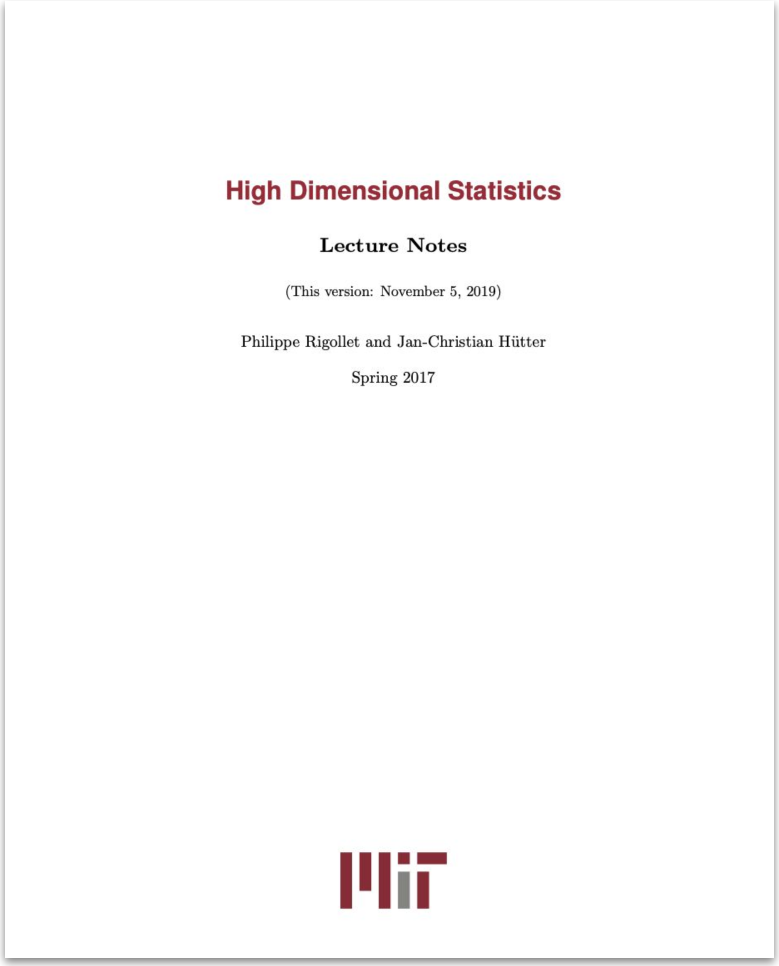
High-Dimensional Probability and Statistics

MATH/STAT/ECE 888: Topics in Mathematical Data Science
Sebastien Roch (Math+Stat)
UW-Madison
Fall 2021

Lecture 25 (11/03/21)

Today's slides based on Rigollet's notes and Wainwright



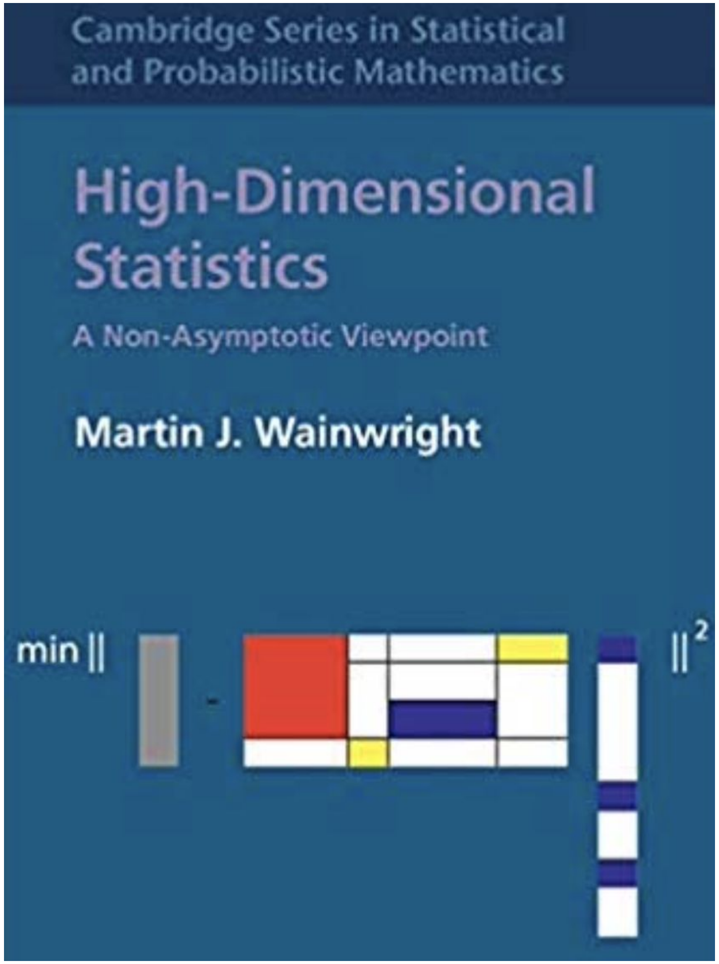
High Dimensional Statistics

Lecture Notes

(This version: November 5, 2019)

