lambda t)$ is positve and convex for every $t \in \mathbb{R}$ and hence $\frac{1}{\lambda}g(\lambda)$ is monotonically decreasing to $g'(0) = 0$ as $\lambda \searrow 0$. Since for $\lambda = 1$ the integrand is assumed to be Q-integrable, the dominated convergence theorem applies and we get the result. □

Theorem 7.11. *Let f be twice continuously differentiable on $(0,\infty)$ with*

$$\limsup_{x\to+\infty} f''(x) < \infty\,.$$

If $\chi^2(P\|Q)<\infty$, then $D_f(\bar\lambda Q+\lambda P\|Q)<\infty$ for all $0\le\lambda<1$ and

$$\lim_{\lambda\to 0}\frac{1}{\lambda^2}D_f(\bar\lambda Q+\lambda P\|Q)=\frac{f''(1)}{2}\chi^2(P\|Q)\,. \tag{7.44}$$

If $\chi^2(P\|Q)=\infty$ and $f''(1)>0$ then (7.44) also holds, i.e. $D_f(\bar\lambda Q+\lambda P\|Q)=\omega(\lambda^2)$.

Remark 7.10. Conditions of the theorem include D, D_{SKL}, H^2, JS, LC and all Rényi-type divergences, with $f(x)=\frac{1}{p-1}(x^p-1)$, of orders $p<2$. A similar result holds also for the case when $f''(x)\to\infty$ with $x\to+\infty$ (e.g. Rényi-type divergences with $p>2$), but then we need to make extra assumptions in order to guarantee applicability of the dominated convergence theorem (often just the finiteness of $D_f(P\|Q)$ is sufficient).

Proof. Assuming that $\chi^2(P\|Q)<\infty$ we must have $P\ll Q$ and hence we can use (7.1) as the definition of D_f. Note that under (7.1) without loss of generality we may assume $f'(1)=f(1)=0$ (indeed, for that we can just add a multiple of $(x-1)$ to $f(x)$, which does not change the value of $D_f(P\|Q)$). From the Taylor expansion we have then

$$f(1+u)=u^2\int_0^1(1-t)f''(1+tu)dt\,.$$

Applying this with $u=\lambda\frac{P-Q}{Q}$ we get

$$D_f(\bar\lambda Q+\lambda P\|Q)=\int dQ\int_0^1 dt(1-t)\lambda^2\left(\frac{P-Q}{Q}\right)^2 f''\left(1+t\lambda\frac{P-Q}{Q}\right)\,. \tag{7.45}$$

Note that for any $\epsilon>0$ we have $\sup_{x\ge\epsilon}|f''(x)|\triangleq C_\epsilon<\infty$. Note that $\frac{P-Q}{Q}\ge -1$ and, thus, for every λ the integrand is non-negative and bounded by

$$\left(\frac{P-Q}{Q}\right)^2 C_{1-\lambda} \tag{7.46}$$

which is integrable over $dQ\times\mathrm{Leb}[0,1]$ (by finiteness of $\chi^2(P\|Q)$ and Fubini, which applies due to non-negativity). Thus, $D_f(\bar\lambda Q+\lambda P\|Q)<\infty$. Dividing (7.45) by λ^2 we see that the integrand is dominated by (7.46) and hence we can apply the dominated convergence theorem to conclude

$$\begin{aligned}\lim_{\lambda\to0}\frac{1}{\lambda^2}D_f(\bar\lambda Q+\lambda P\|Q)&\overset{\text{(a)}}{=}\int_0^1 dt(1-t)\int dQ\left(\frac{P-Q}{Q}\right)^2\lim_{\lambda\to0}f''\left(1+t\lambda\frac{P-Q}{Q}\right)\\&=\int_0^1 dt(1-t)\int dQ\left(\frac{P-Q}{Q}\right)^2 f''(1)=\frac{f''(1)}{2}\chi^2(P\|Q)\,,\end{aligned}$$

which proves (7.44).

We proceed to proving that $D_f(\lambda P+\bar\lambda Q\|Q)=\omega(\lambda^2)$ when $\chi^2(P\|Q)=\infty$. If $P\ll Q$ then this follows by replacing the equality in (a) with $\ge$ due to Fatou lemma. If $P\not\ll Q$, we consider decomposition $P=\mu P_1+\bar\mu P_0$ with $P_1\ll Q$ and $P_0\perp Q$. From definition (7.2) we have (for $\lambda_1=\frac{\lambda\mu}{1-\lambda\bar\mu}$)

$$D_f(\lambda P+\bar\lambda Q\|Q)=(1-\lambda\bar\mu)D_f(\lambda_1P_1+\bar\lambda_1Q\|Q)+\lambda\bar\mu D_f(P_0\|Q)\ge\lambda\bar\mu D_f(P_0\|Q)\,.$$

Note that $D_f(P_0\|Q)>0$ unless $f=\mathrm{const}(x-1)$ (see Prop. 7.1) and the proof is complete. □

7.11 Local expansion of f-divergences in parametric families

In Section 5.3* we have already previewed the fact that in parametric families of distributions, the Hessian of the KL divergence turns out to coincide with the Fisher information. Here we collect such facts and their proofs. These materials form the basis of sharp bounds on parameter estimation that we will study later in Lecture 29.

To start with an example, let us return to the Gaussian location family $P_t \triangleq \mathcal{N}(t,1), t \in \mathbb{R}$. From the identities presented in Section 7.7 we obtain the following asymptotics:

$$\begin{aligned}
\mathrm{TV}(P_t, P_0) &= \frac{|t|}{\sqrt{2\pi}} + o(|t|), & H^2(P_t, P_0) &= \frac{t^2}{4} + o(t^2)\,, \\
\chi^2(P_t\|P_0) &= t^2 + o(t^2), & D(P_t\|P_0) &= \frac{t^2}{2\log e} + o(t^2)\,, \\
\mathrm{LC}(P_t, P_0) &= \frac{1}{4}t^2 + o(t^2)\,.
\end{aligned}$$

We can see that with the exception of TV, other f-divergences behave quadratically under small displacement $t \to 0$. This turns out to be a general fact, and furthermore the coefficient in front of t^2 is given by the Fisher information (at $t = 0$). To proceed carefully, we need some technical assumptions on the family P_t.

Definition 7.3 (Regular single-parameter families)**.** Fix $\tau > 0$, space $\mathcal{X}$ and a family P_t of distributions on $\mathcal{X}$, $t \in [0, \tau)$. We define the following types of conditions that we call regularity at $t = 0$:

a) $P_t(dx) = p_t(x)\mu(dx)$, for some measurable $(t, x) \mapsto p_t(x) \in \mathbb{R}_+$ and a fixed measure μ on $\mathcal{X}$;

b_0) There exists a measurable function $(s, x) \mapsto \dot{p}_s(x)$, $s \in [0, \tau), x \in \mathcal{X}$, such that for μ-almost every x_0 we have $\int_0^\tau |\dot{p}_s(x_0)|ds < \infty$ and

$$p_t(x_0) = p_0(x_0) + \int_0^t \dot{p}_s(x_0)ds\,. \tag{7.47}$$

Furthermore, for μ-almost every x_0 we have $\lim_{t\searrow 0} \dot{p}_t(x_0) = \dot{p}_0(x_0)$.

b_1) We have $\dot{p}_t(x) = 0$ whenever $p_0(x) = 0$ and, furthermore,

$$\int_{\mathcal{X}} \mu(dx) \sup_{0\le t<\tau} \frac{(\dot{p}_t(x))^2}{p_0(x)} < \infty\,. \tag{7.48}$$

c_0) There exists a measurable function $(s, x) \mapsto \dot{h}_s(x)$, $s \in [0, \tau), x \in \mathcal{X}$, such that for μ-almost every x_0 we have $\int_0^\tau |\dot{h}_s(x_0)|ds < \infty$ and

$$h_t(x_0) \triangleq \sqrt{p_t(x_0)} = \sqrt{p_0(x_0)} + \int_0^t \dot{h}_s(x_0)ds\,. \tag{7.49}$$

Furthermore, for μ-almost every x_0 we have $\lim_{t\searrow 0} \dot{h}_t(x_0) = \dot{h}_0(x_0)$.

c_1) The family of functions $\{(\dot{h}_t(x))^2 : t \in [0, \tau)\}$ is uniformly μ-integrable.

Remark 7.11. Recall that the uniform integrability condition c_1) is implied by the following stronger (but easier to verify) condition:

$$\int_{\mathcal{X}} \mu(dx) \sup_{0\le t<\tau} (\dot{h}_t(x))^2 < \infty\,. \tag{7.50}$$

Impressively, if one also assumes continuous differentiability of h_t then the uniform integrability condition becomes equivalent to the continuity of

$$t \mapsto J_F(t) \triangleq 4\int \mu(dx)(\dot{h}_t(x))^2\,.$$

We refer to [Bor99, Appendix V] for this finesse.

