ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 5: Inequalities between f-divergences via their joint range

Lecturer: Yihong Wu *Scribe: Pengkun Yang, Feb 9, 2016 [Ed. Feb 28]*

In the last lecture we proved the Pinkser's inequality that $D(P\|Q) \geq 2d_{\mathrm{TV}}^2(P,Q)$ in an ad hoc manner. The downside of ad hoc approaches is that it is hard to tell whether those inequalities can be improved or not. However, the key step when we proved the Pinkser's inequality, reduction to the case for Bernoulli random variables, is inspiring: is it possible to reduce inequalities between any two f-divergences to the binary case?

5.1 Inequalities via joint range

A systematic method is to prove those inequalities via their joint range. For example, to prove a lower bound of $D(P\|Q)$ by a function of $d_{\mathrm{TV}}(P,Q)$ that $D(P\|Q) \geq F(d_{\mathrm{TV}}(P,Q))$ for some $F : [0,1] \mapsto [0,\infty]$, the best choice, by definition, is the following:

$$F(\epsilon) \triangleq \inf_{(P,Q):d_{\mathrm{TV}}(P,Q)=\epsilon} D(P\|Q).$$

The problem boils to the characterization of the region $\{(d_{\mathrm{TV}}(P,Q), D(P\|Q)) : P, Q\} \subseteq \mathbb{R}^2$, their joint range, whose lower boundary is the function F.

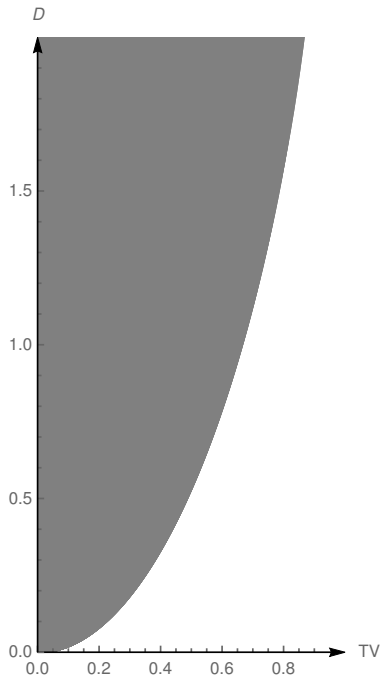

Figure 5.1: Joint range of d_{TV} and D.

Definition 5.1 (Joint range)**.** Consider two f-divergences $D_f(P\|Q)$ and $D_g(P\|Q)$. Their joint range is a subset of $\mathbb{R}^2$ defined by

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

The region $\mathcal{R}$ seems 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. The main theorem we will prove is the following [HV11]:

Theorem 5.1 (Harremoës-Vajda '11)**.**

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

It is easy to obtain a parametric formula of $\mathcal{R}_2$. By Theorem 5.1, the region $\mathcal{R}$ is no more than the convex hull of $\mathcal{R}_2$.

Theorem 5.1 implies that $\mathcal{R}$ is a convex set; however, as a warmup, it is instructive to prove convexity of $\mathcal{R}$ directly, which simply follows from the arbitrariness of the alphabet size of distributions. Given any two points $(D_f(P_0\|Q_0), D_g(P_0\|Q_0))$ and $(D_f(P_1\|Q_1), D_g(P_1\|Q_1))$ in the joint range, it is easy to construct another pair of distributions (P, Q) by alphabet extension that produces any convex combination of those two points.

Theorem 5.2. *$\mathcal{R}$ is convex.*

Proof. 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 define another pair of distributions (P, Q) on $\mathcal{X} \times \{0, 1\}$ by

$$P = \bar{\alpha}(P_0 \times \delta_0) + \alpha(P_1 \times \delta_1),$$
$$Q = \bar{\alpha}(Q_0 \times \delta_0) + \alpha(Q_1 \times \delta_1).$$