Philippe Rigollet and Jan-Christian Hütter

Spring 2017



Cambridge Series in Statistical and Probabilistic Mathematics

High-Dimensional Statistics

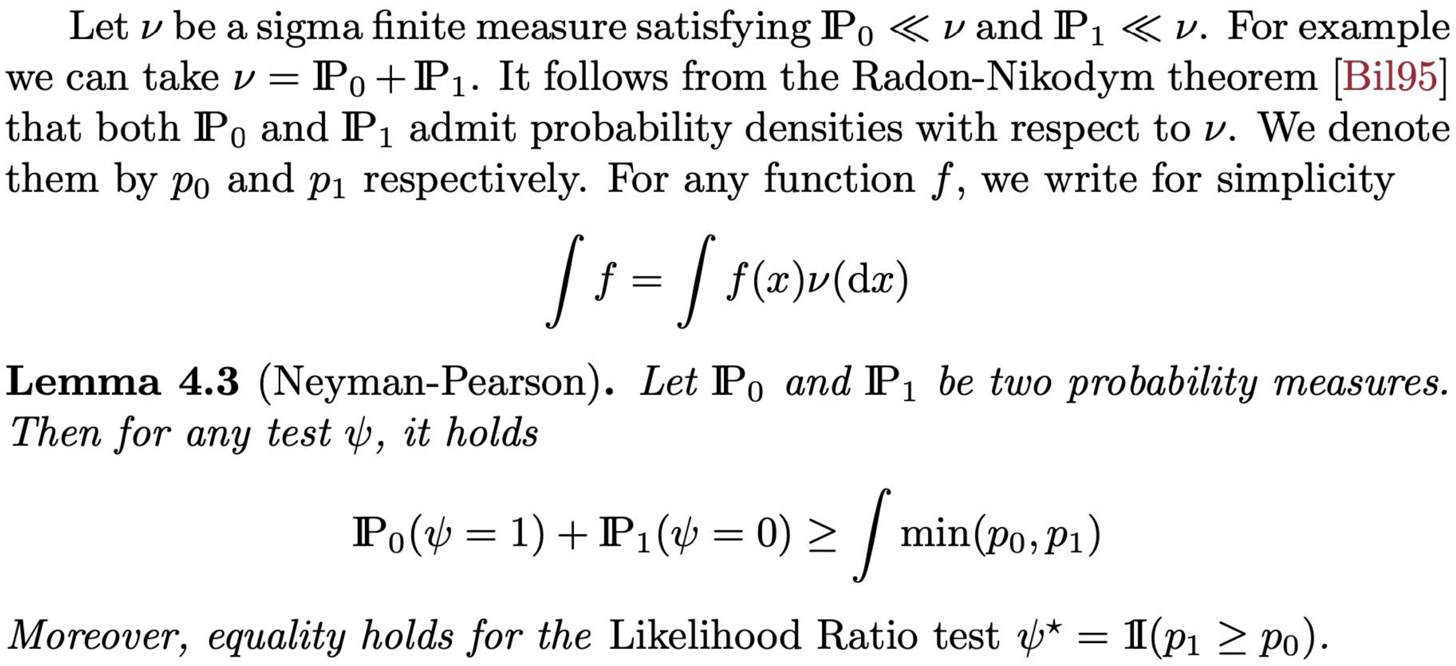
A Non-Asymptotic Viewpoint

Martin J. Wainwright

http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf

Review: Total variation (TV) distance

A lower bound [taken from Rigollet's notes]



Let ν be a sigma finite measure satisfying $\mathbb{P}_0 \ll \nu$ and $\mathbb{P}_1 \ll \nu$. For example we can take $\nu = \mathbb{P}_0 + \mathbb{P}_1$. It follows from the Radon-Nikodym theorem [Bil95] that both $\mathbb{P}_0$ and $\mathbb{P}_1$ admit probability densities with respect to ν. We denote them by p_0 and p_1 respectively. For any function f, we write for simplicity

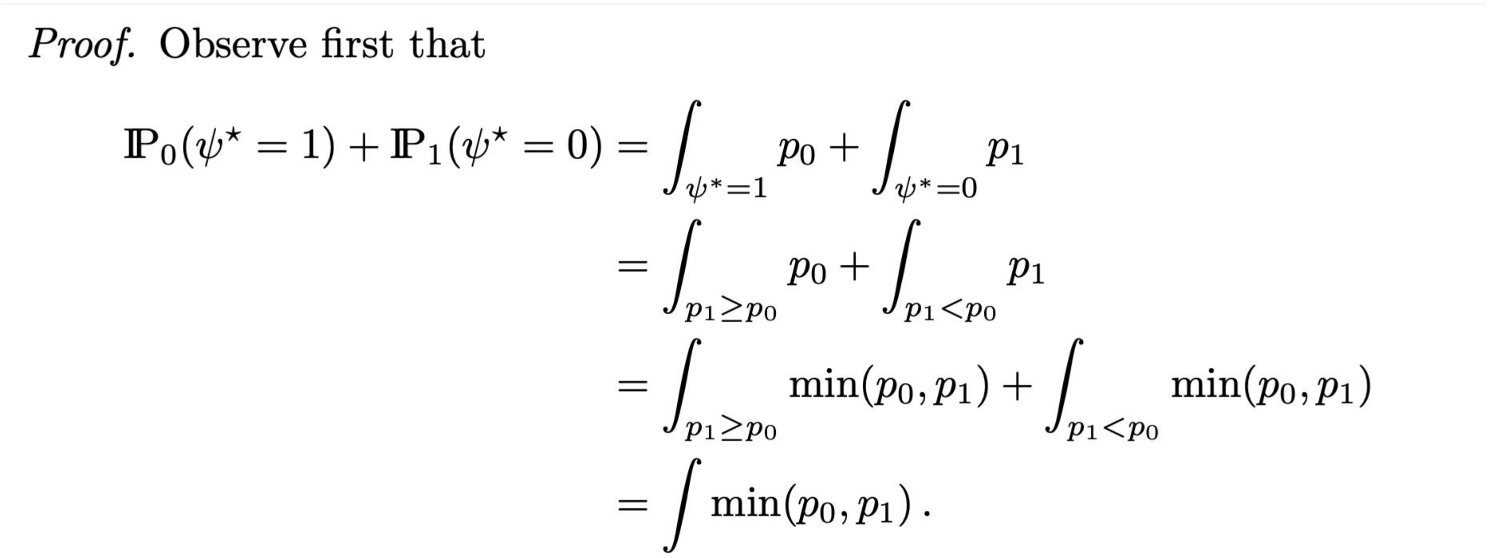
$$\int f = \int f(x)\nu(\mathrm{d}x)$$

Lemma 4.3 (Neyman-Pearson). *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures. Then for any test ψ, it holds*

$$\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \geq \int \min(p_0, p_1)$$

Moreover, equality holds for the Likelihood Ratio test $\psi^\star = \mathbb{1}(p_1 \geq p_0)$.

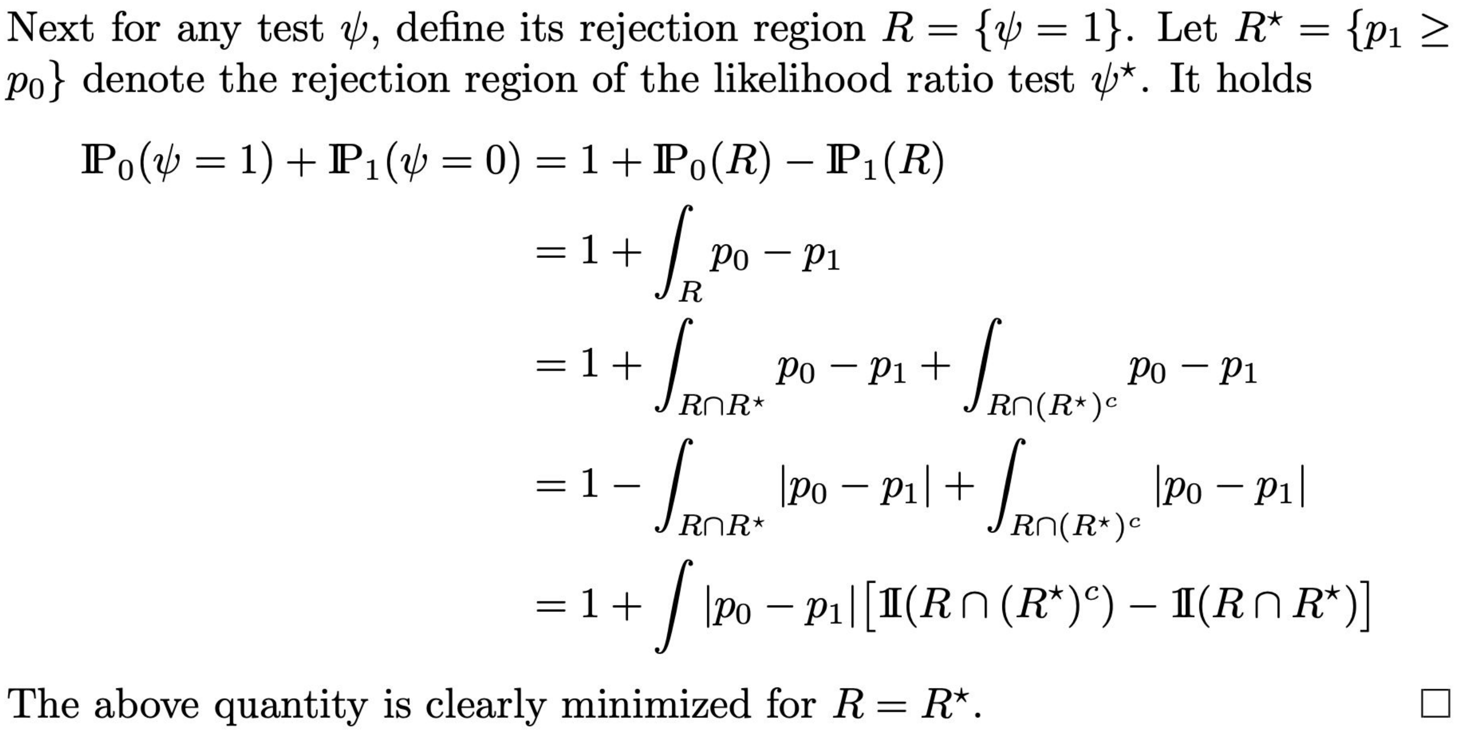
A lower bound cont'd [taken from Rigollet's notes]



Proof. Observe first that

$$
\begin{aligned}
\mathbb{P}_0(\psi^\star = 1) + \mathbb{P}_1(\psi^\star = 0) &= \int_{\psi^* = 1} p_0 + \int_{\psi^* = 0} p_1 \\
&= \int_{p_1 \geq p_0} p_0 + \int_{p_1 < p_0} p_1 \\
&= \int_{p_1 \geq p_0} \min(p_0, p_1) + \int_{p_1 < p_0} \min(p_0, p_1) \\
&= \int \min(p_0, p_1)\,.
\end{aligned}
$$

A lower bound cont'd [taken from Rigollet's notes]

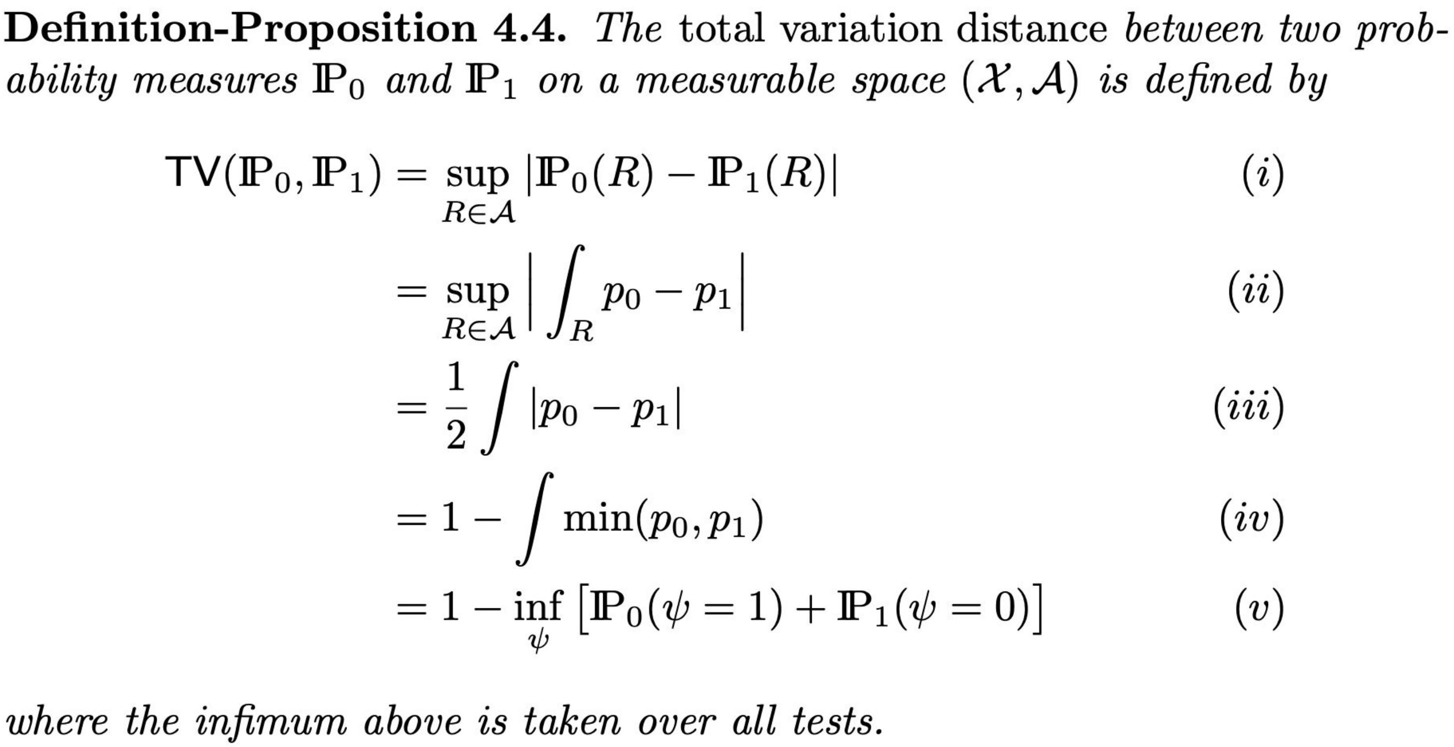


Next for any test ψ, define its rejection region $R = \{\psi = 1\}$. Let $R^\star = \{p_1 \geq p_0\}$ denote the rejection region of the likelihood ratio test $\psi^\star$. It holds

$$\begin{aligned}
\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) &= 1 + \mathbb{P}_0(R) - \mathbb{P}_1(R) \\
&= 1 + \int_R p_0 - p_1 \\
&= 1 + \int_{R \cap R^\star} p_0 - p_1 + \int_{R \cap (R^\star)^c} p_0 - p_1 \\
&= 1 - \int_{R \cap R^\star} |p_0 - p_1| + \int_{R \cap (R^\star)^c} |p_0 - p_1| \\
&= 1 + \int |p_0 - p_1| \big[\mathbb{1}(R \cap (R^\star)^c) - \mathbb{1}(R \cap R^\star) \big]
\end{aligned}$$

The above quantity is clearly minimized for $R = R^\star$. □

Aside: total variation distance [taken from Rigollet's notes]

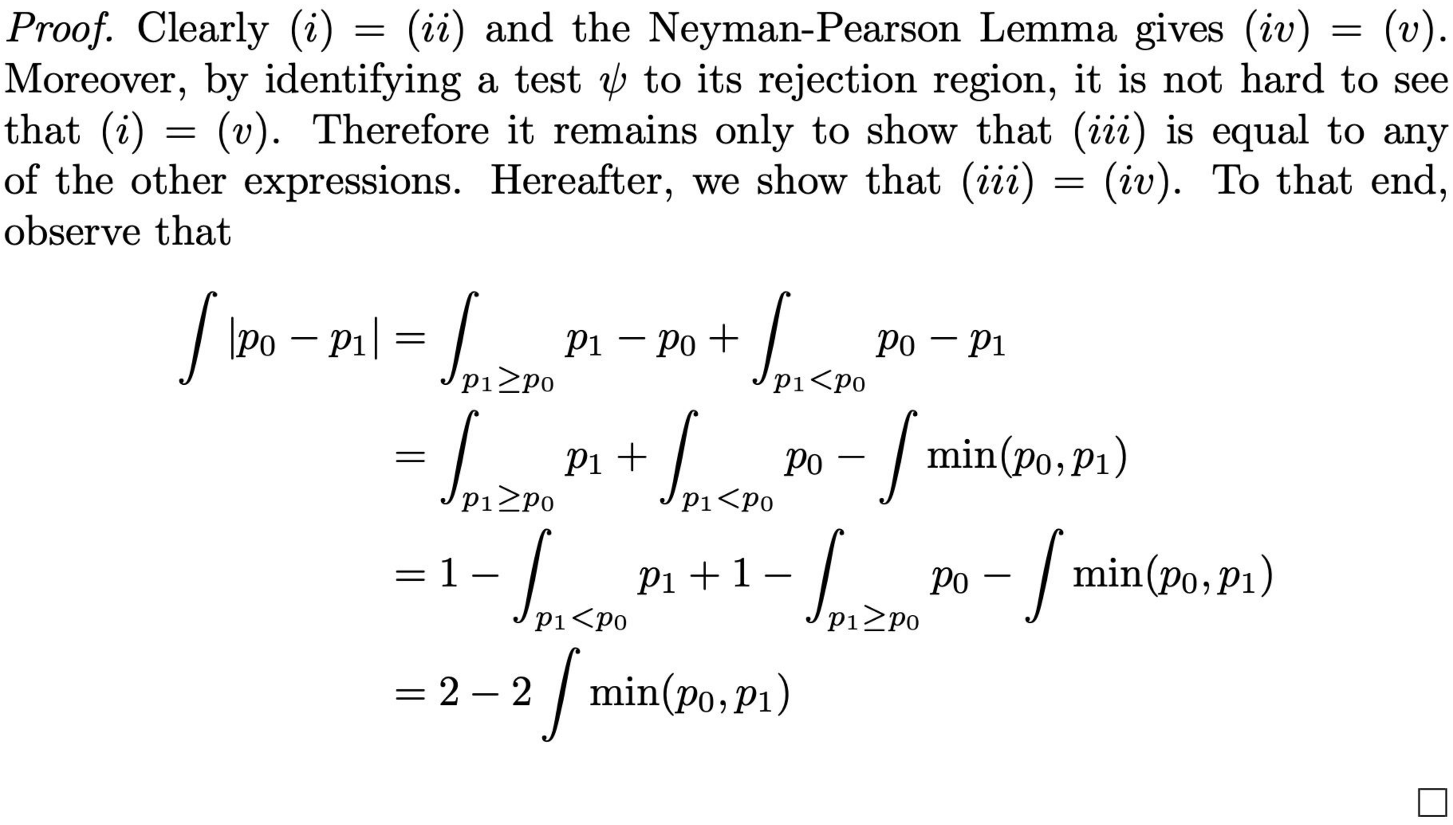


Definition-Proposition 4.4. *The* total variation distance *between two probability measures* $\mathbb{P}_0$ *and* $\mathbb{P}_1$ *on a measurable space* $(\mathcal{X}, \mathcal{A})$ *is defined by*

$$\begin{aligned}
\mathsf{TV}(\mathbb{P}_0, \mathbb{P}_1) &= \sup_{R \in \mathcal{A}} |\mathbb{P}_0(R) - \mathbb{P}_1(R)| && (i)\\
&= \sup_{R \in \mathcal{A}} \Big| \int_R p_0 - p_1 \Big| && (ii)\\
&= \frac{1}{2} \int |p_0 - p_1| && (iii)\\
&= 1 - \int \min(p_0, p_1) && (iv)\\
&= 1 - \inf_{\psi} \big[\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0) \big] && (v)
\end{aligned}$$

where the infimum above is taken over all tests.

Aside: total variation distance [taken from Rigollet's notes]



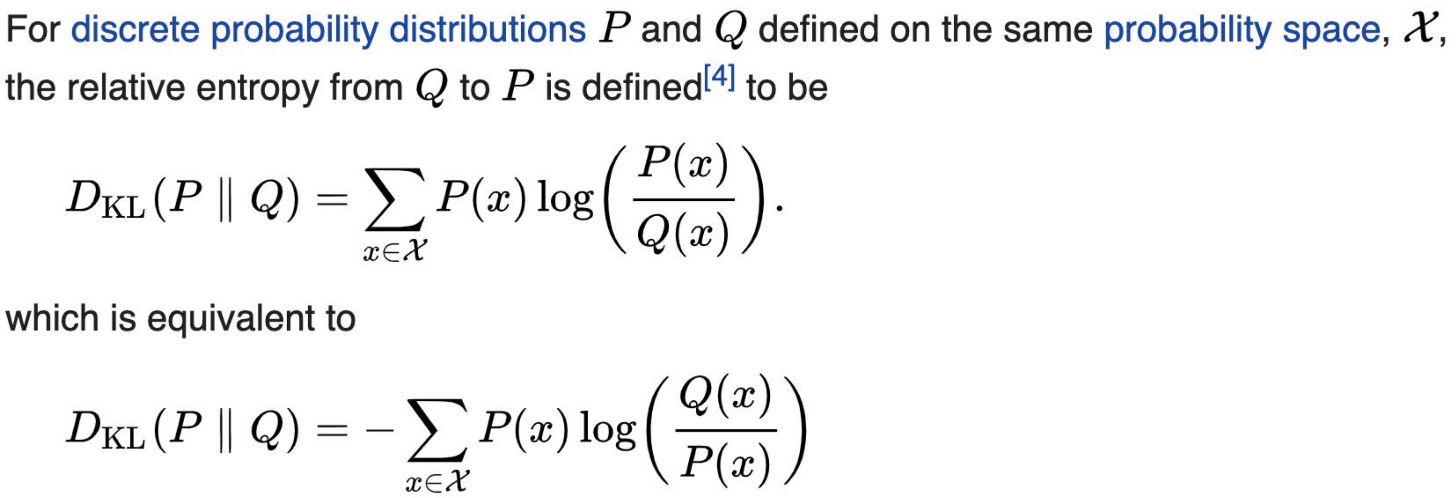
Proof. Clearly $(i) = (ii)$ and the Neyman-Pearson Lemma gives $(iv) = (v)$. Moreover, by identifying a test ψ to its rejection region, it is not hard to see that $(i) = (v)$. Therefore it remains only to show that (iii) is equal to any of the other expressions. Hereafter, we show that $(iii) = (iv)$. To that end, observe that

$$\begin{aligned}
\int |p_0 - p_1| &= \int_{p_1 \geq p_0} p_1 - p_0 + \int_{p_1 < p_0} p_0 - p_1 \\
&= \int_{p_1 \geq p_0} p_1 + \int_{p_1 < p_0} p_0 - \int \min(p_0, p_1) \\
&= 1 - \int_{p_1 < p_0} p_1 + 1 - \int_{p_1 \geq p_0} p_0 - \int \min(p_0, p_1) \\
&= 2 - 2 \int \min(p_0, p_1)
\end{aligned}$$