Theorem 7.12. *Given a family of distributions $\{P_t : t \in [0,\tau)\}$ satisfying the conditions a), b_0) and b_1) in Definition 7.3. Then we have*

$$\chi^2(P_t\|P_0) = J_F(0)t^2 + o(t^2)\,, \tag{7.51}$$

$$D(P_t\|P_0) = \frac{J_F(0)}{2\log e}t^2 + o(t^2)\,, \tag{7.52}$$

where $J_F(0) \triangleq \int_{\mathcal{X}} \mu(dx)\frac{(\dot{p}_0(x))^2}{p_0(x)} < \infty$ is the Fisher information at $t = 0$.

Proof. From assumption b_1) we see that for any x_0 with $p_0(x_0) = 0$ we must have $\dot{p}_t(x_0) = 0$ and thus $p_t(x_0) = 0$ for all $t \in [0,\tau)$. Hence, we may restrict all intergrals below to subset $\{x : p_0(x) > 0\}$, on which the ratio $\frac{(p_t(x_0)-p_0(x_0))^2}{p_0(x_0)}$ is well-defined. Consequently, we have by (7.47)

$$\begin{aligned}
\frac{1}{t^2}\chi^2(P_t\|P_0) &= \frac{1}{t^2}\int \mu(dx)\frac{(p_t(x)-p_0(x))^2}{p_0(x)} \\
&= \frac{1}{t^2}\int \mu(dx)\frac{1}{p_0(x)}\left(t\int_0^1 du\dot{p}_{tu}(x)\right)^2 \\
&\overset{\text{(a)}}{=} \int \mu(dx)\int_0^1 du_1\int_0^1 du_2\frac{\dot{p}_{tu_1}(x)\dot{p}_{tu_2}(x)}{p_0(x)}
\end{aligned}$$

Note that by the continuity assumption in b_1) we have $\dot{p}_{tu_1}(x)\dot{p}_{tu_2}(x) \to \dot{p}_0^2(x)$ for every (u_1,u_2,x) as $t \to 0$. Furthermore, we also have $\left|\frac{\dot{p}_{tu_1}(x)\dot{p}_{tu_2}(x)}{p_0(x)}\right| \le \sup_{0\le t<\tau}\frac{(\dot{p}_t(x_0))^2}{p_0(x_0)}$, which is integrable by (7.48). Consequently, application of the dominated convergence theorem to the integral in (a) concludes the proof of (7.51).

We next show that for any f-divergence with twice continuously differentiable f (and in fact, without assuming (7.48)) we have:

$$\liminf_{t\to 0}\frac{1}{t^2}D_f(P_t\|P_0) \ge \frac{f''(1)}{2}J_F(0)\,. \tag{7.53}$$

Indeed, similar to (7.45) we get

$$D_f(P_t\|P_0) = \int_0^1 dz(1-z)\,\mathbb{E}_{X\sim P_0}\left[f''\left(1+z\frac{p_t(X)-p_0(X)}{p_0(X)}\right)\left(\frac{p_t(X)-p_0(X)}{p_0(X)}\right)^2\right]. \tag{7.54}$$

Dividing by t^2 notice that from b_0) we have $\frac{p_t(X)-p_0(X)}{tp_0(X)} \xrightarrow{\text{a.s.}} \frac{\dot{p}_0(X)}{p_0(X)}$ and thus

$$f''\left(1+z\frac{p_t(X)-p_0(X)}{p_0(X)}\right)\left(\frac{p_t(X)-p_0(X)}{tp_0(X)}\right)^2 \to f''(1)\left(\frac{\dot{p}_0(X)}{p_0}\right)^2 .$$

Thus, applying Fatou's lemma we recover (7.53).

Next, plugging $f(x) = x\log x$ in (7.54) we obtain for the KL divergence

$$\frac{1}{t^2}D(P_t\|P_0) = \frac{1}{\log e}\int_0^1 dz\, \mathbb{E}_{X\sim P_0}\left[\frac{1-z}{1+z\frac{p_t(X)-p_0(X)}{p_0(X)}}\left(\frac{p_t(X)-p_0(X)}{tp_0(X)}\right)^2\right] . \tag{7.55}$$

The first fraction inside the bracket is between 0 and 1 and the second by $\sup_{0<t<\tau}\left(\frac{\dot{p}_t(X)}{p_0(X)}\right)^2$, which is P_0-integrable by b_1). Thus, dominated convergence theorem applies to the double integral in (7.54) and we obtain

$$\lim_{t\to 0}\frac{1}{t^2}D(P_t\|P_0) = \frac{1}{\log e}\int_0^1 dz\, \mathbb{E}_{X\sim P_0}\left[(1-z)\left(\frac{\dot{p}_0(X)}{p_0(X)}\right)^2\right] ,$$

completing the proof of (7.52). □

Theorem 7.12 applies to many cases (e.g. to smooth subfamilies of exponential families, for which one can take $\mu = P_0$ and $p_0(x) \equiv 1$), but it is not sufficiently general. To demonstrate the issue, consider the following example.

Example 7.1 (Location families with compact support)**.** We say that family P_t is a (scalar) location family if $\mathcal{X} = \mathbb{R}$, $\mu = \text{Leb}$ and $p_t(x) = p_0(x-t)$. Consider the following example, for $\alpha > -1$:

$$p_0(x) = C_\alpha \times \begin{cases} x^\alpha, & x\in[0,1] , \\ (2-x)^\alpha, & x\in[1,2] , \\ 0, & \text{otherwise} \end{cases} ,$$

with C_α chosen from normalization. Clearly, here condition (7.48) is not satisfied and both $\chi^2(P_t\|P_0)$ and $D(P_t\|P_0)$ are infinite for $t > 0$, since $P_t \not\ll P_0$. But $J_F(0) < \infty$ whenever $\alpha > 1$ and thus one expects that a certain remedy should be possible. Indeed, one can compute those f-divergences that are finite for $P_t \not\ll P_0$ and find that for $\alpha > 1$ they are quadratic in t. As an illustration, we have

$$H^2(P_t, P_0) = \begin{cases} \Theta(t^{1+\alpha}), & 0\le\alpha<1 \\ \Theta(t^2\log\frac{1}{t}), & \alpha = 1 \\ \Theta(t^2), & \alpha > 1 \end{cases} \tag{7.56}$$

as $t \to 0$. This can be computed directly, or from a more general results of [IK81, Theorem VI.1.1].[4]

The previous example suggests that quadratic behavior as $t \to 0$ can hold even when $P_t \not\ll P_0$, which is the case handled by the next (more technical) result, whose proof we placed in Section 7.14*). One can verify that condition c_1) is indeed satisfied for all $\alpha > 1$ in Example 7.1, thus establishing the quadratic behavior. Also note that the stronger (7.50) only applies to $\alpha \ge 2$.

[4]Statistical significance of this calculation is that if we were to estimate the location parameter t from n iid samples, then precision δ_n^* of the optimal estimator up to constant factors is given by solving $H^2(P_{\delta_n^*}, P_0) \asymp \frac{1}{n}$, cf. [IK81, Chapter VI]. For $\alpha < 1$ we have $\delta_n^* \asymp n^{-\frac{1}{1+\alpha}}$ which is notably better than the empirical mean estimator (attaining precision of only $n^{-\frac{1}{2}}$). For $\alpha = 1/2$ this fact was noted by D. Bernoulli in 1777 as a consequence of his (newly proposed) maximum likelihood estimation.

Theorem 7.13. *Given a family of distributions $\{P_t : t \in [0,\tau)\}$ satisfying the conditions a), c_0) and c_1) of Definition 7.3, we have*

$$\chi^2(P_t \| \bar{\epsilon} P_0 + \epsilon P_t) = t^2 \bar{\epsilon}^2 \left(J_F(0) + \frac{1-4\epsilon}{\epsilon} J^{\#}(0) \right) + o(t^2)\,, \qquad \forall \epsilon \in (0,1) \tag{7.57}$$

$$H^2(P_t, P_0) = \frac{t^2}{4} J_F(0) + o(t^2)\,, \tag{7.58}$$

where $J_F(0) = 4\int \dot{h}_0^2 d\mu < \infty$ is the Fisher information and $J^{\#}(0) = \int \dot{h}_0^2 \mathbf{1}_{\{h_0=0\}} d\mu$ can be called the Fisher defect at $t = 0$.

Example 7.2 (On Fisher defect)**.** Note that in most cases of interest we will have the situation that $t \mapsto h_t(x)$ is actually differentiable for all t in some *two-sided* neighborhood $(-\tau, \tau)$ of 0. In such cases, $h_0(x) = 0$ implies that $t = 0$ is a local minima and thus $\dot{h}_0(x) = 0$, implying that the defect $J_F^{\#} = 0$. However, for other families this will not be so, sometimes even when $p_t(x)$ is smooth on $t \in (-\tau, \tau)$ (but not h_t). Here is such an example.

Consider $P_t = \mathrm{Bern}(t^2)$. A straighforward calculation shows:

$$\chi^2(P_t \| \bar{\epsilon} P_0 + \epsilon P_t) = t^2 \frac{\bar{\epsilon}^2}{\epsilon} + O(t^4), \qquad H^2(P_t, P_0) = 2(1 - \sqrt{1-t^2}) = t^2 + O(t^4)\,.$$

Taking $\mu(\{0\}) = \mu(\{1\}) = 1$ to be the counting measure, we get the following

$$h_t(x) = \begin{cases} \sqrt{1-t^2}, & x = 0 \\ |t|, & x = 1 \end{cases}, \qquad \dot{h}_t(x) = \begin{cases} \frac{-t}{\sqrt{1-t^2}}, & x = 0 \\ \mathrm{sign}(t), & x = 1, t \neq 0 \\ 1, & x = 1, t = 0 \quad \text{(just as an agreement)} \end{cases}.$$

Note that if we view P_t as a family on $t \in [0,\tau)$ for small τ, then all conditions a), c_0) and c_1) are clearly satisfied ($\dot{h}_t$ is bounded on $t \in (-\tau, \tau)$). We have $J_F(0) = 4$ and $J^{\#}(0) = 1$ and thus (7.57) recovers the correct expansion for χ^2 and (7.58) for H^2.

Notice that the non-smoothness of h_t only becomes visible if we extend the domain to $t \in (-\tau, \tau)$. In fact, this issue is not seen in terms of densities p_t. Indeed, let us compute the density p_t and its derivative $\dot{p}_t$ explicitly too:

$$p_t(x) = \begin{cases} 1 - t^2, & x = 0 \\ t^2, & x = 1 \end{cases}, \qquad \dot{p}_t(x) = \begin{cases} -2t, & x = 0 \\ 2t, & x = 1 \end{cases}.$$

Clearly, p_t is continuously differentiable on $t \in (-\tau, \tau)$. Furthermore, the following expectation (typically equal to $J_F(t)$ in (**??**))

$$\mathbb{E}_{X \sim P_t} \left[\left(\frac{\dot{p}_t(X)}{p_t(X)} \right)^2 \right] = \begin{cases} 0, & t = 0 \\ 4 + \frac{4t^2}{1-t^2}, & t \neq 0 \end{cases}$$

is discontinuous at $t = 0$. To make things worse, at $t = 0$ this expectation does not match our definition of the Fisher information $J_F(0)$ in Theorem 7.13, and thus does not yield the correct small-t behavior for either χ^2 or H^2. In general, to avoid difficulties one should restrict to those families with $t \mapsto h_t(x)$ continuously differentiable in $t \in (-\tau, \tau)$.

7.12 Rényi divergences and tensorization

The following family of divergence measures introduced by Rényi is key in many applications involving product measures. Although these measures are not f-divergences, they are obtained as monotone transformation of an appropriate f-divergence and thus satisfy DPI and other properties of f-divergences.

Definition 7.4. For any $\lambda \in \mathbb{R} \setminus 0, 1$ we define the Rényi divergence of order λ as

$$D_\lambda(P\|Q) \triangleq \frac{1}{\lambda - 1} \log \mathbb{E}_Q \left[\left(\frac{dP}{dQ} \right)^\lambda \right] ,$$

where $\mathbb{E}_Q[\cdot]$ is understood as an f-divergence $D_f(P\|Q)$ with $f(x) = x^\lambda$, see Definition 7.1. Conditional Rényi divergence is defined as

$$\begin{aligned} D_\lambda(P_{X|Y}||Q_{X|Y}|P_Y) &\triangleq D_\lambda(P_Y \times P_{X|Y}||P_Y \times Q_{X|Y}) \\ &= \frac{1}{\lambda - 1} \log \mathbb{E}_{Y \sim P_Y} \int_{\mathcal{X}} (dP_{X|Y}(x))^\lambda (dQ_{X|Y}(a))^{1-\lambda} . \end{aligned}$$

Numerous properties of Rényi divergences are known, see [VEH14]. Here we only notice a few:

- Special cases of $\lambda = \frac{1}{2}, 1, 2$: Under mild regularity conditions $\lim_{\lambda \to 1} D_\lambda(P\|Q) = D(P\|Q)$. On the other hand, for D_2 is a monotone transformation of χ^2 in (7.3), while $D_{\frac{1}{2}}$ is a monotone transformation of H^2 in (7.4).

- There is a version of the chain rule:

$$D_\lambda(P_{A,B}||Q_{A,B}) = D_\lambda(P_B||Q_B) + D_\lambda(P_{A|B}||Q_{A|B}|P_B^{(\lambda)}) , \tag{7.59}$$

  where $P_B^{(\lambda)}$ is the λ-tilting of P_B towards Q_B given by

$$P_B^{(\lambda)}(b) \triangleq P_B^\lambda(b) Q_B^{1-\lambda}(b) \exp\{-(\lambda - 1) D_\lambda(P_B||Q_B)\} . \tag{7.60}$$

- However, the key property is additivity under products, or *tensorization*:

$$D_\lambda(\prod_i P_{X_i} \| \prod_i Q_{X_i}) = \sum_i D_\lambda(P_{X_i} \| Q_{X_i}) , \tag{7.61}$$

  which is a simple consequence of (7.59).

The following consequence of the chain rule will be crucial in statistical applications later.

Proposition 7.4. *Consider product channels $P_{Y^n|X^n} = \prod P_{Y_i|X_i}$ and $Q_{Y^n|X^n} = \prod Q_{Y_i|X_i}$. We have (with all optimizations over all possible distributions)*

$$\inf_{P_{X^n}, Q_{X^n}} D_\lambda(P_{Y^n} \| Q_{Y^n}) = \sum_{i=1}^n \inf_{P_{X_i}, Q_{X_i}} D_\lambda(P_{Y_i} \| Q_{Y_i}) \tag{7.62}$$

$$\sup_{P_{X^n}, Q_{X^n}} D_\lambda(P_{Y^n} \| Q_{Y^n}) = \sum_{i=1}^n \sup_{P_{X_i}, Q_{X_i}} D_\lambda(P_{Y_i} \| Q_{Y_i}) = \sum_{i=1}^n \sup_{x, x'} D_\lambda(P_{Y_i|X_i=x} \| Q_{Y_i|X_i=x'}) \tag{7.63}$$

In particular, for any collections of distributions $\{P_\theta, \theta \in \Theta\}$ *and* $\{Q_\theta, \theta \in \Theta\}$*:*

$$\inf_{P \in \mathrm{co}\{P_\theta^{\otimes n}\}, Q \in \mathrm{co}\{Q_\theta^{\otimes n}\}} D_\lambda(P\|Q) \geq n \inf_{P \in \mathrm{co}\{P_\theta\}, Q \in \mathrm{co}\{Q_\theta\}} D_\lambda(P\|Q) \tag{7.64}$$

$$\sup_{P \in \mathrm{co}\{P_\theta^{\otimes n}\}, Q \in \mathrm{co}\{Q_\theta^{\otimes n}\}} D_\lambda(P\|Q) \leq n \sup_{P \in \{P_\theta\}, Q \in \{Q_\theta\}} D_\lambda(P\|Q) \tag{7.65}$$

Remark 7.12. The mnemonic for (7.64)-(7.65) is "mixtures of products are less dissimilar than products of mixtures". The former arise in statistical problems with iid observations.