In other words, we construct a random variable $Z = (X, B)$ with $B \sim \mathrm{Bern}(\alpha)$, where $P_{X|B=i} = P_i$ and $Q_{X|B=i} = Q_i$. Then

$$D_f(P\|Q) = \mathbb{E}_Q\left[f\left(\frac{P}{Q}\right)\right] = \mathbb{E}_B\left[\mathbb{E}_{Q_{Z|B}}\left[f\left(\frac{P}{Q}\right)\right]\right] = \bar{\alpha}D_f(P_0\|Q_0) + \alpha D_f(P_1\|Q_1),$$
$$D_g(P\|Q) = \mathbb{E}_Q\left[g\left(\frac{P}{Q}\right)\right] = \mathbb{E}_B\left[\mathbb{E}_{Q_{Z|B}}\left[g\left(\frac{P}{Q}\right)\right]\right] = \bar{\alpha}D_g(P_0\|Q_0) + \alpha D_g(P_1\|Q_1).$$

Therefore, $\bar{\alpha}(D_f(P_0\|Q_0), D_g(P_0\|Q_0)) + \alpha(D_f(P_1\|Q_1), D_g(P_1\|Q_1)) \in \mathcal{R}$ and thus $\mathcal{R}$ is convex. □

Theorem 5.1 is proved by the following two lemmas:

Lemma 5.1 (non-constructive/existential)**.** $\mathcal{R} = \mathcal{R}_4$.

Lemma 5.2 (constructive/algorithmic)**.**

$$\mathcal{R}_{k+1} = \mathsf{co}(\mathcal{R}_2 \cup \mathcal{R}_k) \quad \textit{for any } k \geq 2$$

and hence

$$\mathcal{R}_k = \mathsf{co}(\mathcal{R}_2), \quad \textit{for any } k \geq 3.$$

5.1.1 Representation of f-divergences

To prove Lemma 5.1 and Lemma 5.2, we first express f-divergences by means of expectation over the likelihood ratio.

Lemma 5.3. *Given two f-divergences $D_f(\cdot\|\cdot)$ and $D_g(\cdot\|\cdot)$, their joint range is*

$$\mathcal{R} = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]) \end{pmatrix} : X \geq 0, \mathbb{E}[X] \leq 1 \right\},$$

$$\mathcal{R}_k = \left\{ \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]) \end{pmatrix} : \begin{array}{c} X \geq 0, \mathbb{E}[X] \leq 1, X \text{ takes at most } k-1 \text{ values}, \\ \text{or } X \geq 0, \mathbb{E}[X] = 1, X \text{ takes at most } k \text{ values} \end{array} \right\},$$

where $\tilde{f}(0) \triangleq \lim_{x\to 0} xf(1/x)$ and $\tilde{g}(0) \triangleq \lim_{x\to 0} xg(1/x)$.

In the statement of Lemma 5.3, we remark that $\tilde{f}(0)$ and $\tilde{g}(0)$ are both well-defined (possibly $+\infty$) by the convexity of $x \mapsto xf(1/x)$ and $x \mapsto xg(1/x)$ (from the last lecture).

Before proving above lemma, we look at the following two examples to understand the correspondence between a point in the joint range and a random variable. The first example is the simple case that $P \ll Q$, when the likelihood ratio of P and Q (or Radon-Nikodym derivative defined on the union of the spaces of P and Q) is well-define.

Example 5.1. Consider the following two distributions P, Q on $[3]$:

	1	2	3
P	0.34	0.34	0.32
Q	0.85	0.1	0.05

Then $D_f(P\|Q) = 0.85f(0.4) + 0.1f(3.4) + 0.05f(6.4)$, which is $\mathbb{E}[f(X)]$ where X is the likelihood ratio of P and Q taking 3 values with the following pmf:

x	0.4	3.4	6.4
$\mu(x)$	0.85	0.1	0.05

On the other direction, given the above pmf of a non-negative, unit-mean random variable $X \sim \mu$ that takes 3 values, we can construct a pair of distribution by $Q(x) = \mu(x)$ and $P(x) = x\mu(x)$.

In general cases P is not necessarily absolutely continuous w.r.t. Q, and the likelihood ratio X may not always exist. However, it is still well-defined on the event $\{Q > 0\}$.

Example 5.2. Consider the following two distributions P, Q on $[2]$:

	1	2
P	0.4	0.6
Q	0	1

Then $D_f(P\|Q) = f(0.6) + 0f(\frac{0.4}{0})$, where $0f(\frac{p}{0})$ is understood as

$$0f\left(\frac{p}{0}\right) = \lim_{x\to 0} xf\left(\frac{p}{x}\right) = p\lim_{x\to 0}\frac{x}{p}f\left(\frac{p}{x}\right) = p\tilde{f}(0).$$

Therefore $D_f(P\|Q) = f(0.6) + 0.4\tilde{f}(0) = \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X])$ where X is defined on $\{Q > 0\}$:

x	0.6
$\mu(x)$	1

On the other direction, given above pmf of a non-negative random variable $X \sim \mu$ with $\mathbb{E}[X] \le 1$ that takes 1 value, we let $Q(x) = \mu(x)$, let $P(x) = x\mu(x)$ on $\{Q > 0\}$ and let P have an extra point mass $1 - \mathbb{E}[X]$.

Proof of Lemma 5.3. We first prove it for $\mathcal{R}$. Given any pair of distributions (P, Q) that produces a point of $\mathcal{R}$, let p, q denote the densities of P, Q under some dominating measure μ, respectively. Let

$$X = \frac{p}{q} \text{ on } \{q > 0\}, \quad \mu_X = Q, \tag{5.1}$$

then $X \ge 0$ and $\mathbb{E}[X] = P[q > 0] \le 1$. Then

$$\begin{aligned} D_f(P\|Q) = \int_{\{q>0\}} f\left(\frac{p}{q}\right)\mathrm{d}Q + \int_{\{q=0\}}\frac{q}{p}f\left(\frac{p}{q}\right)\mathrm{d}P &= \int_{\{q>0\}} f\left(\frac{p}{q}\right)\mathrm{d}Q + \tilde{f}(0)P[q=0] \\ &= \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]), \end{aligned}$$

Analogously,

$$D_g(P\|Q) = \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X]),$$

On the other direction, given any random variable $X \ge 0$ and $\mathbb{E}[X] \le 1$ where $X \sim \mu$, let

$$\mathrm{d}Q = \mathrm{d}\mu, \quad \mathrm{d}P = X\mathrm{d}\mu + (1-\mathbb{E}[X])\delta_*, \tag{5.2}$$

where $*$ is an arbitrary symbol outside the support of X. Then

$$\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q)\end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1-\mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1-\mathbb{E}[X])\end{pmatrix}.$$

Now we consider $\mathcal{R}_k$. Given two probability measures P, Q on $[k]$, the likelihood ratio defined in (5.1) takes at most k values. If $P \ll Q$ then $\mathbb{E}[X] = 1$; if $P \not\ll Q$ then X takes at most $k-1$ values.

On the other direction, if $\mathbb{E}[X] = 1$ then the construction of P, Q in (5.2) are on the same support of X; if $\mathbb{E}[X] < 1$ then the support of P is increased by one. □

5.1.2 Proof of Theorem 5.1

Aside: Fenchel-Eggleston-Carathéodory's theorem: Let $S \subseteq \mathbb{R}^d$ and $x \in \mathsf{co}(S)$. Then there exists a set of $d+1$ points $S' = \{x_1, x_2, \ldots, x_{d+1}\} \in S$ such that $x \in \mathsf{co}(S')$. If S is connected, then d points are enough.

Proof of Lemma 5.1. It suffices to prove that

$$\mathcal{R} \subseteq \mathcal{R}_4.$$

Let $S \triangleq \{(x, f(x), g(x)) : x \geq 0\}$ which is a connected set. For any pair of distributions (P, Q) that produces a point of $\mathcal{R}$, we construct a random variable X as in (5.1), then $(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) \in \mathsf{co}(S)$. By Fenchel-Eggleston-Carathéodory's theorem,[1] there exists $(x_i, f(x_i), g(x_i))$ and the corresponding weight α_i for $i = 1, 2, 3$ such that

$$(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) = \sum_{i=1}^{3} \alpha_i (x_i, f(x_i), g(x_i)).$$

We construct another random variable X' that takes value x_i with probability α_i. Then X takes 3 values and

$$(\mathbb{E}[X], \mathbb{E}[f(X)], \mathbb{E}[g(X)]) = (\mathbb{E}[X'], \mathbb{E}[f(X')], \mathbb{E}[g(X')]). \tag{5.3}$$

By Lemma 5.3 and (5.3),

$$\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(X)] + \tilde{f}(0)(1 - \mathbb{E}[X]) \\ \mathbb{E}[g(X)] + \tilde{g}(0)(1 - \mathbb{E}[X]) \end{pmatrix} = \begin{pmatrix} \mathbb{E}[f(X')] + \tilde{f}(0)(1 - \mathbb{E}[X']) \\ \mathbb{E}[g(X')] + \tilde{g}(0)(1 - \mathbb{E}[X']) \end{pmatrix} \in \mathcal{R}_4.$$

□

We observe from Lemma 5.3 that $D_f(P\|Q)$ only depends on the distribution of X for some $X \geq 0$ and $\mathbb{E}[X] \leq 1$. To find a pair of distributions that produce a point in $\mathcal{R}_k$ it suffices to find a random variable $X \geq 0$ taking k values with $\mathbb{E}[X] = 1$, or taking $k - 1$ values with $\mathbb{E}[X] \leq 1$. In Example 5.1 where (P, Q) produces a point in $\mathcal{R}_3$, we want to show that it also belongs to $\mathsf{co}(\mathcal{R}_2)$. The decomposition of a point in $\mathcal{R}_3$ is equivalent to the decomposition of the likelihood ratio X that

$$\mu_X = \alpha\mu_1 + \bar{\alpha}\mu_2.$$

A solution of such decomposition is that $\mu_X = 0.5\mu_1 + 0.5\mu_2$ where μ_1, μ_2 has the following pmf:

x	0.4	3.4
$\mu_1(x)$	0.8	0.2

x	0.4	6.4
$\mu_2(x)$	0.9	0.1

Then by (5.2) we obtain two pairs of distributions

P_1	0.32	0.68
Q_1	0.8	0.2

P_2	0.36	0.64
Q_2	0.9	0.1

We obtain that

$$\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q) \end{pmatrix} = 0.5 \begin{pmatrix} D_f(P_1\|Q_1) \\ D_g(P_1\|Q_1) \end{pmatrix} + 0.5 \begin{pmatrix} D_f(P_2\|Q_2) \\ D_g(P_2\|Q_2) \end{pmatrix}.$$

[1] To prove Theorem 5.1, it suffices to invoke the basic Carathéodory's theorem, which proves a weaker version of Lemma 5.1 that $\mathcal{R} = \mathcal{R}_5$.

Proof of Lemma 5.2. It suffices to prove the first statement, namely, $\mathcal{R}_{k+1} = \mathsf{co}(\mathcal{R}_2 \cup \mathcal{R}_k)$ for any $k \geq 2$. By the same argument as in the proof of Theorem 5.2 we have $\mathsf{co}(\mathcal{R}_k) \subseteq \mathcal{R}_{k+1}$ and note that $\mathcal{R}_2 \cup \mathcal{R}_k = \mathcal{R}_k$. We only need to prove that

$$\mathcal{R}_{k+1} \subseteq \mathsf{co}(\mathcal{R}_2 \cup \mathcal{R}_k).$$

Given any pair of distributions (P, Q) that produces a point of $(D_f(P\|Q), D_g(P\|Q)) \in \mathcal{R}_{k+1}$, we construct a random variable X as in (5.1) that takes at most $k+1$ values. Let μ denote the distribution of X. We consider two cases that $\mathbb{E}_\mu[X] < 1$ and $\mathbb{E}_\mu[X] = 1$ separately.

- $\mathbb{E}_\mu[X] < 1$. Then X takes at most k values since otherwise $P \ll Q$. Denote the smallest value of X by x and then $x < 1$. Suppose $\mu(x) = q$ and then μ can be represented as

  $$\mu = q\delta_x + \bar{q}\mu',$$

  where μ' is supported on at most $k-1$ values of X other than x. Let $\mu_2 = \delta_x$. We need to construct another probability measure μ_1 such that

  $$\mu = \alpha\mu_1 + \bar{\alpha}\mu_2,$$

  - $\mathbb{E}_{\mu'}[X] \leq 1$. Let $\mu_1 = \mu'$ and let $\alpha = \bar{q}$.
  - $\mathbb{E}_{\mu'}[X] > 1$. Let $\mu_1 = p\delta_x + \bar{p}\mu'$ where $p = \frac{\mathbb{E}_{\mu'}[X]-1}{\mathbb{E}_{\mu'}[X]-x}$ such that $\mathbb{E}_{\mu_1}[X] = 1$. Let $\alpha = \frac{\mathbb{E}_\mu[X]-x}{1-x}$.

- $\mathbb{E}_\mu[X] = 1$.[2] Denote the smallest value of X by x and the largest value by y, respectively, and then $x \leq 1, y \geq 1$. Suppose $\mu(x) = r$ and $\mu(y) = s$ and then μ can be represented as

  $$\mu = r\delta_x + (1-r-s)\mu' + s\delta_y,$$

  where μ' is supported on at most $k-1$ values of X other than x, y. Let $\mu_2 = \beta\delta_x + \bar{\beta}\delta_y$ where $\beta = \frac{y-1}{y-x}$ such that $\mathbb{E}_{\mu_2}[X] = 1$. We need to construct another probability measure μ_1 such that

  $$\mu = \alpha\mu_1 + \bar{\alpha}\mu_2,$$

  - $\mathbb{E}_{\mu'}[X] \leq 1$. Let $\mu_1 = p\delta_y + \bar{p}\mu'$ where $p = \frac{1-\mathbb{E}_{\mu'}[X]}{y-\mathbb{E}_{\mu'}[X]}$ such that $\mathbb{E}_{\mu_1}[X] = 1$. Let $\bar{\alpha} = r/\beta$.
  - $\mathbb{E}_{\mu'}[X] > 1$. Let $\mu_1 = p\delta_x + \bar{p}\mu'$ where $p = \frac{\mathbb{E}_{\mu'}[X]-1}{\mathbb{E}_{\mu'}[X]-x}$ such that $\mathbb{E}_{\mu_1}[X] = 1$. Let $\bar{\alpha} = s/\bar{\beta}$.

Applying the construction in (5.2) with μ_1 and μ_2, we obtain two pairs of distributions (P_1, Q_1) supported on k values and (P_2, Q_2) supported on two values, respectively. Then

$$\begin{aligned}
\begin{pmatrix} D_f(P\|Q) \\ D_g(P\|Q) \end{pmatrix} &= \begin{pmatrix} \mathbb{E}_\mu[f(X)] + \tilde{f}(0)(1-\mathbb{E}_\mu[X]) \\ \mathbb{E}_\mu[g(X)] + \tilde{g}(0)(1-\mathbb{E}_\mu[X]) \end{pmatrix} \\
&= \alpha \begin{pmatrix} \mathbb{E}_{\mu_1}[f(X)] + \tilde{f}(0)(1-\mathbb{E}_{\mu_1}[X]) \\ \mathbb{E}_{\mu_1}[g(X)] + \tilde{g}(0)(1-\mathbb{E}_{\mu_1}[X]) \end{pmatrix} + \bar{\alpha} \begin{pmatrix} \mathbb{E}_{\mu_2}[f(X)] + \tilde{f}(0)(1-\mathbb{E}_{\mu_2}[X]) \\ \mathbb{E}_{\mu_2}[g(X)] + \tilde{g}(0)(1-\mathbb{E}_{\mu_2}[X]) \end{pmatrix} \\
&= \alpha \begin{pmatrix} D_f(P_1\|Q_1) \\ D_g(P_1\|Q_1) \end{pmatrix} + \bar{\alpha} \begin{pmatrix} D_f(P_2\|Q_2) \\ D_g(P_2\|Q_2) \end{pmatrix}.
\end{aligned}$$

□

[2] Many thanks to Pengkun Yang for correcting the error in the original proof.