□

Kullback-Liebler (KL) divergence

Kullback–Leibler divergence (from Wikipedia)



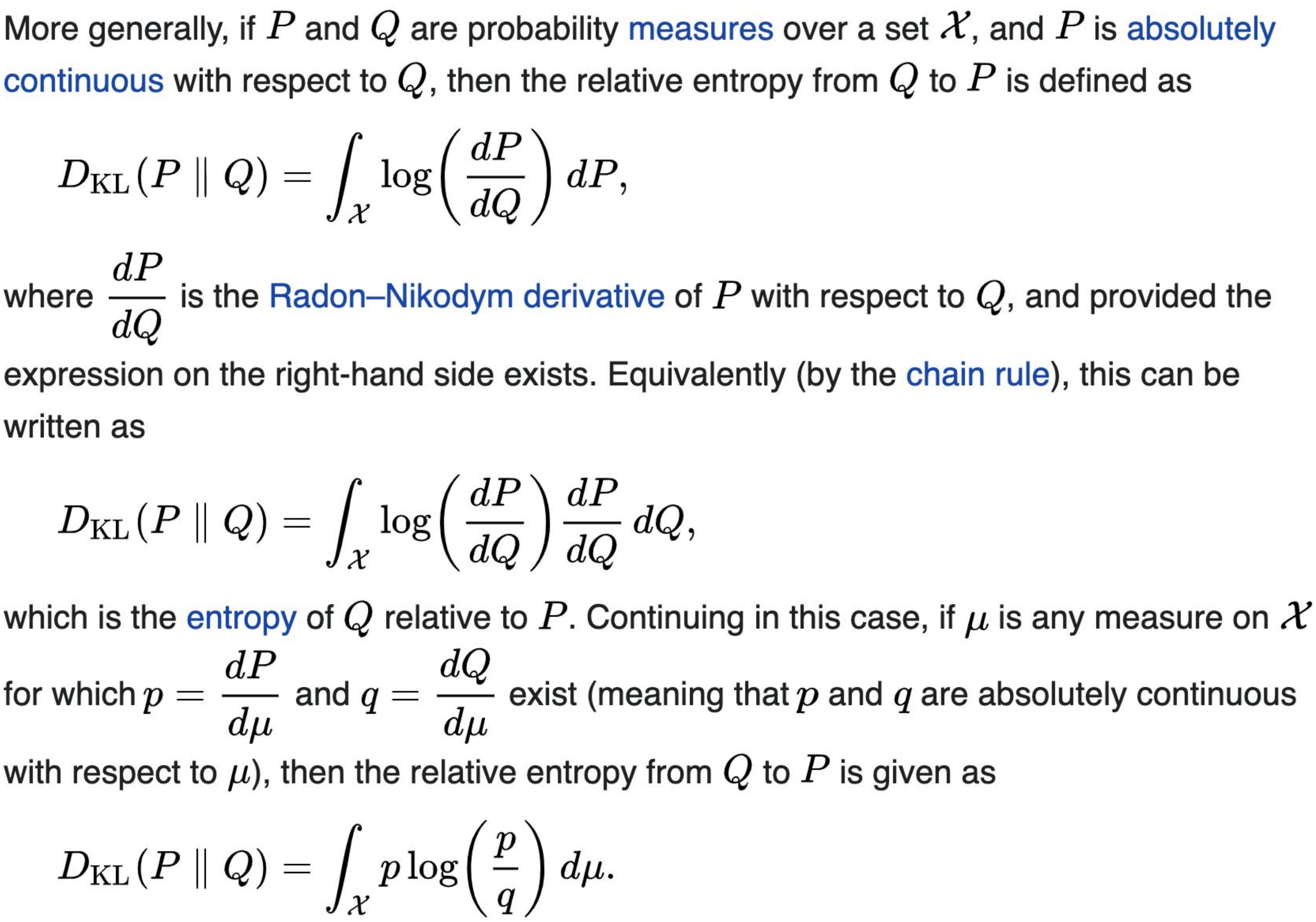
For discrete probability distributions P and Q defined on the same probability space, $\mathcal{X}$, the relative entropy from Q to P is defined[4] to be

$$D_{\text{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

which is equivalent to

$$D_{\text{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

Kullback–Leibler divergence cont'd (from Wikipedia)



More generally, if P and Q are probability measures over a set $\mathcal{X}$, and P is absolutely continuous with respect to Q, then the relative entropy from Q to P is defined as

$$D_{\text{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP,$$

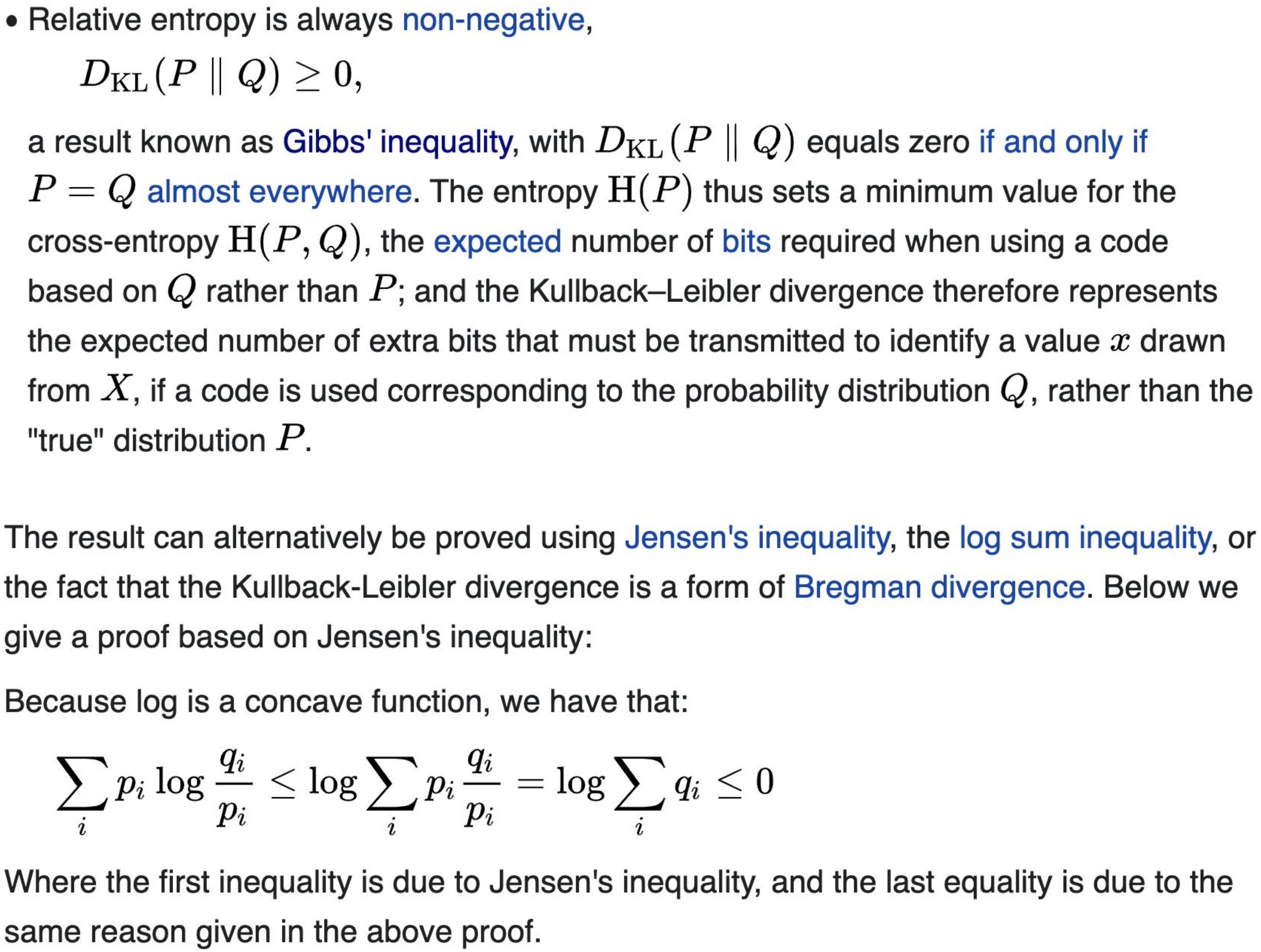
where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative of P with respect to Q, and provided the expression on the right-hand side exists. Equivalently (by the chain rule), this can be written as

$$D_{\text{KL}}(P \parallel Q) = \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) \frac{dP}{dQ}\, dQ,$$

which is the entropy of Q relative to P. Continuing in this case, if μ is any measure on $\mathcal{X}$ for which $p = \frac{dP}{d\mu}$ and $q = \frac{dQ}{d\mu}$ exist (meaning that p and q are absolutely continuous with respect to μ), then the relative entropy from Q to P is given as

$$D_{\text{KL}}(P \parallel Q) = \int_{\mathcal{X}} p \log\left(\frac{p}{q}\right) d\mu.$$

Kullback–Leibler divergence cont’d (from Wikipedia)



- Relative entropy is always non-negative,

  $$D_{\mathrm{KL}}(P \parallel Q) \geq 0,$$

  a result known as Gibbs' inequality, with $D_{\mathrm{KL}}(P \parallel Q)$ equals zero if and only if $P = Q$ almost everywhere. The entropy $\mathrm{H}(P)$ thus sets a minimum value for the cross-entropy $\mathrm{H}(P, Q)$, the expected number of bits required when using a code based on Q rather than P; and the Kullback–Leibler divergence therefore represents the expected number of extra bits that must be transmitted to identify a value x drawn from X, if a code is used corresponding to the probability distribution Q, rather than the "true" distribution P.

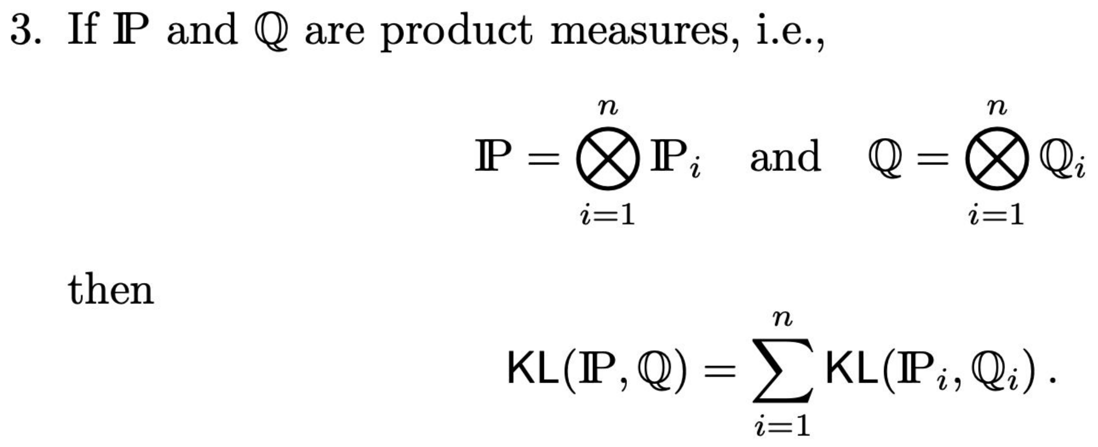
The result can alternatively be proved using Jensen's inequality, the log sum inequality, or the fact that the Kullback-Leibler divergence is a form of Bregman divergence. Below we give a proof based on Jensen's inequality:

Because log is a concave function, we have that:

$$\sum_i p_i \log \frac{q_i}{p_i} \leq \log \sum_i p_i \frac{q_i}{p_i} = \log \sum_i q_i \leq 0$$

Where the first inequality is due to Jensen's inequality, and the last equality is due to the same reason given in the above proof.

Product measures



3. If $\mathbb{P}$ and $\mathbb{Q}$ are product measures, i.e.,

$$\mathbb{P} = \bigotimes_{i=1}^{n} \mathbb{P}_i \quad \text{and} \quad \mathbb{Q} = \bigotimes_{i=1}^{n} \mathbb{Q}_i$$

then

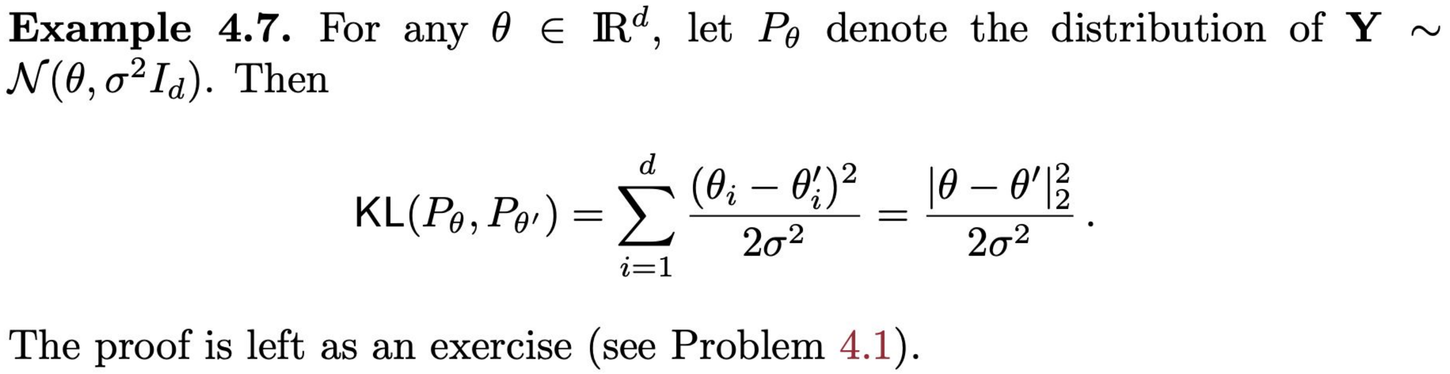
$$\mathsf{KL}(\mathbb{P}, \mathbb{Q}) = \sum_{i=1}^{n} \mathsf{KL}(\mathbb{P}_i, \mathbb{Q}_i)\,.$$

Product measures cont'd

3. Note that if $X = (X_1, \ldots, X_n)$,

$$\begin{aligned}\mathsf{KL}(\mathbb{P}, \mathbb{Q}) &= \mathbb{E} \log\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(X)\right)\\ &= \sum_{i=1}^{n} \int \log\left(\frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\mathbb{Q}_i}(X_i)\right) \mathrm{d}\mathbb{P}_1(X_1) \cdots \mathrm{d}\mathbb{P}_n(X_n)\\ &= \sum_{i=1}^{n} \int \log\left(\frac{\mathrm{d}\mathbb{P}_i}{\mathrm{d}\mathbb{Q}_i}(X_i)\right) \mathrm{d}\mathbb{P}_i(X_i)\\ &= \sum_{i=1}^{n} \mathsf{KL}(\mathbb{P}_i, \mathbb{Q}_i)\end{aligned}$$

Example 4.7

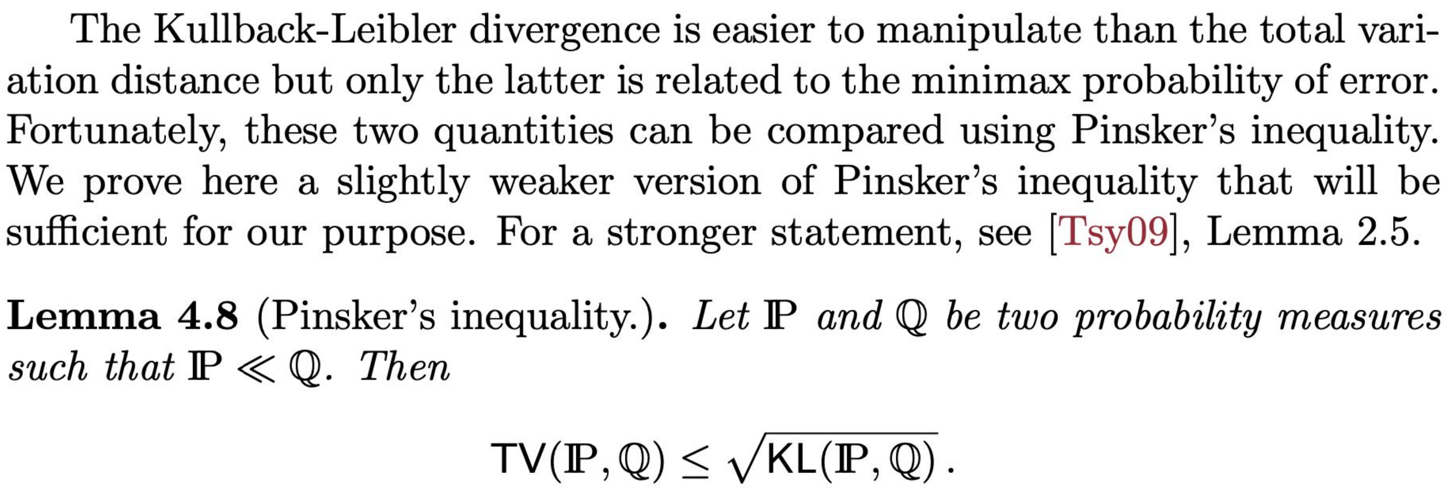


Example 4.7. For any $\theta \in \mathbb{R}^d$, let P_θ denote the distribution of $\mathbf{Y} \sim \mathcal{N}(\theta, \sigma^2 I_d)$. Then

$$\mathsf{KL}(P_\theta, P_{\theta'}) = \sum_{i=1}^{d} \frac{(\theta_i - \theta_i')^2}{2\sigma^2} = \frac{|\theta - \theta'|_2^2}{2\sigma^2} .$$

The proof is left as an exercise (see Problem 4.1).

Pinsker



The Kullback-Leibler divergence is easier to manipulate than the total variation distance but only the latter is related to the minimax probability of error. Fortunately, these two quantities can be compared using Pinsker's inequality. We prove here a slightly weaker version of Pinsker's inequality that will be sufficient for our purpose. For a stronger statement, see [Tsy09], Lemma 2.5.

Lemma 4.8 (Pinsker's inequality.)**.** *Let* $\mathbb{P}$ *and* $\mathbb{Q}$ *be two probability measures such that* $\mathbb{P} \ll \mathbb{Q}$*. Then*

$$\mathsf{TV}(\mathbb{P}, \mathbb{Q}) \leq \sqrt{\mathsf{KL}(\mathbb{P}, \mathbb{Q})}\,.$$

Pinsker cont'd

Proof. Note that

$$\begin{aligned}
\mathsf{KL}(\mathbb{P},\mathbb{Q}) &= \int_{pq>0} p\log\left(\frac{p}{q}\right)\\
&= -2\int_{pq>0} p\log\left(\sqrt{\frac{q}{p}}\right)\\
&= -2\int_{pq>0} p\log\left(\left[\sqrt{\frac{q}{p}}-1\right]+1\right)\\
&\geq -2\int_{pq>0} p\left[\sqrt{\frac{q}{p}}-1\right] \qquad \text{(by Jensen)}\\
&= 2-2\int\sqrt{pq}
\end{aligned}$$

Pinsker cont’d

Next, note that

$$\begin{aligned}\Big(\int \sqrt{pq}\Big)^2 &= \Big(\int \sqrt{\max(p,q)\min(p,q)}\Big)^2\\ &\le \int \max(p,q) \int \min(p,q) \qquad \text{(by Cauchy-Schwarz)}\\ &= \big[2 - \int \min(p,q)\big] \int \min(p,q)\\ &= \big(1 + \mathsf{TV}(\mathbb{P},\mathbb{Q})\big)\big(1 - \mathsf{TV}(\mathbb{P},\mathbb{Q})\big)\\ &= 1 - \mathsf{TV}(\mathbb{P},\mathbb{Q})^2\end{aligned}$$

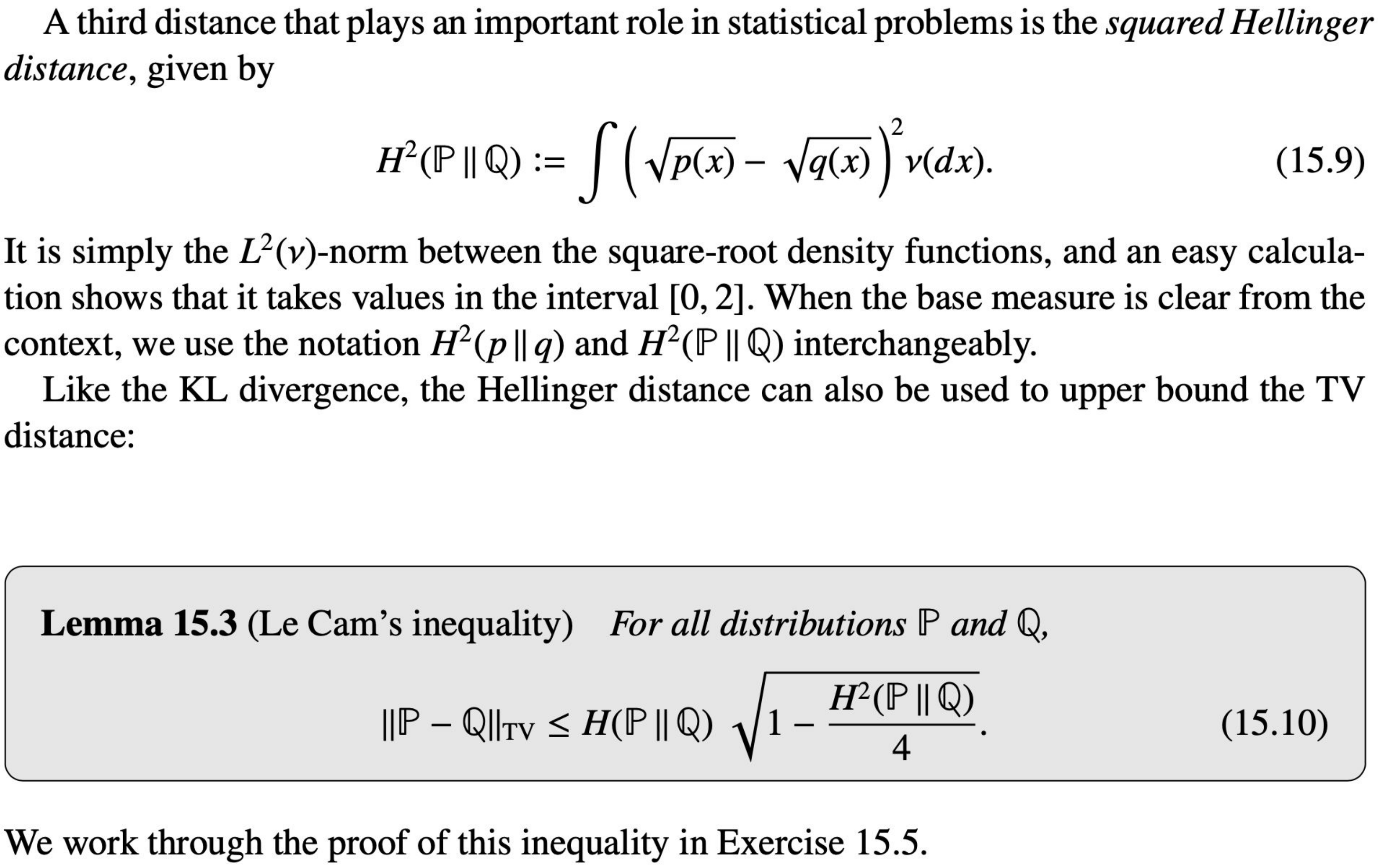
The two displays yield

$$\mathsf{KL}(\mathbb{P},\mathbb{Q}) \ge 2 - 2\sqrt{1 - \mathsf{TV}(\mathbb{P},\mathbb{Q})^2} \ge \mathsf{TV}(\mathbb{P},\mathbb{Q})^2\,,$$

where we used the fact that $0 \le \mathsf{TV}(\mathbb{P},\mathbb{Q}) \le 1$ and $\sqrt{1-x} \le 1 - x/2$ for $x \in [0,1]$. □

Hellinger distance

Squared Hellinger distance



A third distance that plays an important role in statistical problems is the *squared Hellinger distance*, given by

$$H^2(\mathbb{P} \,\|\, \mathbb{Q}) := \int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 \nu(dx). \tag{15.9}$$

It is simply the $L^2(\nu)$-norm between the square-root density functions, and an easy calculation shows that it takes values in the interval $[0, 2]$. When the base measure is clear from the context, we use the notation $H^2(p \,\|\, q)$ and $H^2(\mathbb{P} \,\|\, \mathbb{Q})$ interchangeably.

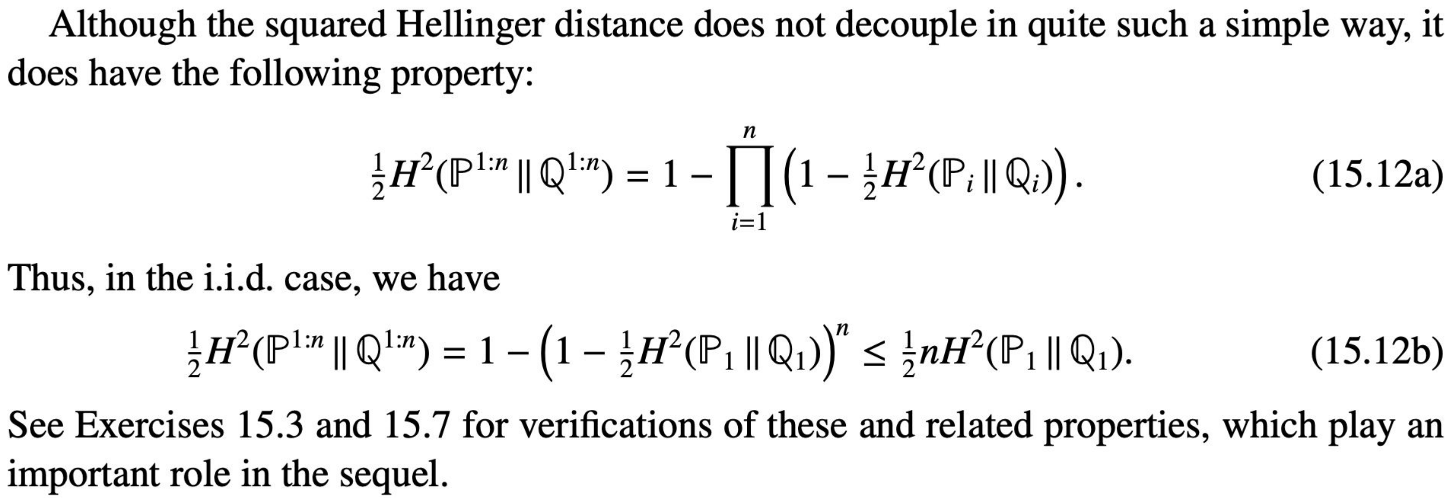
Like the KL divergence, the Hellinger distance can also be used to upper bound the TV distance:

Lemma 15.3 (Le Cam's inequality) *For all distributions* $\mathbb{P}$ *and* $\mathbb{Q}$*,*

$$\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} \leq H(\mathbb{P} \,\|\, \mathbb{Q}) \sqrt{1 - \frac{H^2(\mathbb{P} \,\|\, \mathbb{Q})}{4}}. \tag{15.10}$$

We work through the proof of this inequality in Exercise 15.5.

Squared Hellinger distance cont'd



Although the squared Hellinger distance does not decouple in quite such a simple way, it does have the following property:

$$\tfrac{1}{2}H^2(\mathbb{P}^{1:n} \,\|\, \mathbb{Q}^{1:n}) = 1 - \prod_{i=1}^{n}\Big(1 - \tfrac{1}{2}H^2(\mathbb{P}_i \,\|\, \mathbb{Q}_i)\Big). \tag{15.12a}$$

Thus, in the i.i.d. case, we have

$$\tfrac{1}{2}H^2(\mathbb{P}^{1:n} \,\|\, \mathbb{Q}^{1:n}) = 1 - \Big(1 - \tfrac{1}{2}H^2(\mathbb{P}_1 \,\|\, \mathbb{Q}_1)\Big)^n \le \tfrac{1}{2}nH^2(\mathbb{P}_1 \,\|\, \mathbb{Q}_1). \tag{15.12b}$$

See Exercises 15.3 and 15.7 for verifications of these and related properties, which play an important role in the sequel.