Proof. The second equality in (7.63) follows from the fact that D_λ is an increasing function of an f-divergence, and thus maximization should be attained at an extreme point of the space of probabilities, which are just the single-point masses. The main equalities (7.62)-(7.63) follow from a) restricting optimizations to product distributions and invoking (7.61); and b) the chain rule for D_λ. For example for $n = 2$, we fix P_{X^2} and Q_{X^2}, which (via channels) induce joint distributions P_{X^2,Y^2} and Q_{X^2,Y^2}. Then we have

$$D_\lambda(P_{Y_1|Y_2=y}\|Q_{Y_1|Y_2=y'}) \geq \inf_{\tilde{P}_{X_1}, \tilde{Q}_{X_1}} D_\lambda(\tilde{P}_{Y_1}\|\tilde{Q}_{Y_1})\,,$$

since $P_{Y_1|Y_2=y}$ is a distribution induced by taking $\tilde{P}_{X_1} = P_{X_1|Y_2=y}$, and similarly for $Q_{Y_1|Y_2=y'}$. In all, we get

$$D_\lambda(P_{Y^2}\|Q_{Y^2}) = D_\lambda(P_{Y_2}\|Q_{Y_2}) + D_\lambda(P_{Y_1|Y_2}\|Q_{Y_1|Y_2}|P_{Y_2}^{(\lambda)}) \geq \sum_{i=1}^{2} \inf_{P_{X_i}, Q_{X_i}} D_\lambda(P_{Y_i}\|Q_{Y_i})\,,$$

as claimed. The case of sup is handled similarly.

From (7.62)-(7.63), we get (7.64)-(7.65) by taking $\mathcal{X} = \Theta$ and specializing inf, sup to diagonal distributions P_{X^n} and Q_{X^n}, i.e. those with the property that $\mathbb{P}[X_1 = \cdots = X_n] = 1$ and $\mathbb{Q}[X_1 = \cdots = X_n] = 1$). □

7.13 Variational representation of f-divergences

In Theorem 4.5 we had a very useful variational representation of KL-divergence due to Donsker and Varadhan. In this section we show how to derive such representations for other f-divergences in a principled way. The proofs are slightly technical and given in Section 7.14* at the end of this chapter.

Let $f : (0, +\infty) \to \mathbb{R}$ be a convex function. The convex conjugate $f^* : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ of f is defined by:

$$f^*(y) = \sup_{x \in \mathbb{R}_+} xy - f(x)\,, \qquad y \in \mathbb{R}\,. \tag{7.66}$$

Denote the domain of f^* by $\mathrm{dom}(f^*) \triangleq \{y : f^*(y) < \infty\}$. Two important properties of the convex conjugates are

1. f^* is also convex (which holds regardless of f being convex or not);
2. Biconjugation: $(f^*)^* = f$, which means

$$f(x) = \sup_y xy - f^*(y)$$

and implies the following (for all $x > 0$ and y)

$$f(x) + f^*(g) \geq xy\,.$$

Similarly, we can define a convex conjugate for any convex functional $\Psi(P)$ defined on the space of measures, by setting

$$\Psi^*(g) = \sup_P \int g dP - \Psi(P)\,. \tag{7.67}$$

Under appropriate conditions (e.g. finite $\mathcal{X}$), biconjugation then yields the sought-after variational representation

$$\Psi(P) = \sup_g \int g dP - \Psi^*(g)\,. \tag{7.68}$$

Next we will now compute these conjugates for $\Psi(P) = D_f(P\|Q)$. It turns out to be convenient to first extend the definition of $D_f(P\|Q)$ to all finite signed measures P then compute the conjugate. To this end, let $f_{\text{ext}} : \mathbb{R} \to \cup\{+\infty\}$ be an extension of f, such that $f_{\text{ext}}(x) = f(x)$ for $x \geq 0$ and f_{ext} is convex on $\mathbb{R}$. In general, we can always choose $f_{\text{ext}}(x) = \infty$ for all $x < 0$. In special cases e.g. $f(x) = |x-1|/2$ or $f(x) = (x-1)^2$ we can directly take $f_{\text{ext}}(x) = f(x)$ for all x. Now we can define $D_f(P\|Q)$ for all signed measure measures P in the same way as in Definition 7.1 using f_{ext} in place of f.

For each choice of f_{ext} we have a variational representation of f-divergence:

Theorem 7.14. *Let P and Q be probability measures on $\mathcal{X}$. Fix an extension f_{ext} of f. Then*

$$D_f(P\|Q) = \sup_{g:\mathcal{X}\to\text{dom}(f^*_{\text{ext}})} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*_{\text{ext}}(g(X))]. \tag{7.69}$$

where f^*_{ext} is the conjugate of f_{ext}, i.e., $f^*_{\text{ext}}(y) = \sup_{x\in\mathbb{R}} xy - f_{\text{ext}}(x)$.

As a consequence of the variational characterization, we get the following properties for f-divergences:

1. Convexity: First of all, note that $D_f(P\|Q)$ is expressed as a supremum of affine functions (since the expectation is a linear operation). As a result, we get that $(P, Q) \mapsto D_f(P\|Q)$ is convex, which was proved in previous lectures using different method.

2. Weak lower semicontinuity: Recall the example in Remark 4.3, where $\{X_i\}$ are i.i.d. Rademachers (± 1), and
$$\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \xrightarrow{\text{D}} \mathcal{N}(0,1)$$
by the central limit theorem; however, by Proposition 7.1, for all n,
$$D_f\left(\frac{P_{X_1+X_2+\ldots+X_n}}{\sqrt{n}}\middle\|\mathcal{N}(0,1)\right) = f(0) + f'(\infty) > 0,$$
since the former distribution is discrete and the latter is continuous. Therefore similar to the KL divergence, the best we can hope for f-divergence is semicontinuity. Indeed, if $\mathcal{X}$ is a nice space (e.g., Euclidean space), in (7.69) we can restrict the function g to continuous bounded functions, in which case $D_f(P\|Q)$ is expressed as a supremum of weakly continuous functionals (note that $f^* \circ g$ is also continuous and bounded since f^* is continuous) and is hence weakly lower semi-continuous, i.e., for any sequence of distributions P_n and Q_n such that $P_n \xrightarrow{w} P$ and $Q_n \xrightarrow{w} Q$, we have
$$\liminf_{n\to\infty} D_f(P_n\|Q_n) \geq D_f(P\|Q).$$

3. Relation to DPI: As we discussed in (4.2) variational representations can be thought of as extensions of the DPI. As an exercise, one should try to derive the estimate

$$|P[A] - Q[A]| \le \sqrt{Q[A] \cdot \chi^2(P\|Q)}$$

via both the DPI and (7.73).

Example 7.3 (Total variation)**.** For total variation, we have $f(x) = \frac{1}{2}|x-1|$. Consider the extension $f_{\text{ext}}(x) = \frac{1}{2}|x-1|$ for $x \in \mathbb{R}$. Then

$$f^*_{\text{ext}}(y) = \sup_x \left\{ xy - \frac{1}{2}|x-1| \right\} = \begin{cases} +\infty & \text{if } |y| > \frac{1}{2} \\ y & \text{if } |y| \le \frac{1}{2} \end{cases} .$$

Thus (7.69) gives

$$\text{TV}(P,Q) = \sup_{g:|g|\le\frac{1}{2}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)], \tag{7.70}$$

which previously appeared in (7.14).

Example 7.4 (χ^2-divergence)**.** For χ^2-divergence we have $f(x) = (x-1)^2$. Take $f_{\text{ext}}(x) = (x-1)^2$, whose conjugate is $f^*_{\text{ext}}(y) = y + \frac{y^2}{4}$. Applying (7.69) yields

$$\chi^2(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q\left[g(X) + \frac{g^2(X)}{4} \right] \tag{7.71}$$

$$= \sup_{g:\mathcal{X}\to\mathbb{R}} 2\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g^2(X)] - 1 \tag{7.72}$$

where the last step follows from a change of variable ($g \leftarrow \frac{1}{2}g - 1$).

To get another equivalent, but much more memorable representation, we notice that (**??**) it is not scale-invariant. To make it so, setting $g = \lambda h$ and optimizing over the $\lambda \in \mathbb{R}$ first we get

$$\chi^2(P\|Q) = \sup_{h:\mathcal{X}\to\mathbb{R}} \frac{(\mathbb{E}_P[h(X)] - \mathbb{E}_Q[h(X)])^2}{\text{Var}_Q(h(X))}. \tag{7.73}$$

The statistical interpretation of (7.73) is as follows: if a test statistic $h(X)$ is such that the separation between its expectation under P and Q far exceeds its standard deviation, then this suggests the two hypothesis can be distinguished reliably. The representation (7.73) will turn out useful in statistical applications in Lecture 29 for deriving the Hammersley-Chapman-Robbins (HCR) lower bound as well as its Bayesian version, see Section 29.5, and ultimately the Cramer-Rao and van Trees lower bounds.

Example 7.5 (KL-divergence)**.** In this case we have $f(x) = x\log x$. Consider the extension $f_{\text{ext}}(x) = \infty$ for $x < 0$, whose convex conjugate is $f^*(y) = \frac{\log e}{e}\exp(y)$. Hence (7.69) yields

$$D(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - (\mathbb{E}_Q[\exp\{g(X)\}] - 1)\log e \tag{7.74}$$

Note that in the last example, the variational representation (7.74) we obtained for the KL divergence is not the same as the Donsker-Varadhan identity in Theorem 4.5, that is,

$$D(P\|Q) = \sup_{g:\mathcal{X}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[\exp\{g(X)\}] . \tag{7.75}$$

In fact, (7.74) is weaker than (7.75) in the sense that for each choice of g, the obtained lower bound on $D(P\|Q)$ in the RHS is smaller. Furthermore, regardless of the choice of f_{ext}, the Donsker-Varadhan representation can never be obtained from Theorem 7.14 because, unlike (7.75), the second term in (7.69) is always linear in Q. It turns out if we define $D_f(P\|Q) = \infty$ for all non-probability measure P, and compute its convex conjugate, we obtain in the next theorem a different type of variational representation, which, specialized to KL divergence in Example 7.5, recovers exactly the Donsker-Varadhan identity.

Theorem 7.15. *Consider the extension f_{ext} of f such that $f_{\text{ext}}(x) = \infty$ for $x < 0$. Let $S = \{x : q(x) > 0\}$ where q is as in (7.2). Then*

$$D_f(P\|Q) = f'(\infty)P[S^c] + \sup_g \mathbb{E}_P[g1_S] - \Psi^*_{Q,P}(g)\,, \tag{7.76}$$

where

$$\Psi^*_{Q,P}(g) \triangleq \inf_{a\in\mathbb{R}} \mathbb{E}_Q[f^*_{\text{ext}}(g(X) - a)] + aP[S].$$

In the special case $f'(\infty) = \infty$, we have

$$D_f(P\|Q) = \sup_g \mathbb{E}_P[g] - \Psi^*_Q(g), \qquad \Psi^*_Q(g) \triangleq \inf_{a\in\mathbb{R}} \mathbb{E}_Q[f^*_{\text{ext}}(g(X) - a)] + a. \tag{7.77}$$

Remark 7.13 (Marton's divergence)**.** Recall that in Theorem 7.5 we have shown both the sup and inf characterizations for the TV. Do other f-divergences also possess inf characterizations? The only other known example (to us) is due to Marton. Let

$$D_m(P\|Q) = \int dQ \left(1 - \frac{dP}{dQ}\right)^2_+,$$

which is clearly an f-divergence with $f(x) = (1-x)^2_+$. We have the following [BLB04, Lemma 8.3]:

$$D_m(P\|Q) = \inf\{\mathbb{E}[P^2[X \neq Y|Y]] : X \sim P, Y \sim Q\}\,,$$

where infimum is over all couplings. To prove this, a simple calculation shows that the same coupling used for (7.16) attains $\mathbb{E}[P^2[X \neq Y|Y]] = D_m(P\|Q)$. On the other hand, for the case of discrete P, Q (which can be considered without loss of generality due to Theorem 7.4), we have $P_{X,Y}(x_0, x_0) \leq P_X(x_0) = P(x_0)$ for every x_0. Thus dividing by $P_Y(x_0) = Q(x_0)$ we get

$$\mathbb{P}[X \neq Y|Y] \geq 1 - \frac{P(Y)}{Q(Y)}\,,$$

taking positive part of this, squaring and taking the expectation over $Y \sim Q$ completes the proof of

$$\mathbb{E}[\mathbb{P}^2[X \neq Y|Y]] \leq D_m(P\|Q)\,.$$

Marton's D_m divergence plays a crucial role in concentration of measure [BLB04, Chapter 8]. Note also that while Theorem 7.11 does not apply to D_m, due to absence of twice continuous differentiability, it does apply to $D_m(P\|Q) + D_m(Q\|P)$.

7.14* Technical proofs: convexity, local expansions and variational representations

In this section we collect proofs of some technical theorems from this chapter.

Proof of Theorem 7.13. By definition we have

$$L(t) \triangleq \frac{1}{\bar{\epsilon}^2 t^2}\chi^2(P_t\|\bar{\epsilon}P_0 + \epsilon P_t) = \frac{1}{t^2}\int_{\mathcal{X}} \mu(dx)\frac{(p_t(x)-p_0(x))^2}{\bar{\epsilon}p_0(x)+\epsilon p_t(x)} = \frac{1}{t^2}\int \mu(dx) g(t,x)^2\,, \tag{7.78}$$

where

$$g(t,x) \triangleq \frac{p_t(x)-p_0(x)}{\sqrt{\bar{\epsilon}p_0(x)+\epsilon p_t(x)}} = \phi(h_t(x);x)\,, \qquad \phi(h;x) \triangleq \frac{h^2-p_0(x)}{\sqrt{\bar{\epsilon}p_0(x)+\epsilon h^2}}\,.$$

By c_0) the function $t \mapsto h_t(x) \triangleq \sqrt{p_t(x)}$ is absolutely continuous (for μ-a.e. x). Below we will show that $\|\phi(\cdot;x)\|_{\mathrm{Lip}} = \sup_{h\geq 0}|\phi'(h;x)| \leq \frac{2-\epsilon}{(1-\epsilon)\sqrt{\epsilon}}$. This implies that $t \mapsto g(t,x)$ is also absolutely continuous and hence differentiable almost everywhere. Consequently, we have

$$g(t,x) = t\int_0^1 du \dot{g}(tu,x), \qquad \dot{g}(t,x) \triangleq \phi'(h_t(x);x)\dot{h}_t(x)\,,$$

Since $\phi'(\cdot;x)$ is continuous with

$$\phi'(h_0(x);x) = \begin{cases} 2, & x: h_0(x) > 0, \\ \frac{1}{\sqrt{\epsilon}}, & x: h_0(x) = 0 \end{cases} \tag{7.79}$$

(we verify these facts below too), we conclude that

$$\lim_{s\to 0}\dot{g}(s,x) = \dot{g}(0,x) = \dot{h}_0(x)\left(2\cdot 1\{h_0(x)>0\} + \frac{1}{\sqrt{\epsilon}}1\{h_0(x)=0\}\right)\,, \tag{7.80}$$

where we also used continuity $\dot{h}_t(x) \to \dot{h}_0(x)$ by assumption c_0).

Substituting the integral expression for $g(t,x)$ into (7.78) we obtain

$$L(t) = \int \mu(dx)\int_0^1 du_1 \int_0^1 du_2 \dot{g}(tu_1,x)\dot{g}(tu_2,x)\,. \tag{7.81}$$

Since $|\dot{g}(s,x)| \leq C|h_s(x)|$ for some $C = C(\epsilon)$, we have from Cauchy-Schwarz

$$\int \mu(dx)|g(s_1,x)\dot{g}(s_2,x)| \leq C^2 \sup_t \int_{\mathcal{X}} \mu(dx)\dot{h}_t(x)^2 < \infty\,. \tag{7.82}$$

where the last inequality follows from the uniform integrability assumption c_1). This implies that Fubini's theorem applies in (7.81) and we obtain

$$L(t) = \int_0^1 du_1 \int_0^1 du_2 G(tu_1,tu_2)\,, \qquad G(s_1,s_2) \triangleq \int \mu(dx)\dot{g}(s_1,x)\dot{g}(s_2,x)\,.$$

Notice that if a family of functions $\{f_\alpha(x): \alpha \in I\}$ is uniformly square-integrable, then the family $\{f_\alpha(x)f_\beta(x): \alpha\in I, \beta\in I\}$ is uniformly integrable simply because apply $|f_\alpha f_\beta| \leq \frac{1}{2}(f_\alpha^2 + f_\beta^2)$.

Consequently, from the assumption c_1) we see that the integral defining $G(s_1, s_2)$ allows passing the limit over s_1, s_2 inside the integral. From (7.80) we get as $t \to 0$

$$G(tu_1, tu_2) \to G(0,0) = \int \mu(dx)\dot{h}_0(x)^2 \left(4 \cdot 1\{h_0 > 0\} + \frac{1}{\epsilon} 1\{h_0 = 0\}\right) = J_F(0) + \frac{1-4\epsilon}{\epsilon} J^{\#}(0)\,.$$

From (7.82) we see that $G(s_1, s_2)$ is bounded and thus, the bounded convergence theorem applies and

$$\lim_{t\to 0} \int_0^1 du_1 \int_0^1 du_2 G(tu_1, tu_2) = G(0,0)\,,$$

which thus concludes the proof of $L(t) \to J_F(0)$ and of (7.57) assuming facts about ϕ. Let us verify those.

For simplicity, in the next paragraph we omit the argument x in $h_0(x)$ and $\phi(\cdot; x)$. A straightforward differentiation yields

$$\phi'(h) = 2h \frac{h_0^2(1-\frac{\epsilon}{2}) + \frac{\epsilon}{2}h^2}{(\bar{\epsilon}h_0^2 + \epsilon h^2)^{3/2}}.$$

Since $\frac{h}{\sqrt{\bar{\epsilon}h_0^2 + \epsilon h^2}} \le \frac{1}{\sqrt{\epsilon}}$ and $\frac{h_0^2(1-\frac{\epsilon}{2}) + \frac{\epsilon}{2}h^2}{\bar{\epsilon}h_0^2 + \epsilon h^2} \le \frac{1-\epsilon/2}{1-\epsilon}$ we obtain the finiteness of ϕ'. For the continuity of ϕ' notice that if $h_0 > 0$ then clearly the function is continuous, whereas for $h_0 = 0$ we have $\phi'(h) = \frac{1}{\sqrt{\epsilon}}$ for all h.

We next proceed to the Hellinger distance. Just like in the argument above, we define

$$M(t) \triangleq \frac{1}{t^2} H^2(P_t, P_0) = \int \mu(dx) \int_0^1 du_1 \int_0^1 du_2 \dot{h}_{tu_1}(x)\dot{h}_{tu_2}(x)\,.$$

Exactly as above from Cauchy-Schwarz and $\sup_t \int \mu(dx)\dot{h}_t(x)^2 < \infty$ we conclude that Fubini applies and hence

$$M(t) = \int_0^1 du_1 \int_0^1 du_2 H(tu_1, tu_2)\,, \qquad H(s_1, s_2) \triangleq \int \mu(dx)\dot{h}_{s_1}(x)\dot{h}_{s_2}(x)\,.$$

Again, the family $\{\dot{h}_{s_1}\dot{h}_{s_2} : s_1 \in [0,\tau), s_2 \in [0,\tau\}$ is uniformly integrable and thus from c_0) we conclude $H(tu_1, tu_2) \to \frac{1}{4}J_F(0)$. Furthermore, similar to (7.82) we see that $H(s_1, s_2)$ is bounded and thus

$$\lim_{t\to 0} M(t) = \int_0^1 du_1 \int_0^1 du_2 \lim_{t\to 0} H(tu_1, tu_2) = \frac{1}{4} J_F(0)\,,$$

concluding the proof of (7.58). □

Proceeding to variational representations, we prove the counterpart of Gelfand-Yaglom-Perez Theorem 4.7.

Proof of Theorem 7.4. The lower bound $D_f(P\|Q) \ge D_f(P_{\mathcal{E}}\|Q_{\mathcal{E}})$ follows from the DPI. To prove an upper bound, first we reduce to the case of $f \ge 0$ by property 6 in Prop. 7.1. Then define sets $S = \text{supp}Q$, $F_\infty = \{\frac{dP}{dQ} = 0\}$ and for a fixed $\epsilon > 0$ let

$$F_m = \left\{\epsilon m \le f\left(\frac{dP}{dQ}\right) < \epsilon(m+1)\right\}, m = 0, 1, \dots\,.$$

We have

$$\epsilon \sum_m mQ[F_m] \le \int_S dQ f\left(\frac{dP}{dQ}\right) \le \epsilon \sum_m (m+1)Q[F_m] + f(0)Q[F_\infty]$$
$$\le \epsilon \sum_m mQ[F_m] + f(0)Q[F_\infty] + \epsilon\,. \tag{7.83}$$

Notice that on the interval $I_m^+ = \{x > 1 : \epsilon m \le f(x) < \epsilon(m+1)\}$ the function f is increasing and on $I_m^- = \{x \le 1 : \epsilon m \le f(x) < \epsilon(m+1)\}$ it is decreasing. Thus partition further every F_m into $F_m^+ = \{\frac{dP}{dQ} \in I_m^+\}$ and $F_m^- = \{\frac{dP}{dQ} \in I_m^-\}$. Then, we see that

$$f\left(\frac{P[F_m^\pm]}{Q[F_m^\pm]}\right) \ge \epsilon m\,.$$

Consequently, for a fixed n define the partition consisting of sets $\mathcal{E} = \{F_0^+, F_0^-, \ldots, F_n^+, F_n^-, F_\infty, S^c, \cup_{m>n} F_m\}$. For this partition we have, by the previous display:

$$D(P_\mathcal{E}\|Q_\mathcal{E}) \ge \epsilon \sum_{m\le n} mQ[F_m] + f(0)Q[F_\infty] + f'(\infty)P[S^c]\,. \tag{7.84}$$

We next show that with sufficiently large n and sufficiently small ϵ the RHS of (7.84) approaches $D_f(P\|Q)$. If $f(0)Q[F_\infty] = \infty$ (and hence $D_f(P\|Q) = \infty$) then clearly (7.84) is also infinite. Thus, assume that $f(0)Q[F_\infty] < \infty$.

If $\int_S dQ f\left(\frac{dP}{dQ}\right) = \infty$ then the sum over m on the RHS of (7.83) is also infinite, and hence for any $N > 0$ there exists some n such that $\sum_{m\le n} mQ[F_m] \ge N$, thus showing that RHS for (7.84) can be made arbitrarily large. Thus assume $\int_S dQ f\left(\frac{dP}{dQ}\right) < \infty$. Considering LHS of (7.83) we conclude that for some large n we have $\sum_{m>n} mQ[F_m] \le \frac{1}{2}$. Then, we must have again from (7.83)

$$\epsilon \sum_{m\le n} mQ[F_m] + f(0)Q[F_\infty] \ge \int_S dQ f\left(\frac{dP}{dQ}\right) - \frac{3}{2}\epsilon\,.$$

Thus, we have shown that for arbitrary $\epsilon > 0$ the RHS of (7.84) can be made greater than $D_f(P\|Q) - \frac{3}{2}\epsilon$. □

Proof of Theorem 7.14. Armed with Theorem 7.4, it suffices to show (7.69) for finite $\mathcal{X}$. Indeed, for general $\mathcal{X}$, given a finite partition $\mathcal{E} = \{E_1, \ldots, E_n\}$ of $\mathcal{X}$, we say a function $g : \mathcal{X} \to \mathbb{R}$ is $\mathcal{E}$-compatible if g is constant on each $E_i \in \mathcal{E}$. Taking the supremum over all finite partitions $\mathcal{E}$ we get

$$\begin{aligned} D_f(P\|Q) &= \sup_\mathcal{E} D_f(P_\mathcal{E}\|Q_\mathcal{E}) \\ &= \sup_\mathcal{E} \sup_{\substack{g:\mathcal{X}\to \mathrm{dom}(f^*_{\mathrm{ext}}) \\ g\ \mathcal{E}\text{-compatible}}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*_{\mathrm{ext}}(g(X))] \\ &= \sup_{g:\mathcal{X}\to \mathrm{dom}(f^*_{\mathrm{ext}})} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*_{\mathrm{ext}}(g(X))], \end{aligned}$$

where the last step follows is because the two sumprema combined is equivalent to the supremum over all simple (finitely-valued) functions g, which are dense in all measurable functions.

Next, consider finite $\mathcal{X}$. Let $S = \{x \in \mathcal{X} : Q(x) > 0\}$ denote the support of Q. We show the following statement

$$D_f(P\|Q) = \sup_{g:S\to \mathrm{dom}(f^*_{\mathrm{ext}})} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*_{\mathrm{ext}}(g(X))] + f'(\infty)P(S^c), \tag{7.85}$$

which is equivalent to (7.69) simple because $\sup\{x : x \in \mathrm{dom}(f^*_{\mathrm{ext}})\} = \lim_{x\to\infty} \frac{f_{\mathrm{ext}}(x)}{x} = f'(\infty)$. By definition,

$$D_f(P\|Q) = \underbrace{\sum_{x\in S} Q(x) f_{\mathrm{ext}}\left(\frac{P(x)}{Q(x)}\right)}_{\triangleq \Psi(P)} + f'(\infty)\cdot P(S^c),$$

Consider the functional $\Psi(P)$ defined above where P takes values over all signed measures on S, which can be identified with $\mathbb{R}^S$. The convex conjugate of $\Psi(P)$ is as follows: for any $g : S \to \mathbb{R}$,

$$\begin{aligned}
\Psi^*(g) &= \sup_P \sum_x P(x)g(x) - Q(x)\left\{\sup_{h\in\mathrm{dom}(f^*_{\mathrm{ext}})} \frac{P(x)}{Q(x)}h - f^*_{\mathrm{ext}}(h)\right\} \\
&= \sup_P \inf_{h:S\to\mathrm{dom}(f^*_{\mathrm{ext}})} \sum_x P(x)(g(x)-h(x)) + Q(x) f^*_{\mathrm{ext}}(h(x)) \\
&\overset{(a)}{=} \inf_{h:S\to\mathrm{dom}(f^*_{\mathrm{ext}})} \sup_P \sum_x P(x)(g(x)-h(x)) + \mathbb{E}_Q[f^*_{\mathrm{ext}}(h)] \\
&= \begin{cases} \mathbb{E}_Q[f^*_{\mathrm{ext}}(g(X))] & g : S \to \mathrm{dom}(f^*_{\mathrm{ext}}) \\ +\infty & \text{otherwise} \end{cases}.
\end{aligned}$$

where (a) follows from the minimax theorem (which applies due to finiteness of $\mathcal{X}$). Applying the convex duality in (7.68) yields the proof of the desired (7.85). □

Proof of Theorem 7.15. First we argue that the supremum in the right-hand side of (7.76) can be taken over all simple functions g. Then thanks to Theorem 7.4, it will suffice to consider finite alphabet $\mathcal{X}$. To that end, fix any g. For any δ, there exists a such that $\mathbb{E}_Q[f^*_{\mathrm{ext}}(g-a)] - aP[S] \le \Psi^*_{Q,P}(g) + \delta$. Since $\mathbb{E}_Q[f^*_{\mathrm{ext}}(g-a_n)]$ can be approximated arbitrarily well by simple functions we conclude that there exists a simple function $\tilde{g}$ such that simultaneously $\mathbb{E}_P[\tilde{g}1_S] \ge \mathbb{E}_P[g1_S] - \delta$ and

$$\Psi^*_{Q,P}(\tilde{g}) \le \mathbb{E}_Q[f^*_{\mathrm{ext}}(\tilde{g}-a)] - aP[S] + \delta \le \Psi^*_{Q,P}(g) + 2\delta\,.$$

This implies that restricting to simple functions in the supremization in (7.76) does not change the right-hand side.

Next consider finite $\mathcal{X}$. We proceed to compute the conjugate of Ψ, where $\Psi(P) \triangleq D_f(P\|Q)$ if P is a probability measure on $\mathcal{X}$ and $+\infty$ otherwise. Then for any $g : \mathcal{X} \to \mathbb{R}$, maximizing over all

Bibliography

[AFTS01] I.C. Abou-Faycal, M.D. Trott, and S. Shamai. The capacity of discrete-time memoryless rayleigh-fading channels. *IEEE Transaction Information Theory*, 47(4):1290 – 1301, 2001.

[Ahl82] Rudolf Ahlswede. An elementary proof of the strong converse theorem for the multiple-access channel. *J. Combinatorics, Information and System Sciences*, 7(3), 1982.

[AK01] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[Alo81] Noga Alon. On the number of subgraphs of prescribed type of graphs with a given number of edges. *Israel J. Math.*, 38(1-2):116–130, 1981.

[AMS04] Shiri Artstein, Vitali Milman, and Stanisław J Szarek. Duality of metric entropy. *Annals of mathematics*, pages 1313–1328, 2004.

[AN07] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[AS08] Noga Alon and Joel H. Spencer. *The Probabilistic Method*. John Wiley & Sons, 3rd edition, 2008.

[Ash65] Robert B. Ash. *Information Theory*. Dover Publications Inc., New York, NY, 1965.

[Ban92] Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.

[BC15] Sergey Bobkov and Gennadiy P Chistyakov. Entropy power inequality for the Rényi entropy. *IEEE Transactions on Information Theory*, 61(2):708–714, 2015.

[BF14] Ahmad Beirami and Faramarz Fekri. Fundamental limits of universal lossless one-to-one compression of parametric sources. In *Information Theory Workshop (ITW), 2014 IEEE*, pages 212–216. IEEE, 2014.

[Bir83] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(2):181–237, 1983.

[Bla74] R. E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Inf. Theory*, 20(4):405–417, 1974.

[BLB04] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, pages 208–240. Springer, 2004.

[BNO03] Dimitri P Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar. *Convex analysis and optimization*. Athena Scientific, Belmont, MA, USA, 2003.

[Boh38] H. F. Bohnenblust. Convex regions and projections in Minkowski spaces. *Ann. Math.*, 39(2):301–308, 1938.

[Bor99] A.A. Borovkov. *Mathematical Statistics*. CRC Press, 1999.

[Bre73] Lev M Bregman. Some properties of nonnegative matrices and their permanents. *Soviet Math. Dokl.*, 14(4):945–949, 1973.

[Bro86] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. In S. S. Gupta, editor, *Lecture Notes-Monograph Series*, volume 9. Institute of Mathematical Statistics, Hayward, CA, 1986.

[CB90] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory*, 36(3):453–471, 1990.

[CB94] Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

[Ç11] Erhan Çinlar. *Probability and Stochastics*. Springer, New York, 2011.

[Cha05] Sourav Chatterjee. An error bound in the Sudakov-Fernique inequality. *arXiv preprint arXiv:0510424*, 2005.

[Che56] Herman Chernoff. Large-sample theory: Parametric case. *The Annals of Mathematical Statistics*, 27(1):1–22, 1956.

[Cho56] Noam Chomsky. Three models for the description of language. *IRE Trans. Inform. Th.*, 2(3):113–124, 1956.

[CK81a] I. Csiszár and J. Körner. Graph decomposition: a new key to coding theorems. *IEEE Trans. Inf. Theory*, 27(1):5–12, 1981.

[CK81b] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Academic, New York, 1981.

[CR+51] Douglas G Chapman, Herbert Robbins, et al. Minimum variance estimation without regularity assumptions. *The Annals of Mathematical Statistics*, 22(4):581–586, 1951.

[CS83] J. Conway and N. Sloane. A fast encoding method for lattice codes and quantizers. *IEEE Transactions on Information Theory*, 29(6):820–824, Nov 1983.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:229–318, 1967.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[Doo53] Joseph L. Doob. *Stochastic Processes*. New York Wiley, 1953.

[DSC] P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Probab.*, 6(3):695–750.

[Dud99] Richard M Dudley. *Uniform central limit theorems*. Number 63. Cambridge university press, 1999.

[DV83] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[Egg58] H. G. Eggleston. *Convexity*, volume 47 of *Tracts in Math and Math. Phys.* Cambridge University Press, 1958.

[Eli55] Peter Elias. Coding for noisy channels. *IRE Convention Record*, 3:37–46, 1955.

[Eli72] P. Elias. The efficient construction of an unbiased random sequence. *Annals of Mathematical Statistics*, 43(3):865–870, 1972.

[ELZ05] Uri Erez, Simon Litsyn, and Ram Zamir. Lattices which are good for (almost) everything. *IEEE Transactions on Information Theory*, 51(10):3401–3416, Oct. 2005.

[ES03] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[EZ04] U. Erez and R. Zamir. Achieving $\frac{1}{2}\log(1+\mathrm{SNR})$ on the AWGN channel with lattice encoding and decoding. *IEEE Trans. Inf. Theory*, IT-50:2293–2314, Oct. 2004.

[Fel70] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. Wiley, New York, third edition, 1970.

[Fer67] T. S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York, NY, 1967.

[FHT03] A. A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *Information Theory, IEEE Transactions on*, 49(6):1491–1498, Jun. 2003.

[FJ89] G.D. Forney Jr. Multidimensional constellations. II. Voronoi constellations. *IEEE Journal on Selected Areas in Communications*, 7(6):941–958, Aug 1989.

[FK98] Ehud Friedgut and Jeff Kahn. On the number of copies of one hypergraph in another. *Israel J. Math.*, 105:251–256, 1998.

[FMG92] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Trans. Inf. Theory*, 38(4):1258–1270, 1992.

[Gil10] Gustavo L Gilardoni. On pinsker's and vajda's type inequalities for csiszár's-divergences. *Information Theory, IEEE Transactions on*, 56(11):5377–5386, 2010.

[GKY56] I. M. Gel'fand, A. N. Kolmogorov, and A. M. Yaglom. On the general definition of the amount of information. *Dokl. Akad. Nauk. SSSR*, 11:745–748, 1956.

[GL95a] R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

[GL95b] Richard D Gill and Boris Y Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, pages 59–79, 1995.

[Ham50] John M Hammersley. On estimating restricted parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(2):192–240, 1950.

[Har] Sergiu Hart. Overweight puzzle. http://www.ma.huji.ac.il/~hart/puzzle/overweight.html.

[Hoe65] Wassily Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, pages 369–401, 1965.

[HV11] P. Harremoës and I. Vajda. On pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory*, 57(6):3230–3235, Jun. 2011.

[HWX17] B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. *IEEE Trans. on Information Theory*, 63(8):4729 – 4745, 2017.

[IK81] I. A. Ibragimov and R. Z. Khas'minskiĭ. *Statistical Estimation: Asymptotic Theory*. Springer, 1981.

[IS03] Y. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer, New York, NY, 2003.

[Joh11] I.M. Johnstone. *Gaussian estimation: Sequence and wavelet models*. 2011. Available at http://www-stat.stanford.edu/~imj/.

[KF+09] Christof Külske, Marco Formentin, et al. A symmetric entropy bound on the non-reconstruction regime of markov chains on galton-watson trees. *Electronic Communications in Probability*, 14:587–596, 2009.

[KO94] M.S. Keane and G.L. O'Brien. A Bernoulli factory. *ACM Transactions on Modeling and Computer Simulation*, 4(2):213–219, 1994.

[Kos63] VN Koshelev. Quantization with minimal entropy. *Probl. Pered. Inform*, 14:151–156, 1963.

[KS14] Oliver Kosut and Lalitha Sankar. Asymptotics and non-asymptotics for universal fixed-to-variable source coding. *arXiv preprint arXiv:1412.4444*, 2014.

[KT59] A.N. Kolmogorov and V.M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959. reprinted in.

[KV14] Ioannis Kontoyiannis and Sergio Verdú. Optimal lossless data compression: Non-asymptotics and asymptotics. *IEEE Trans. Inf. Theory*, 60(2):777–795, 2014.

[LC86] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, NY, 1986.

[LM03] Amos Lapidoth and Stefan M Moser. Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels. *IEEE Transactions on Information Theory*, 49(10):2426–2467, 2003.

[Loe97] Hans-Andrea Loeliger. Averaging bounds for lattices and linear codes. *IEEE Transactions on Information Theory*, 43(6):1767–1773, Nov. 1997.

[LZ94] Tamás Linder and Ram Zamir. On the asymptotic tightness of the Shannon lower bound. *IEEE Trans. Inf. Theory*, 40(6):2026 – 2031, Nov. 1994.

[Mas74] James Massey. On the fractional weight of distinct binary n-tuples (corresp.). *IEEE Transactions on Information Theory*, 20(1):131–131, 1974.

[MF98] Neri Merhav and Meir Feder. Universal prediction. *IEEE Trans. Inf. Theory*, 44(6):2124–2147, 1998.

[MP05] Elchanan Mossel and Yuval Peres. New coins from old: computing with unknown bias. *Combinatorica*, 25(6):707–724, 2005.

[MT10] Mokshay Madiman and Prasad Tetali. Information inequalities for joint distributions, with interpretations and applications. *IEEE Trans. Inf. Theory*, 56(6):2699–2713, 2010.

[OE15] O. Ordentlich and U. Erez. A simple proof for the existence of "good" pairs of nested lattices. *IEEE Transactions on Information Theory*, Submitted Aug. 2015.

[OPS48] BM Oliver, JR Pierce, and CE Shannon. The philosophy of pcm. *Proceedings of the IRE*, 36(11):1324–1331, 1948.

[Per92] Yuval Peres. Iterating von Neumann's procedure for extracting random bits. *Annals of Statistics*, 20(1):590–597, 1992.

[Pis99] G. Pisier. *The volume of convex bodies and Banach space geometry*. Cambridge University Press, 1999.

[PPV10a] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[PPV10b] Y. Polyanskiy, H. V. Poor, and S. Verdú. Feedback in the non-asymptotic regime. *IEEE Trans. Inf. Theory*, April 2010. submitted for publication.

[PPV11] Y. Polyanskiy, H. V. Poor, and S. Verdú. Minimum energy to send k bits with and without feedback. *IEEE Trans. Inf. Theory*, 57(8):4880–4902, August 2011.

[PV10] Y. Polyanskiy and S. Verdú. Arimoto channel coding converse and Rényi divergence. In *Proceedings of the Forty-eighth Annual Allerton Conference on Communication, Control, and Computing*, pages 1327–1333, 2010.

[PW14] Y. Polyanskiy and Y. Wu. Peak-to-average power ratio of good codes for Gaussian channel. *IEEE Trans. Inf. Theory*, 60(12):7655–7660, December 2014.

[PW17] Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and Bayesian networks. In Eric Carlen, Mokshay Madiman, and Elisabeth M. Werner, editors, *Convexity and Concentration. The IMA Volumes in Mathematics and its Applications, vol 161*, pages 211–249. Springer, New York, NY, 2017.

[Rad97] Jaikumar Radhakrishnan. An entropy proof of Bregman's theorem. *J. Combin. Theory Ser. A*, 77(1):161–164, 1997.

[Ree65] Alec H Reeves. The past present and future of PCM. *IEEE Spectrum*, 2(5):58–62, 1965.

[Rin76] Yosef Rinott. On convexity of measures. *Annals of Probability*, 4(6):1020–1026, 1976.

[RSU01] Thomas J. Richardson, Mohammad Amin Shokrollahi, and Rüdiger L. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *IEEE Transactions on Information Theory*, 47(2):619–637, 2001.

[RU96] Bixio Rimoldi and Rüdiger Urbanke. A rate-splitting approach to the gaussian multiple-access channel. *Information Theory, IEEE Transactions on*, 42(2):364–375, 1996.

[RZ86] Ryabko B. Reznikova Zh. Analysis of the language of ants by information-theoretical methods. *Problemi Peredachi Informatsii*, 22(3):103–108, 1986. English translation: http://reznikova.net/R-R-entropy-09.pdf.

[SF11] Ofer Shayevitz and Meir Feder. Optimal feedback communication via posterior matching. *IEEE Trans. Inf. Theory*, 57(3):1186–1222, 2011.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, July/October 1948.

[Sio58] Maurice Sion. On general minimax theorems. *Pacific J. Math*, 8(1):171–176, 1958.

[Smi71] J. G. Smith. The information capacity of amplitude and variance-constrained scalar Gaussian channels. *Information and Control*, 18:203 – 219, 1971.

[Spe15] Spectre. SPECTRE: Short packet communication toolbox. https://github.com/yp-mit/spectre, 2015. GitHub repository.

[Spi96] Daniel A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, 42(6):1723–1731, 1996.

[Spi97] Daniel A. Spielman. The complexity of error-correcting codes. In *Fundamentals of Computation Theory*, pages 67–84. Springer, 1997.

[Str65] V. Strassen. The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 36(2):423–439, 1965.

[SV11] Wojciech Szpankowski and Sergio Verdú. Minimum expected length of fixed-to-variable lossless compression without prefix constraints. *IEEE Trans. Inf. Theory*, 57(7):4017–4025, 2011.

[SV16] Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[TE97] Giorgio Taricco and Michele Elia. Capacity of fading channel with no side information. *Electronics Letters*, 33(16):1368–1370, 1997.

[Top00] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

[Tre68] Harry L. Van Trees. Detection, estimation, and modulation theory. I, 1968.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[TV05] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[UR98] R. Urbanke and B. Rimoldi. Lattice codes can achieve capacity on the AWGN channel. *IEEE Transactions on Information Theory*, 44(1):273–278, 1998.

[Vaj70] Igor Vajda. Note on discrimination information and variation (corresp.). *IEEE Transactions on Information Theory*, 16(6):771–773, 1970.

[vdV02] Aad van der Vaart. The statistical work of Lucien Le Cam. *Annals of Statistics*, pages 631–682, 2002.

[VEH14] Tim Van Erven and Peter Harremoës. Rényi divergence and kullback-leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014.

[Ver07] S. Verdú. *EE528–Information Theory, Lecture Notes*. Princeton Univ., Princeton, NJ, 2007.

[Vit61] Anatoliĭ Georgievich Vitushkin. *Theory of the Transmission and Processing of Information*. Pergamon Press, 1961.

[vN51] J. von Neumann. Various techniques used in connection with random digits. *Monte Carlo Method, National Bureau of Standards, Applied Math Series*, (12):36–38, 1951.

[Yek04] Sergey Yekhanin. Improved upper bound for the redundancy of fix-free codes. *IEEE Trans. Inf. Theory*, 50(11):2815–2818, 2004.

[Yos03] Nobuyuki Yoshigahara. *Puzzles 101: A Puzzlemaster's Challenge*. A K Peters, Natick, MA, USA, 2003.

[Zam14] Ram Zamir. *Lattice Coding for Signals and Networks*. Cambridge University Press, Cambridge, 2014.

[ZY97] Zhen Zhang and Raymond W Yeung. A non-Shannon-type conditional inequality of information quantities. *IEEE Trans. Inf. Theory*, 43(6):1982–1986, 1997.

[ZY98] Zhen Zhang and Raymond W Yeung. On characterization of entropy function via information inequalities. *IEEE Trans. Inf. Theory*, 44(4):1440–1452, 1998.