Remark 5.1. Theorem 5.1 is a direct consequence of Krein-Milman's theorem. Consider the space of $\{P_X : X \geq 0, \mathbb{E}[X] \leq 1\}$, which has only two types of extreme points:

1. $X = x$ for $0 \leq x \leq 1$;
2. X takes two values x_1, x_2 with probability α_1, α_2, respectively, and $\mathbb{E}[X] = 1$.

For the first case, let $P = \mathrm{Bern}(x)$ and $Q = \delta_1$; for the second case, let $P = \mathrm{Bern}(\alpha_2 x_2)$ and $Q = \mathrm{Bern}(\alpha_2)$.

5.2 Examples

5.2.1 Hellinger distance versus total variation

The upper and lower bound we mentioned in the last lecture is the following [Tsy09, Sec. 2.4]:

$$\frac{1}{2}H^2(P,Q) \leq d_{\mathrm{TV}}(P,Q) \leq H(P,Q)\sqrt{1 - H^2(P,Q)/4}. \tag{5.4}$$

Their joint range $\mathcal{R}_2$ has a parametric formula

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

and is the gray region in Fig. 5.2. The joint range $\mathcal{R}$ is the convex hull of $\mathcal{R}_2$ (grey region, non-convex) and exactly described by (5.4); so (5.4) is not improvable. Indeed, with t ranges from 0 to 1,

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

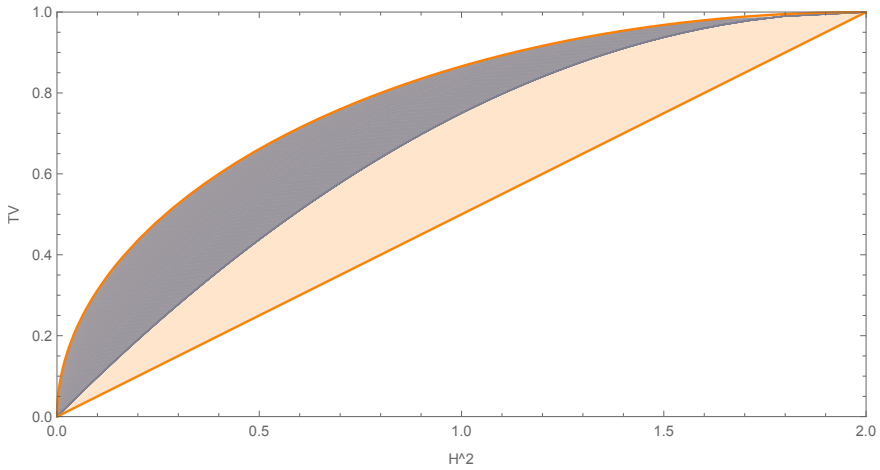

Figure 5.2: Joint range of d_{TV} and H^2.

5.2.2 KL divergence versus total variation

Pinsker's inequality states that

$$D(P\|Q) \geq 2d_{\mathrm{TV}}^2(P,Q). \tag{5.5}$$

There are various kinds of improvements of Pinsker's inequality. Now we know that the best lower bound is the lower boundary of Fig. 5.1, which is exactly the boundary of $\mathcal{R}_2$. Therefore a paremetric formula of the lower boundary is easy to write down, but there is no known close-form expression. Here is a corollary that we will use later:

$$D(P\|Q) \geq d_{\mathrm{TV}}(P,Q) \log \frac{1+TV(P,Q)}{1-TV(P,Q)}.$$

Consequences:

- The original Pinsker's inequality shows that $D \to 0 \Rightarrow d_{\mathrm{TV}} \to 0$.
- $d_{\mathrm{TV}} \to 1 \Rightarrow D \to \infty$. Thus $D = O(1) \Rightarrow d_{\mathrm{TV}}$ is bounded away from one. This is not obtainable from Pinsker's inequality.

Also from Fig. 5.1 we know that it is impossible to have an upper bound of $D(P\|Q)$ using a function of $d_{\mathrm{TV}}(P,Q)$ due to the lack of upper boundary.

For more examples see [Tsy09, Sec. 2.4].

References

[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.

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