# $f$-Divergence Inequalities via Functional Domination


Igal Sason
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion-Israel Institute of Technology
Haifa 32000, Israel
E-mail: sason@ee.technion.ac.il

Sergio Verdú
Department of Electrical Engineering
Princeton University
New Jersey 08544, USA
E-mail: verdu@princeton.edu



***Abstract*—This paper considers derivation of $f$-divergence inequalities via the approach of functional domination. Bounds on an $f$-divergence based on one or several other $f$-divergences are introduced, dealing with pairs of probability measures defined on arbitrary alphabets. In addition, a variety of bounds are shown to hold under boundedness assumptions on the relative information.**[1]



**Index Terms** – $f$-divergence, relative entropy, relative information, reverse Pinsker inequalities, reverse Samson's inequality, total variation distance, $\chi^2$ divergence.


## I. Basic Definitions

We assume throughout that the probability measures $P$ and $Q$ are defined on a common measurable space $(\mathcal{A}, \mathscr{F})$, and $P \ll Q$ denotes that $P$ is *absolutely continuous* with respect to $Q$.

*Definition 1:* If $P \ll Q$, the *relative information* provided by $a \in \mathcal{A}$ according to $(P, Q)$ is given by[2]

$$\imath_{P\|Q}(a) \triangleq \log \frac{\mathrm{d}P}{\mathrm{d}Q}\,(a). \tag{1}$$

Introduced by Ali-Silvey [1] and Csiszár ([4]), a useful generalization of the relative entropy, which retains some of its major properties (and, in particular, the data processing inequality), is the class of $f$-divergences. A general definition of $f$-divergence is given in [14, p. 4398], specialized next to the case where $P \ll Q$.

*Definition 2:* Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function, and suppose that $P \ll Q$. The *$f$-divergence* from $P$ to $Q$ is given by

$$D_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}Q = \mathbb{E}\big[f(Z)\big] \tag{2}$$

with

$$Z = \exp\big(\imath_{P\|Q}(Y)\big), \quad Y \sim Q. \tag{3}$$

In (2), we take the continuous extension[3]

$$f(0) = \lim_{t \downarrow 0} f(t) \in (-\infty, +\infty]. \tag{4}$$

If $p$ and $q$ denote, respectively, the densities of $P$ and $Q$ with respect to a $\sigma$-finite measure $\mu$ (i.e., $p = \frac{\mathrm{d}P}{\mathrm{d}\mu}$, $q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$), then we can write (2) as

$$D_f(P\|Q) = \int_{\{q>0\}} q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu. \tag{5}$$

*Remark 1:* Different functions may lead to the same $f$-divergence for all $(P, Q)$: if for an arbitrary $b \in \mathbb{R}$, we have

$$f_b(t) = f_0(t) + b\,(t-1), \quad t \geq 0 \tag{6}$$

then

$$D_{f_0}(P\|Q) = D_{f_b}(P\|Q). \tag{7}$$

Relative entropy is $D_r(P\|Q)$ where $r$ is given by

$$r(t) = t \log t + (1-t) \log e, \tag{8}$$

and the total variation distance $|P - Q|$ and $\chi^2$ divergence $\chi^2(P\|Q)$ are $f$-divergences with $f(t) = (t-1)^2$ and $f(t) = |t-1|$, respectively.

The following key property of $f$-divergences follows from Jensen's inequality.

*Proposition 1:* If $f\colon (0,\infty) \to \mathbb{R}$ is convex and $f(1) = 0$, $P \ll Q$, then

$$D_f(P\|Q) \geq 0. \tag{9}$$

If, furthermore, $f$ is strictly convex at $t = 1$, then equality in (9) holds if and only if $P = Q$.

The reader is referred to [19] for a survey on general properties of $f$-divergences, and also to the textbook by Liese and Vajda [13].

The full paper version of this work, which includes several other approaches for the derivation of $f$-divergence inequalities, is available in [17].

## II. Functional Domination

Let $f$ and $g$ be convex functions on $(0,\infty)$ with $f(1) = g(1) = 0$, and let $P$ and $Q$ be probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$. If, for $\alpha > 0$, $f(t) \leq \alpha g(t)$ for all $t \in (0,\infty)$ then, it follows from Definition 2 that

$$D_f(P\|Q) \leq \alpha\, D_g(P\|Q). \tag{10}$$

This simple observation leads to a proof of several inequalities with the aid of Remark 1.


[1] This work has been supported by the Israeli Science Foundation (ISF) under Grant 12/12, by the US National Science Foundation under Grant CCF-1016625, and in part by the Center for Science of Information, an NSF Science and Technology Center under Grant CCF-0939370.


[2] $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon-Nikodym derivative (or density) of $P$ with respect to $Q$. Logarithms have an arbitrary common base, and the exponent indicates the inverse function of the logarithm with that base.

[3] The convexity of $f\colon (0,\infty) \to \mathbb{R}$ implies its continuity on $(0,\infty)$.

### A. Basic Tool

We start this section by proving a general result, which will be helpful in proving various tight bounds among $f$-divergences.

*Theorem 1:* Let $P \ll Q$, and assume

- $f$ is convex on $(0, \infty)$ with $f(1) = 0$;
- $g$ is convex on $(0, \infty)$ with $g(1) = 0$;
- $g(t) > 0$ for all $t \in (0, 1) \cup (1, \infty)$.

Denote the function $\kappa \colon (0, 1) \cup (1, \infty) \to \mathbb{R}$

$$\kappa(t) = \frac{f(t)}{g(t)}, \quad t \in (0, 1) \cup (1, \infty) \tag{11}$$

and

$$\bar{\kappa} = \sup_{t \in (0,1) \cup (1,\infty)} \kappa(t). \tag{12}$$

Then,

a)
$$D_f(P\|Q) \leq \bar{\kappa}\, D_g(P\|Q). \tag{13}$$

b) If, in addition, $f'(1) = g'(1) = 0$, then
$$\sup_{P \neq Q} \frac{D_f(P\|Q)}{D_g(P\|Q)} = \bar{\kappa}. \tag{14}$$

*Proof:* See [17, Theorem 1]. ■

*Remark 2:* Beyond the restrictions in Theorem 1a), the only operative restriction imposed by Theorem 1b) is the differentiability of the functions $f$ and $g$ at $t = 1$. Indeed, we can invoke Remark 1 and add $f'(1)\,(1 - t)$ to $f(t)$, without changing $D_f$ (and likewise with $g$) and thereby satisfying the condition in Theorem 1b); the stationary point at 1 must be a minimum of both $f$ and $g$ because of the assumed convexity, which implies their non-negativity on $(0, \infty)$.

*Remark 3:* It is useful to generalize Theorem 1b) by dropping the assumption on the existence of the derivatives at 1. As it is explained in [17], it is enough to require that the left derivatives of $f$ and $g$ at 1 be equal to 0. Analogously, if $\bar{\kappa} = \sup_{0<t<1} \kappa(t)$, it is enough to require that the right derivatives of $f$ and $g$ at 1 be equal to 0.

### B. Relationships Among $D(P\|Q)$, $\chi^2(P\|Q)$ and $|P - Q|$

*Theorem 2:*

a) If $P \ll Q$ and $c_1, c_2 \geq 0$, then
$$D(P\|Q) \leq \left(c_1\, |P - Q| + c_2\, \chi^2(P\|Q)\right) \log e \tag{15}$$
holds if $(c_1, c_2) = (0, 1)$ and $(c_1, c_2) = \left(\frac{1}{4}, \frac{1}{2}\right)$. Furthermore, if $c_1 = 0$ then $c_2 = 1$ is optimal, and if $c_2 = \frac{1}{2}$ then $c_1 = \frac{1}{4}$ is optimal.

b) If $P \ll\gg Q$ and $P \neq Q$, then
$$\frac{D(P\|Q) + D(Q\|P)}{\chi^2(P\|Q) + \chi^2(Q\|P)} \leq \tfrac{1}{2}\, \log e \tag{16}$$
and the constant in the right side of (16) is the best possible.

*Proof:* See [17, Theorem 2]. ■

*Remark 4:* Inequality (15) strengthens the bound in [9, (2.8)],

$$D(P\|Q) \leq \tfrac{1}{2}\left(|P - Q| + \chi^2(P\|Q)\right) \log e. \tag{17}$$

Note that the short outline of the suggested proof in [9, p. 710] leads not (17) but to the weaker upper bound $|P - Q| + \frac{1}{2}\,\chi^2(P\|Q)$ nats.

### C. An Alternative Proof of Samson's Inequality

For the purpose of this sub-section, we introduce *Marton's divergence* [15]:

$$d_2^2(P, Q) = \min \mathbb{E}\left[\mathbb{P}^2[X \neq Y \mid Y]\right] \tag{18}$$

where the minimum is over all probability measures $P_{XY}$ with respective marginals $P_X = P$ and $P_Y = Q$. From [15, pp. 558–559]

$$d_2^2(P, Q) = D_s(P\|Q) \tag{19}$$

with

$$s(t) = (t - 1)^2\ 1\{t < 1\}. \tag{20}$$

Note that Marton's divergence satisfies the triangle inequality [15, Lemma 3.1], and $d_2(P, Q) = 0$ implies $P = Q$; however, due to its asymmetry, it is not a distance measure.

An analog of Pinsker's inequality, which comes in handy for the proof of Marton's conditional transportation inequality [3, Lemma 8.4], is the following bound due to Samson [16, Lemma 2]:

*Theorem 3:* If $P \ll Q$, then

$$d_2^2(P, Q) + d_2^2(Q, P) \leq \tfrac{2}{\log e}\ D(P\|Q). \tag{21}$$

In [17, Section 3.D], we provide an alternative proof of Theorem 3, in view of Theorem 1b), with the following advantages:

a) This proof yields the optimality of the constant in (21), i.e., we prove that
$$\sup_{P \neq Q} \frac{d_2^2(P, Q) + d_2^2(Q, P)}{D(P\|Q)} = \tfrac{2}{\log e} \tag{22}$$
where the supremum is over all probability measures $P, Q$ such that $P \neq Q$ and $P \ll\gg Q$.

b) A simple adaptation of this proof results in a reverse inequality to (21), which holds under the boundedness assumption of the relative information (see Section III-D).

### D. Ratio of $f$-Divergence to Total Variation Distance

Let $f \colon (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and let $f^\star \colon (0, \infty) \to \mathbb{R}$ be given by

$$f^\star(t) = t\, f\left(\tfrac{1}{t}\right) \tag{23}$$

for all $t > 0$. Note that $f^\star$ is also convex, $f^\star(1) = 0$, and $D_f(P\|Q) = D_{f^\star}(Q\|P)$ if $P \ll\gg Q$. By definition, we take

$$f^\star(0) = \lim_{t \downarrow 0} f^\star(t) = \lim_{u \to \infty} \frac{f(u)}{u}. \tag{24}$$

Vajda [18, Theorem 2] showed that the range of an $f$-divergence is given by

$$0 \le D_f(P\|Q) \le f(0) + f^\star(0) \tag{25}$$

where every value in this range is attainable by a suitable pair of probability measures $P \ll Q$. Recalling Remark 1, note that $f_b(0) + f_b^\star(0) = f(0) + f^\star(0)$ with $f_b(\cdot)$ defined in (6). Basu *et al.* [2, Lemma 11.1] strengthened (25), showing that

$$D_f(P\|Q) \le \tfrac{1}{2}\left(f(0) + f^\star(0)\right) |P - Q|. \tag{26}$$

If $f(0)$ and $f^\star(0)$ are finite, (26) yields a counterpart to a result by Csiszár (see [6, Theorem 3.1]) which implies that if $f\colon (0,\infty) \to \mathbb{R}$ is a strictly convex function, then there exists a real-valued function $\psi_f$ such that $\lim_{x\downarrow 0} \psi_f(x) = 0$, and

$$|P - Q| \le \psi_f\big(D_f(P\|Q)\big). \tag{27}$$

Next, we demonstrate that the constant in (26) cannot be improved.

*Theorem 4:* If $f\colon (0,\infty) \to \mathbb{R}$ is convex with $f(1) = 0$, then

$$\sup_{P \ne Q} \frac{D_f(P\|Q)}{|P - Q|} = \tfrac{1}{2}\left(f(0) + f^\star(0)\right) \tag{28}$$

where the supremum is over all probability measures $P, Q$ such that $P \ll Q$ and $P \ne Q$.

*Proof:* See [17, Theorem 5]. ■

*Remark 5:* Csiszár [5, Theorem 2] showed that if $f(0)$ and $f^\star(0)$ are finite and $P \ll Q$, then there exists a constant $C_f > 0$ which depends only on $f$ such that $D_f(P\|Q) \le C_f\sqrt{|P - Q|}$. Note that, if $|P - Q| < 1$, then this inequality is superseded by (26) where the constant is not only explicit but is the best possible according to Theorem 4.

A direct application of Theorem 4 yields

*Corollary 1:*

$$\sup_{P \ne Q} \frac{d_2^2(P,Q)}{|P - Q|} = \frac{1}{2}, \tag{29}$$

$$\sup_{P \ne Q} \frac{d_2^2(P,Q) + d_2^2(Q,P)}{|P - Q|} = 1 \tag{30}$$

where the supremum in (29) is over all $P \ll Q$ with $P \ne Q$, and the supremum in (30) is over all $P \ll\gg Q$ with $P \ne Q$.

*Proof:* See [17, Corollary 1]. ■

*Remark 6:* The results in (29) and (30) form counterparts of (22).

## III. Bounded Relative Information

In this section we show that it is possible to find bounds among $f$-divergences without requiring a strong condition of functional domination (see Section II) as long as the relative information is upper and/or lower bounded almost surely.

### A. Definition of $\beta_1$ and $\beta_2$.

The following notation is used throughout the rest of the paper. Given a pair of probability measures $(P, Q)$ on the same measurable space, denote $\beta_1, \beta_2 \in [0,1]$ by

$$\beta_1 = \exp\big(-D_\infty(P\|Q)\big), \tag{31}$$

$$\beta_2 = \exp\big(-D_\infty(Q\|P)\big) \tag{32}$$

with the convention that if $D_\infty(P\|Q) = \infty$, then $\beta_1 = 0$, and if $D_\infty(Q\|P) = \infty$, then $\beta_2 = 0$. Note that if $\beta_1 > 0$, then $P \ll Q$, while $\beta_2 > 0$ implies $Q \ll P$. Furthermore, if $P \ll\gg Q$, then with $Y \sim Q$,

$$\beta_1 = \operatorname{ess\,inf} \frac{\mathrm{d}Q}{\mathrm{d}P}(Y) = \left(\operatorname{ess\,sup} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\right)^{-1}, \tag{33}$$

$$\beta_2 = \operatorname{ess\,inf} \frac{\mathrm{d}P}{\mathrm{d}Q}(Y) = \left(\operatorname{ess\,sup} \frac{\mathrm{d}Q}{\mathrm{d}P}(Y)\right)^{-1}. \tag{34}$$

The following examples illustrate important cases in which $\beta_1$ and $\beta_2$ are positive.

*Example 1:* (*Gaussian distributions.*) Let $P$ and $Q$ be Gaussian probability measures with equal means, and variances $\sigma_0^2$ and $\sigma_1^2$ respectively. Then,

$$\beta_1 = \frac{\sigma_0}{\sigma_1} 1\{\sigma_0 \le \sigma_1\}, \tag{35}$$

$$\beta_2 = \frac{\sigma_1}{\sigma_0} 1\{\sigma_1 \le \sigma_0\}. \tag{36}$$

*Example 2:* (*Shifted Laplace distributions.*) Let $P$ and $Q$ be the probability measures whose probability density functions are, respectively, given by $f_\lambda(\cdot - a_0)$ and $f_\lambda(\cdot - a_1)$ with

$$f_\lambda(x) = \tfrac{\lambda}{2}\, \exp(-\lambda|x|), \quad x \in \mathbb{R} \tag{37}$$

where $\lambda > 0$. In this case, (37) gives

$$\frac{\mathrm{d}P}{\mathrm{d}Q}(x) = \exp\big(\lambda(|x - a_1| - |x - a_0|)\big), \quad x \in \mathbb{R} \tag{38}$$

which yields

$$\beta_1 = \beta_2 = \exp\big(-\lambda\, |a_1 - a_0|\big) \in (0, 1]. \tag{39}$$

### B. Basic Tool

Since $\beta_1 = 1 \Leftrightarrow \beta_2 = 1 \Leftrightarrow P = Q$, it is advisable to avoid trivialities by excluding that case.

*Theorem 5:* Let $f$ and $g$ satisfy the assumptions in Theorem 1, and assume that $(\beta_1, \beta_2) \in [0,1)^2$. Then,

$$D_f(P\|Q) \le \kappa^\star\, D_g(P\|Q) \tag{40}$$

where

$$\kappa^\star = \sup_{\beta \in (\beta_2, 1) \cup (1, \beta_1^{-1})} \kappa(\beta) \tag{41}$$

and $\kappa(\cdot)$ is defined in (11).

*Proof:* See [17, Theorem 5]. ■

Note that if $\beta_1 = \beta_2 = 0$, then Theorem 5 does not improve upon Theorem 1a).

*Remark 7:* In the application of Theorem 5, it is often convenient to make use of the freedom afforded by Remark 1 and choose the corresponding offsets such that:

- the positivity property of $g$ required by Theorem 5 is satisfied;
- the lowest $\kappa^\star$ is obtained.

*Remark 8:* Similarly to the proof of Theorem 1b), under the conditions therein, one can verify that the constants in Theorem 5 are the best possible among all probability measures $P, Q$ with given $(\beta_1, \beta_2) \in [0,1)^2$.

*Remark 9:* Note that if we swap the assumptions on $f$ and $g$ in Theorem 5, the same result translates into

$$\inf_{\beta \in (\beta_2, 1) \cup (1, \beta_1^{-1})} \kappa(\beta) \cdot D_g(P\|Q) \leq D_f(P\|Q). \tag{42}$$

Furthermore, provided both $f$ and $g$ are positive (except at $t = 1$) and $\kappa$ is monotonically increasing, Theorem 5 and (42) result in

$$\kappa(\beta_2)\, D_g(P\|Q) \leq D_f(P\|Q) \tag{43}$$

$$\leq \kappa(\beta_1^{-1})\, D_g(P\|Q). \tag{44}$$

In this case, if $\beta_1 > 0$, sometimes it is convenient to replace $\beta_1 > 0$ with $\beta_1' \in (0, \beta_1)$ at the expense of loosening the bound. A similar observation applies to $\beta_2$.

*Example 3:* If $f(t) = (t-1)^2$ and $g(t) = |t-1|$, we get

$$\chi^2(P\|Q) \leq \max\{\beta_1^{-1} - 1, 1 - \beta_2\}\ |P - Q|. \tag{45}$$

### C. Bounds on $\frac{D(P\|Q)}{D(Q\|P)}$

The remaining part of this section is devoted to various applications of Theorem 5. From this point, we make use of the definition of $r\colon (0,\infty) \to [0,\infty)$ in (8).

An illustrative application of Theorem 5 gives upper and lower bounds on the ratio of relative entropies.

*Theorem 6:* Let $P \ll\gg Q$, $P \neq Q$, and $(\beta_1, \beta_2) \in (0,1)^2$. Let $\kappa\colon (0,1) \cup (1,\infty) \to (0,\infty)$ be defined as

$$\kappa(t) = \frac{t \log t + (1-t)\, \log e}{(t-1) \log e - \log t}. \tag{46}$$

Then,

$$\kappa(\beta_2) \leq \frac{D(P\|Q)}{D(Q\|P)} \leq \kappa(\beta_1^{-1}). \tag{47}$$

*Proof:* See [17, Theorem 6]. ■

### D. Reverse Samson's Inequality

The next result gives a counterpart to Samson's inequality (21).

*Theorem 7:* Let $(\beta_1, \beta_2) \in (0,1)^2$. Then,

$$\inf \frac{d_2^2(P,Q) + d_2^2(Q,P)}{D(P\|Q)} = \min\{\kappa(\beta_1^{-1}),\ \kappa(\beta_2)\} \tag{48}$$

where the infimum is over all $P \ll Q$ with given $(\beta_1, \beta_2)$, and where $\kappa\colon (0,1) \cup (1,\infty) \to \left(0, \frac{2}{\log e}\right)$ is given by

$$\kappa(t) = \frac{(t-1)^2}{r(t)\ \max\{1, t\}}, \quad t \in (0,1) \cup (1,\infty). \tag{49}$$

*Proof:* See [17, Theorem 7]. ■

### E. Local Behavior of $f$-Divergences

Another application of Theorem 5 shows that the local behavior of $f$-divergences differs by only a constant, provided that the first distribution approaches the reference measure in a certain strong sense.

*Theorem 8:* Suppose that $\{P_n\}$, a sequence of probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$, converges to $Q$ (another probability measure on the same space) in the sense that, for $Y \sim Q$,

$$\lim_{n\to\infty} \operatorname{ess\,sup} \frac{\mathrm{d}P_n}{\mathrm{d}Q}(Y) = 1 \tag{50}$$

where it is assumed that $P_n \ll Q$ for all sufficiently large $n$. If $f$ and $g$ are convex on $(0,\infty)$ and they are positive except at $t = 1$ (where they are 0), then

$$\lim_{n\to\infty} D_f(P_n\|Q) = \lim_{n\to\infty} D_g(P_n\|Q) = 0, \tag{51}$$

and

$$\min\{\kappa(1^-), \kappa(1^+)\} \leq \lim_{n\to\infty} \frac{D_f(P_n\|Q)}{D_g(P_n\|Q)} \leq \max\{\kappa(1^-), \kappa(1^+)\} \tag{52}$$

where we have indicated the left and right limits of the function $\kappa(\cdot)$, defined in (11), at 1 by $\kappa(1^-)$ and $\kappa(1^+)$, respectively.

*Proof:* See [17, Theorem 9]. ■

*Corollary 2:* Let $\{P_n \ll Q\}$ converge to $Q$ in the sense of (50). Then, $D(P_n\|Q)$ and $D(Q\|P_n)$ vanish as $n \to \infty$ with

$$\lim_{n\to\infty} \frac{D(P_n\|Q)}{D(Q\|P_n)} = 1. \tag{53}$$

*Corollary 3:* Let $\{P_n \ll Q\}$ converge to $Q$ in the sense of (50). Then, $\chi^2(P_n\|Q)$ and $D(P_n\|Q)$ vanish as $n \to \infty$ with

$$\lim_{n\to\infty} \frac{D(P_n\|Q)}{\chi^2(P_n\|Q)} = \tfrac{1}{2} \log e. \tag{54}$$

Note that (54) is known in the finite alphabet case [7, Theorem 4.1]).

### F. Strengthened Jensen's inequality

Bounding away from zero a certain density between two probability measures enables the following strengthened version of Jensen's inequality, which generalizes a result in [11, Theorem 1].

*Lemma 1:* Let $f\colon \mathbb{R} \to \mathbb{R}$ be a convex function, $P_1 \ll P_0$ be probability measures defined on a measurable space $(\mathcal{A}, \mathscr{F})$, and fix an arbitrary random transformation $P_{Z|X}\colon \mathcal{A} \to \mathbb{R}$. Denote[4] $P_0 \to P_{Z|X} \to P_{Z_0}$, and $P_1 \to P_{Z|X} \to P_{Z_1}$. Then,

$$\beta \left(\mathbb{E}\left[f(\mathbb{E}[Z_0|X_0])\right] - f(\mathbb{E}[Z_0])\right)$$

$$\leq \mathbb{E}[f(\mathbb{E}[Z_1|X_1])] - f(\mathbb{E}[Z_1]) \tag{55}$$

[4]We follow the notation in [20] where $P_0 \to P_{Z|X} \to P_{Z_0}$ means that the marginal probability measures of the joint distribution $P_0 P_{Z|X}$ are $P_0$ and $P_{Z_0}$.

where $X_0 \sim P_0$, $X_1 \sim P_1$, and

$$\beta \triangleq \operatorname{ess\,inf} \frac{\mathrm{d}P_1}{\mathrm{d}P_0}\,(X_0). \tag{56}$$

*Proof:* See [17, Lemma 1]. ■

*Remark 10:* Letting $Z = X$, and choosing $P_0$ so that $\beta = 0$ (e.g., $P_1$ is a restriction of $P_0$ to an event of $P_0$-probability less than 1), (55) becomes Jensen's inequality $f(\mathbb{E}[X_1]) \leq \mathbb{E}[f(X_1)]$.

Lemma 1 finds the following application to the derivation of $f$-divergence inequalities.

*Theorem 9:* Let $f\colon (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. Fix $P \ll Q$ on the same space with $(\beta_1, \beta_2) \in [0,1)^2$ and let $X \sim P$. Then,

$$\begin{aligned} \beta_2\, D_f(P\|Q) &\leq \mathbb{E}\left[f\left(\exp(\imath_{P\|Q}(X))\right)\right] - f\big(1+\chi^2(P\|Q)\big) \\ &\leq \beta_1^{-1}\, D_f(P\|Q). \end{aligned} \tag{57}$$

Specializing Theorem 9 to the convex function on $(0,\infty)$ where $f(t) = -\log t$ sharpens the inequality

$$D(P\|Q) \leq \log\left(1+\chi^2(P\|Q)\right) \tag{58}$$
$$\leq \chi^2(P\|Q)\log e. \tag{59}$$

under the assumption of bounded relative information.

*Theorem 10:* Fix $P \ll\gg Q$ such that $(\beta_1, \beta_2) \in (0,1)^2$. Then,

$$\beta_2\, D(Q\|P) \leq \log\big(1+\chi^2(P\|Q)\big) - D(P\|Q) \tag{60}$$
$$\leq \beta_1^{-1}\, D(Q\|P). \tag{61}$$

## IV. Reverse Pinsker Inequalities

It is not possible to lower bound $|P-Q|$ solely in terms of $D(P\|Q)$ since for an arbitrary small $\epsilon > 0$ and an arbitrary large $\lambda > 0$, we can construct examples with $|P-Q| < \epsilon$ and $\lambda < D(P\|Q) < \infty$. As in Section III, the following result involves the bounds on the relative information.

*Theorem 11:* If $\beta_1 \in (0,1)$ and $\beta_2 \in [0,1)$, then,

$$D(P\|Q) \leq \tfrac{1}{2}\left(\varphi(\beta_1^{-1}) - \varphi(\beta_2)\right)\,|P-Q| \tag{62}$$

where $\varphi\colon [0,\infty) \to [0,\infty)$ is given by

$$\varphi(t) = \begin{cases} 0 & t = 0 \\ \frac{t\log t}{t-1} & t \in (0,1)\cup(1,\infty) \\ \log e & t = 1. \end{cases} \tag{63}$$

*Proof:* See [17, Theorem 23]. ■

*Remark 11:* Note that for Theorem 11 to give a nontrivial result, it is necessary that the relative information be upper bounded, namely $\beta_1 > 0$. However, we still get a nontrivial bound if $\beta_2 = 0$.

In the following, we assume that $P$ and $Q$ are probability measures defined on a common finite set $\mathcal{A}$, and $Q$ is strictly positive on $\mathcal{A}$ with $|\mathcal{A}| \geq 2$.

*Theorem 12:* Let $Q_{min} = \min_{a\in\mathcal{A}} Q(a)$, then

$$D(P\|Q) \leq \log\left(1 + \frac{|P-Q|^2}{2Q_{\min}}\right). \tag{64}$$

Furthermore, if $Q \ll P$ and $\beta_2$ is defined as in (32), then the following tightened bound holds:

$$D(P\|Q) \leq \log\left(1 + \frac{|P-Q|^2}{2Q_{\min}}\right) - \tfrac{1}{2}\beta_2|P-Q|^2\log e. \tag{65}$$

*Proof:* See [17, Theorem 25]. ■

*Remark 12:* The result in (64) improves the inequality by Csiszár and Talata [8, p. 1012]:

$$D(P\|Q) \leq \left(\frac{\log e}{Q_{\min}}\right)\cdot|P-Q|^2. \tag{66}$$

For further reverse Pinsker Inequalities and some of their implications, see [17, Section 6].

## References


[1] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistics Society*, series B, vol. 28, no. 1, pp. 131–142, 1966.

[2] A. Basu, H. