The goal of this short note is to discuss the relation between Kullback–Leibler divergence and total variation distance, starting with the celebrated Pinsker's inequality relating the two, before switching to a simple, yet (arguably) more useful inequality, apparently not as well known. We summarize the gist of it below:

Theorem 1 (The BH Bound)**.** *For every two probability distributions $\mathbf{p}, \mathbf{q}$, we have the simple yet never vacuous bound*

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{1 - e^{-\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})}} \tag{1}$$

and this is never worse than Pinsker's inequality (except for a factor $\sqrt{2}$ for $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \ll 1$.)

While establishing this bound and discussing its various aspects, we will consider probability distributions over some set Ω, and conveniently ignore any measurability or absolute continuity issue – the reader is encouraged to think of discrete Ω for concreteness. Everything does however apply to the general setting, given the suitable insertion of the words "Radon–Nikodym derivative" and "measurable" in appropriate locations.

Total variation distance and Kullback–Leibler divergence. The TV distance and KL divergence (in nats) between two probability distributions $\mathbf{p}, \mathbf{q}$ over Ω are given respectively by

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sup_{S \subseteq \Omega} (\mathbf{p}(S) - \mathbf{q}(S)) = \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1 \in [0, 1]$$

and

$$\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) = \sum_{x \in \Omega} \mathbf{p}(x) \log \frac{\mathbf{p}(x)}{\mathbf{q}(x)} \in [0, \infty)$$

where log is the natural logarithm, with the convention that $0 \log 0 = 0$. Both TV distance and KL divergence are special cases of what is known as *f-divergences*, and they both enjoy a lot of crucial properties, such as the data processing inequality, which we will not get into here.[1]

Organisation. We start by a (very brief) review of Pinsker's inequality, and its shortcomings, in Section 1, before stating and deriving the BH bound in Section 2. The reader asking themselves why we should care at all about this improved bound can skip directly to Section 3 for some motivation and applications, and those keen on the Donsker–Varadhan formula (or looking for an open question) might enjoy Section 4. Finally, Section 5 provides some pointers, and discusses a slightly more refined (albeit much more unwieldy) bounds.

1 Pinsker's inequality

We first state our baseline, Pinsker's inequality, a fundamental relation between KL divergence and total variation distance originally due to, well, Pinsker [Pin64], although in a weaker form and with suboptimal constants: the constant was then independently improved to the optimal $1/\sqrt{2}$ by Kullback, Csizsár, and Kemperman. See [Tsy09, Section 2.8] for a discussion.

Lemma 2 (Pinsker's Inequality)**.** *For every $\mathbf{p}, \mathbf{q}$ on Ω,*

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{\frac{1}{2} \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})}\,. \tag{2}$$

[1]The KL divergence, annoyingly, is not symmetric in its arguments and thus not a real metric, but it makes up for it sometimes.

There are many proofs of Pinsker's inequality: e.g., [Tsy09, Lemma 2.5], or a very clever argument due to Pollard, or even in Section 4 of this very note, using the Donsker–Varadhan formula (also known by me as "Thomas' Favourite Lemma"). One can for instance consult [Raz] for a list; we will here follow an argument from Yihong Wu's lecture notes [Wu20, Theorem 4.5], which has the advantage of being nearly magical.

Proof. Consider first the binary case, i.e., where $\mathbf{p}$,$\mathbf{q}$ are Bernoulli distributions $\mathrm{Bern}(p)$ and $\mathrm{Bern}(q)$, respectively. Then $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = |p - q|$ and so we are left to prove

$$2(p-q)^2 \leq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \tag{3}$$

Note that the cases where either p or q is in $\{0, 1\}$ are easily checked *(verify it!)*, so we can assume $p, q \in (0, 1)$. To prove (3) in this case, we introduce the function $f\colon (0,1) \to \mathbb{R}$ defined by $f(x) = p \log x + (1-p) \log(1-x)$, and observe that the RHS of (3) is exactly $f(p) - f(q)$. We then can write

$$f(p) - f(q) = \int_q^p f'(x)dx = \int_q^p \frac{p-x}{x(1-x)}dx \geq 4 \int_q^p (p-x)dx = 4 \cdot \frac{1}{2}(p-q)^2$$

establishing (3) (note that we used the fact that $x(1-x) \leq 1/4$ for $x \in (0, 1)$).

Turning to the general case, let $\mathbf{p}, \mathbf{q}$ be distributions on an arbitrary domain Ω, and fix any measurable subset $S \subseteq \Omega$. For X, Y distributed according to $\mathbf{p}$ and $\mathbf{q}$, the random variables $\mathbb{1}_S(X)$ and $\mathbb{1}_S(Y)$ are have distributions $\mathbf{p}' \coloneqq \mathrm{Bern}(\mathbf{p}(S))$ and $\mathbf{q}' \coloneqq \mathrm{Bern}(\mathbf{q}(S))$ respectively, and therefore

$$2(\mathbf{p}(S) - \mathbf{q}(S))^2 = 2\mathrm{d}_{\mathrm{TV}}(\mathbf{p}', \mathbf{q}')^2 \leq \mathrm{KL}(\mathbf{p}' \,\|\, \mathbf{q}') \leq \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})$$

where the first inequality is (3), and the second is the data processing inequality. Since this inequality holds for every S, taking a supremum over S leads to

$$2\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})^2 = 2 \sup_{S \subseteq \Omega} (\mathbf{p}(S) - \mathbf{q}(S))^2 \leq \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \,,$$

establishing Pinsker's inequality. $\square$

Before we try to improve upon Pinsker's inequality, let us note that one particular avenue is doomed: specifically, the constant $1/\sqrt{2}$ in (2) cannot be replaced by any $c < 1/\sqrt{2}$. To see why, fix any $\varepsilon \in (0, 1/4)$, and observe that for $\mathbf{p} = \mathrm{Bern}(1/2)$ and $\mathbf{q} = \mathrm{Bern}(1/2 + \varepsilon)$ we have $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \varepsilon$ and $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) = \frac{1}{2} \log \frac{1}{1-4\varepsilon^2}$, so that

$$\frac{\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})}{\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})^2} = -\frac{\log(1-4\varepsilon^2)}{2\varepsilon^2} \xrightarrow[\varepsilon \to 0^+]{} 2 \,. \tag{4}$$

Still, in spite of its multiple applications in Statistics and information theory and its "optimality" shown above, Pinsker's inequality suffers a major drawback: by definition, the TV distance is always at most 1, yet the RHS of (2) grows unbounded with the KL divergence. In other terms, the bound is totally and utterly useless for any $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) > 2$, as depicted in Figure 1.

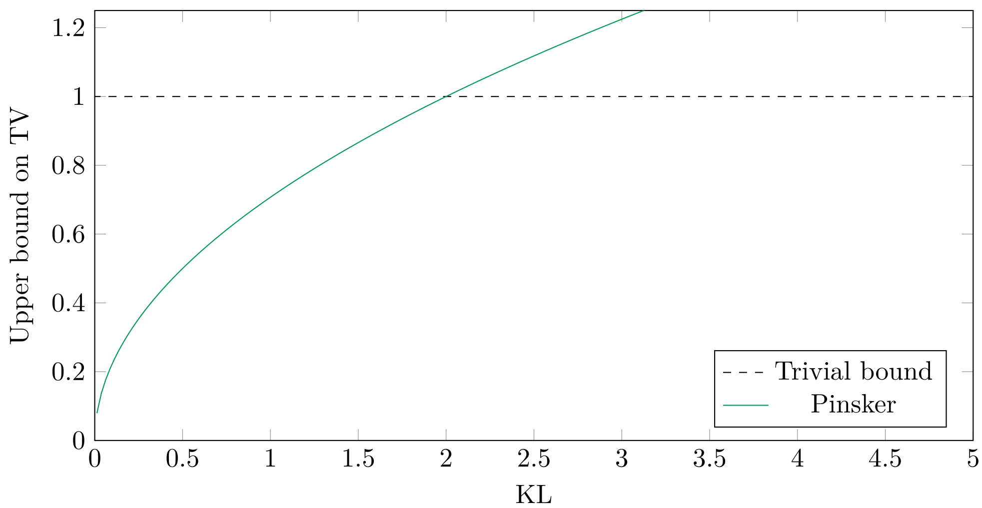

Figure 1: Pinsker's inequality becomes vacuous for $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) > 2$. That's a downer.

To see why one would care about this issue (without jumping yet to Section 3), consider the following very simple and intuitive fact: *"if* $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) < \infty$*, then* $\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) < 1$.*"* While absolutely true, this claim cannot be proven from Pinsker's inequality. Even worse, using Pinsker's one cannot even establish that if $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) < 2.01$ then the two distributions $\mathbf{p}$ and $\mathbf{q}$ have TV distance bounded away from 1!

2 The Bretagnolle–Huber bound

In view of the above, can we hope for some better inequality which does not leap into vacuousness when the KL divergence gets large? The answer is, thankfully, yes.

Lemma 3 (The BH Bound)**.** *For every* $\mathbf{p}, \mathbf{q}$ *on* Ω*,*

$$\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{1 - e^{-\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})}}\,. \tag{5}$$

Proof. We follow the original argument of [BH78, Lemma 2.1]: fixing $\mathbf{p}, \mathbf{q}$, we define, for X distributed according to $\mathbf{p}$, the random variables $U := \frac{\mathbf{q}(X)}{\mathbf{p}(X)}$, $V := (U-1)_+$, and $W := 1 + V - U = (1-U)_+$. One can check that

$$\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\mathbb{E}_{\mathbf{p}}[|U-1|] = \mathbb{E}_{\mathbf{p}}[V] = \mathbb{E}_{\mathbf{p}}[W]$$

and that by construction $(1+V)(1-W) = U$, so that $\log U = \log(1+V) + \log(1-W)$. Moreover, since $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) = -\sum_{x\in\Omega} \mathbf{p}(x) \log \frac{\mathbf{q}(x)}{\mathbf{p}(x)} = -\mathbb{E}_{\mathbf{p}}[\log U]$, we get by Jensen's inequality that

$$-\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) = \mathbb{E}_{\mathbf{p}}[\log(1+V)] + \mathbb{E}_{\mathbf{p}}[\log(1-W)] \leq \log(1+\mathbb{E}_{\mathbf{p}}[V]) + \log(1-\mathbb{E}_{\mathbf{p}}[W]) = \log(1-\mathrm{d_{TV}}(\mathbf{p}, \mathbf{q})^2)$$

which, exponentiating both sides, rearranging and taking the square root, proves the lemma. $\square$

Now, instead of the above bound, one may encounter the following weaker one, for instance in Tsybakov's monograph [Tsy09].[2] It is unclear to me what advantage this looser inequality holds over (3), but as we shall see in Figure 2 it at least behaves in a satisfying way for large values of $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})$, and is never bigger than the trivial bound of 1.

[2]which is worth reading by itself, as it countains many gems, insights, and useful discussions.

Corollary 4 (Tsybakov's version)**.** *For every* $\mathbf{p}, \mathbf{q}$ *on* Ω*,*

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq 1 - \frac{1}{2} e^{-\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})} . \tag{6}$$

Proof. This readily frollows from Lemma 3, upon noting that $\sqrt{1 - e^{-x}} \leq 1 - \frac{1}{2}e^{-x}$ for all $x \in [0, \infty)$ (just square both sides and expand the RHS). ☐

Because it is somewhat fun to do, we also reproduce Tsybakov's proof of Corollary 4, and show how it can be used to derive Lemma 3 (so I am at a loss as to why Tsybakov would only state the weaker version in his monograph).[3]

Proof of Lemma 3 from the argument of [Tsy09], Lemma 2.6. Fix $\mathbf{p}, \mathbf{q}$. First, we observe that one can write

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = 1 - \sum_{x \in \Omega} \min(\mathbf{p}(x), \mathbf{q}(x)) = \sum_{x \in \Omega} \max(\mathbf{p}(x), \mathbf{q}(x)) - 1 \tag{7}$$

(this is a useful trick, check it!), and therefore by Cauchy–Schwarz

$$\begin{aligned} 1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})^2 &= (1 + \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}))(1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})) = \left(\sum_{x \in \Omega} \max(\mathbf{p}(x), \mathbf{q}(x))\right)\left(\sum_{x \in \Omega} \min(\mathbf{p}(x), \mathbf{q}(x))\right) \\ &\geq \left(\sum_{x \in \Omega} \sqrt{\max(\mathbf{p}(x), \mathbf{q}(x)) \min(\mathbf{p}(x), \mathbf{q}(x))}\right)^2 = \left(\sum_{x \in \Omega} \sqrt{\mathbf{p}(x)\mathbf{q}(x)}\right)^2 \end{aligned} \tag{8}$$

which will come handy very soon. Indeed, what Tsybakov does show is the following:

$$\begin{aligned} \left(\sum_{x \in \Omega} \sqrt{\mathbf{p}(x)\mathbf{q}(x)}\right)^2 &= e^{2 \log \sum_x \sqrt{\mathbf{p}(x)\mathbf{q}(x)}} = e^{2 \log \sum_x \mathbf{p}(x) \sqrt{\frac{\mathbf{q}(x)}{\mathbf{p}(x)}}} \\ &= e^{2 \log \mathbb{E}_{\mathbf{p}}\left[\sqrt{\frac{\mathbf{q}(X)}{\mathbf{p}(X)}}\right]} \geq e^{2\mathbb{E}_{\mathbf{p}}\left[\log \sqrt{\frac{\mathbf{q}(X)}{\mathbf{p}(X)}}\right]} && \text{(Jensen's inequality)} \\ &= e^{\mathbb{E}_{\mathbf{p}}\left[\log \frac{\mathbf{q}(X)}{\mathbf{p}(X)}\right]} = e^{-\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})} \end{aligned} \tag{9}$$

(to be precise, sums and expectations are restricted to the support of $\mathbf{p}$, to avoid dividing by zero). Combining (8) and (9) yields Lemma 3. ☐

To see how these bounds (2), (5), and (6) compare, let us look at a plot.

[3] We observe that the same "extension" of Tsybakov's argument can be found in the proof of [GHRZ19, Lemma 6].

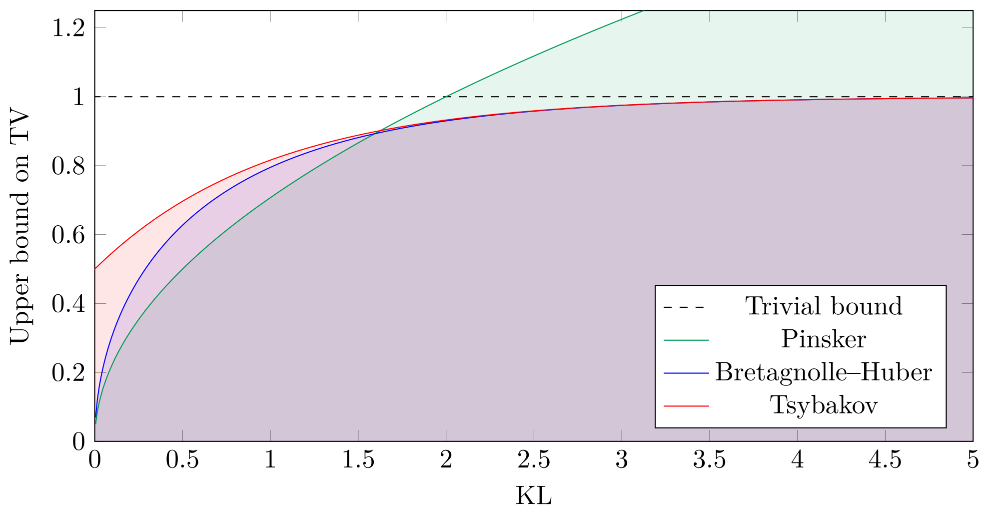

Figure 2: The four upper bounds we have: shaded regions correspond to values of TV still allowed by the corresponding bounds (so smaller shaded areas are better). As we can see, Pinsker's bound (2) is useful for small values of $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})$ only. Tsybakov's bound (6) is much better for large $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})$, and in particular have the right behaviour as $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \to \infty$, never becoming worse than the trivial bound. However, it is now useless for small $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})$, and does not even go to 0 as $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \to 0^+$. The clear winner is the Bretagnolle–Huber bound (5), which not only is never worse than Tsybakov's (obviously), but also has the right behaviour for small values of $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})$, being essentially equivalent (up to a constant factor) to Pinsker's in that regime.

Figure 2 clearly hints that the BH bound obtained in Lemma 3 is never much worse than the one from Pinsker's inequality, but let us make this formal. First, a Taylor approximation shows that $\sqrt{1-e^{-x}} = \sqrt{x} + o(\sqrt{x})$ as $x \to 0^+$, so for small TV our new bound is worse than Pinsker's by only a factor $\sqrt{2}$. It is actually easy to see that this is always the case, as the inequality $\sqrt{1-e^{-x}} \leq \sqrt{2} \cdot \sqrt{\frac{x}{2}}$ (for $x \geq 0$) is equivalent to $1 - x \leq e^{-x}$, which holds by convexity. We can summarize this as follows:

> The BH bound (5) is never vacuous, has the right behaviour when $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \to \infty$ *and* $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \to 0^+$, and is at worst a $\sqrt{2}$ factor off from Pinsker's bound (2).

To provide a complementary view of the depiction of the bounds from Figure 2 (which showed the *upper* bounds on TV, as a function of KL, implied by the Pinsker, BH, and Tsybakov inequalities), we give in Figure 3 the corresponding *lower* bounds on KL as a function of TV.

So close, yet so far? Interestingly, one can derive an inequality looking similar to the BH bound from Pinsker's inequality, with a major caveat. Note that (2) can be equivalently rephrased as follows:

$$1 - e^{-2\mathrm{d_{TV}}(\mathbf{p},\mathbf{q})^2} \leq 1 - e^{-\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})} . \tag{10}$$

Using the (tight) inequality $x \leq (1-e^{-2})^{-1}(1-e^{-2x})$, which holds for all $x \in [0,1]$, we then get

$$\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) \leq \frac{1}{\sqrt{1-e^{-2}}} \cdot \sqrt{1 - e^{-\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})}} , \tag{11}$$

which, except for this leading factor $\frac{1}{\sqrt{1-e^{-2}}} \approx 1.075$, looks very much like (5). Unfortunately, this leading factor is exactly what makes (11) useless, as the bound is still vacuous whenever $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) > 2$, and further is strictly weaker than Pinsker's for $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) < 2$. See Figure 4 for an illustration.

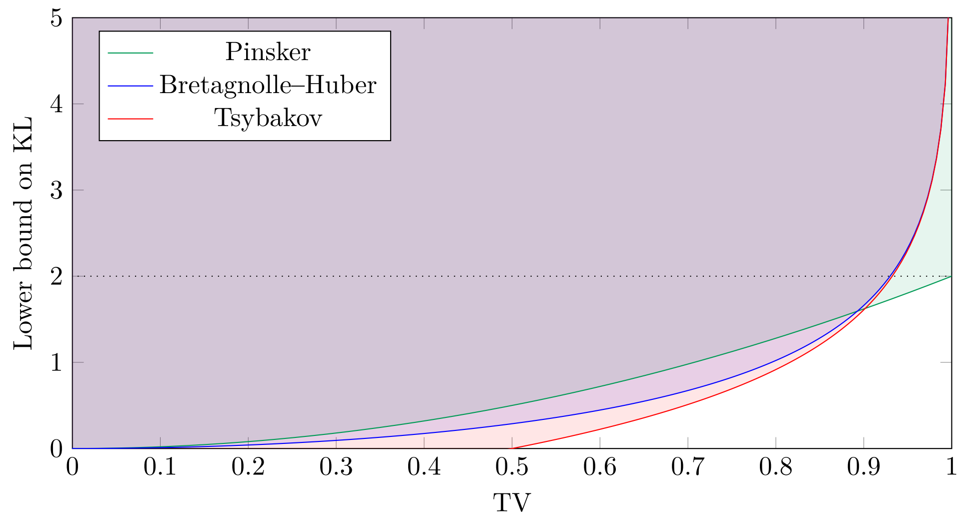

Figure 3: The lower bounds which (2), (5), and (6) give on the KL divergence. The shaded areas are the values of KL (as a function of TV) still allowed by the corresponding inqualities, so smaller shaded area is better: as one can see, Pinsker's inequality is unable to rule out any value of KL greater than 2, while the bound given by Tsybakov only kicks in for TV $\geq 1/2$.

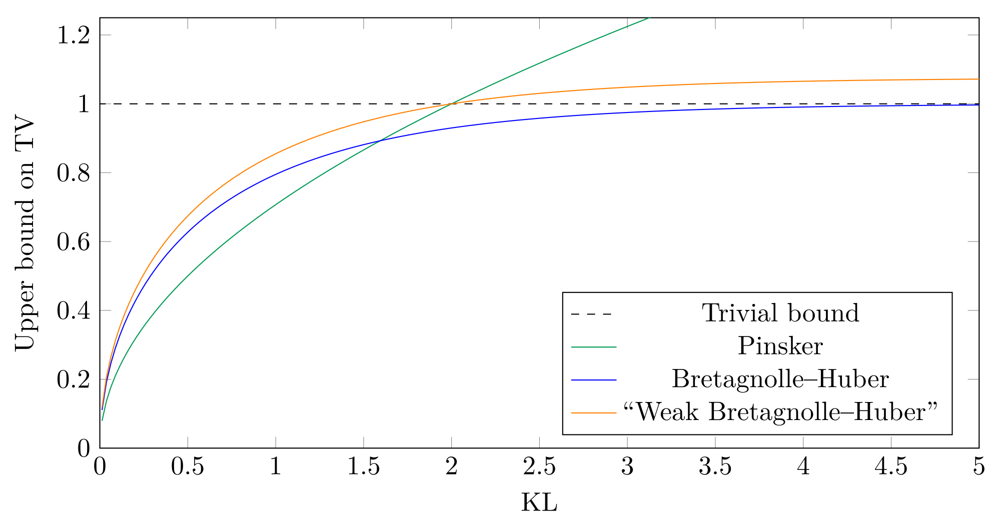

Figure 4: A "weak Bretagnolle–Huber bound" (11) can be derived from Pinsker's inequality, but it is not a good idea.

3 Why do we care?

As we saw, the BH bound (5) and the weaker bound (6) both improve on Pinsker's inequality (2) in the regime of large KL. Since for vanishingly small KL Pinsker's inequality is tight in general, and provides a good bound for small constant values as well, it is natural to wonder why we may care about the regime KL $\gg 1$. One example lies in proving sample complexity lower bounds. As an example and motivation, consider the following (true) fact:

Fact 5. *The number of independent tosses required to distinguish with probability at least* $1-\delta$ *between a* *fair coin (i.e.,* $\mathrm{Bern}(1/2)$*) and an ε-biased coin (i.e.,* $\mathrm{Bern}(1/2+\varepsilon)$*) is* $\Omega\big(\log(1/\delta)/\varepsilon^2\big)$.

To prove this for constant $\delta > 0$, say $\delta = 1/10$, the standard way to proceed is to observe that, by a relatively standard argument, we need the number n of samples (tosses) to satisfy

$$\mathrm{d}_{\mathrm{TV}}\Big(\mathrm{Bern}(1/2)^{\otimes n}, \mathrm{Bern}(1/2+\varepsilon)^{\otimes n}\Big) \geq 1-2\delta \tag{12}$$

and we can then use Pinsker's inequality and additivity of KL divergence for product distributions to get

$$\begin{aligned}
(1-2\delta)^2 &\leq \mathrm{d}_{\mathrm{TV}}\Big(\mathrm{Bern}(1/2)^{\otimes n}, \mathrm{Bern}(1/2+\varepsilon)^{\otimes n}\Big)^2 \\
&\leq \frac{1}{2}\mathrm{KL}\Big(\mathrm{Bern}(1/2)^{\otimes n} \parallel \mathrm{Bern}(1/2+\varepsilon)^{\otimes n}\Big) && \text{(Pinsker)} \\
&= n \cdot \frac{1}{2}\mathrm{KL}(\mathrm{Bern}(1/2) \parallel \mathrm{Bern}(1/2+\varepsilon)) \\
&= n \cdot \frac{1}{2}\log\frac{1}{1-4\varepsilon^2} && \text{(Direct computation of KL)}
\end{aligned}$$

which is at most $4n\varepsilon^2$ for ε small enough, e.g., $0 < \varepsilon < 1/3$. This shows the $\Omega\big(1/\varepsilon^2\big)$ for constant $\delta \in (0, 1/2)$. It is not hard to see, unfortunately, that this approach will never yield any bound better than $\Omega\big(1/\varepsilon^2\big)$, even as $\delta \to 0^+$; exactly because Pinsker's inequality does not allow us to discrimate between "moderately large KL" and "KL going to ∞." But if we were to use the BH bound instead, then the exact same argument shows that we need

$$\begin{aligned}
(1-2\delta)^2 &\leq \mathrm{d}_{\mathrm{TV}}\Big(\mathrm{Bern}(1/2)^{\otimes n}, \mathrm{Bern}(1/2+\varepsilon)^{\otimes n}\Big)^2 \\
&\leq 1 - e^{-\mathrm{KL}\left(\mathrm{Bern}(1/2)^{\otimes n} \parallel \mathrm{Bern}(1/2+\varepsilon)^{\otimes n}\right)} && \text{(BH)} \\
&= 1 - e^{-n\cdot\frac{1}{2}\mathrm{KL}(\mathrm{Bern}(1/2) \parallel \mathrm{Bern}(1/2+\varepsilon))} \\
&= 1 - e^{-n\cdot\frac{1}{2}\log\frac{1}{1-4\varepsilon^2}}
\end{aligned}$$

or, reorganizing, $n \geq \frac{2}{\log\frac{1}{1-4\varepsilon^2}}\log\frac{1}{1-(1-2\delta)^2} \geq \frac{1}{2\varepsilon^2}\log\frac{1}{2\delta}$ (the last inequality again for $\varepsilon \in (0, 1/3)$), which proves Fact 5.

Remark 6. One can also use (6) to prove Fact 5 in a similar fashion, which turns out to be even (marginally) simpler: verify it!

4 The TFL

Let us switch gears a little, and consider the relation between these inequalities and the fundamental lemma below, sometimes known as the *Gibbs variational principle*, or the *Donsker–Varadhan formula* [DV75], and which I have been told is a special case of Fenchel duality. We will refer to it, succinctly, as *Thomas' Favourite Lemma*, thus named after Thomas Steinke and his fondness for this result.

Lemma 7 (Thomas' Favourite Lemma (TFL))**.** *For every* $\mathbf{q} \ll \mathbf{p}$*,*

$$\mathrm{KL}(\mathbf{p} \parallel \mathbf{q}) = \sup_f \Big(\mathbb{E}_{\mathbf{p}}[f(X)] - \log \mathbb{E}_{\mathbf{q}}[e^{f(Y)}]\Big)$$

where the supremum is over all (measurable) $f\colon \Omega \to \mathbb{R}$*.*

From TFL to Pinsker. Consider any bounded function $f\colon \Omega \to \mathbb{R}$. From Lemma 7 followed by an application of Hoeffding's Lemma, we can write

$$\begin{aligned}\mathbb{E}_{\mathbf{p}}[f(X)] &\leq \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) + \log \mathbb{E}_{\mathbf{q}}[e^{f(Y)}] && \text{(TFL)}\\ &\leq \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) + \mathbb{E}_{\mathbf{q}}[f(Y)] + \frac{1}{2}\|f\|_\infty^2 && \text{(Hoeffding's Lemma)}\end{aligned}$$

so, reorganizing and taking the supremum over all f such that $\|f\|_\infty = \lambda$ (for some $\lambda > 0$ to be carefully chosen), we get

$$\sup_{f\colon \|f\|_\infty = \lambda} (\mathbb{E}_{\mathbf{p}}[f(X)] - \mathbb{E}_{\mathbf{q}}[f(Y)]) \leq \mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) + \frac{1}{2}\lambda^2 \,. \tag{13}$$

Noting then that the LHS is exactly equal to $2\lambda \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$,[4] from (13) we are left with the following, true for all $\lambda > 0$:

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \frac{1}{2\lambda}\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) + \frac{\lambda}{4} \,. \tag{14}$$

Optimizing for $\lambda > 0$, we choose $\lambda := \sqrt{2\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})}$ and obtain

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \sqrt{\frac{1}{2}\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q})} \,, \tag{15}$$

retrieving Pinsker's inequality (2).

It is, however, unclear if one can obtain the Bretagnolle–Huber inequality in a similar fashion. At the very least, I do not know how.

Question 8. Can one derive (5) from the TFL?

5 Discussion and pointers

This note is only a succinct, non-exhaustive discussion of possible improvements to Pinsker's inequality, and barely scratches the surface of the many results on this and related questions. We conclude with a few pointers for the interested and fearless reader: Reid and Williamson [RW09] provide a generalization of Pinsker-type inequalities for other f-divergences, as well as an (optimal) integral form of the inequality. Speaking of inequalities between f-divergences, Sason and Verdú develop in [SV16] techniques to obtain many bounds, among which the Bretagnolle–Huber one. Finally, the recent book of Lattimore and Szepesvári [LS20] covers the BH bound in its chapter on relative entropy (Theorem 14.2), where it also provides some context and discussion.

We could not conclude without mentioning that the excellent lecture notes of Yihong Wu [Wu20] devote an entire chapter (Section 5) to inequalities between f-divergences, including a wonderful theorem due to Harremoës and Vajda (Theorem 5.1), which essentially states that to prove *any* such inequality it suffices to prove it for Bernoulli random variables. Those lecture notes also provide, in Section 5.2.2, a handy (albeit short) discussion of Pinsker's inequality, and states the following improvement due to Vajda [Vaj70]:

$$\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \geq \log \frac{1 + \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})}{1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})} - \frac{2\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})}{1 + \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})} \tag{16}$$

This is even tighter than the BH bound (Theorem 1) we spent so much time covering here, and which only states that $\mathrm{KL}(\mathbf{p} \,\|\, \mathbf{q}) \geq \log \frac{1}{1 - \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})^2}$;[5] however, it is (at least in my eyes) *much* more cumbersome to use.

[4] By definition of TV distance as integral probability metric [Mÿ7], or, without using those fancy terms, checking that

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sup_{S \subseteq \Omega} (\mathbb{E}_{\mathbf{p}}[\mathbb{1}_S(X)] - \mathbb{E}_{\mathbf{q}}[\mathbb{1}_S(Y)]) = \frac{1}{2} \sup_{f\colon \|f\|_\infty \leq 1} (\mathbb{E}_{\mathbf{p}}[f(X)] - \mathbb{E}_{\mathbf{q}}[f(Y)]) \,.$$

[5] In particular, (16) does not lose that asymptotic factor $\sqrt{2}$ over Pinsker's for small KL, unlike the BH bound.

References

[BH78] J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. In *Séminaire de Probabilités, XII (Univ. Strasbourg, Strasbourg, 1976/1977)*, volume 649 of *Lecture Notes in Math.*, pages 342–363. Springer, Berlin, 1978.

[DV75] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. I. II. *Comm. Pure Appl. Math.*, 28:1–47; ibid. 28 (1975), 279–301, 1975.

[GHRZ19] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 503–513. Curran Associates, Inc., 2019.

[LS20] T. Lattimore and Cs. Szepesvári. *Bandit algorithms.* Cambridge U. Press, 2020. Available at https://tor-lattimore.com/downloads/book/book.pdf.

[M̈97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.*, 29(2):429–443, 1997.

[Pin64] M. S. Pinsker. *Information and information stability of random variables and processes.* Translated and edited by Amiel Feinstein. Holden-Day, Inc., San Francisco, Calif.-London-Amsterdam, 1964.

[Raz] Ilya Razenshteyn. Two reference requests: Pinsker's inequality and Pontryagin duality. MathOverflow. URL:https://mathoverflow.net/q/42667 (version: 2018-11-01).

[RW09] Mark D. Reid and Robert C. Williamson. Generalised pinsker inequalities. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[SV16] Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Trans. Inform. Theory*, 62(11):5973–6006, 2016. Available at https://arxiv.org/abs/1508.00335.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[Vaj70] Igor Vajda. Note on discrimination information and variation. *IEEE Trans. Inform. Theory*, IT-16:771–773, 1970.

[Wu20] Yihong Wu. Lecture notes on: Information-theoretic methods for high-dimensional statistics. online, 2020. URL:http://www.stat.yale.edu/~yw562/teaching/it-stats.pdf (accessed 2020-12-18).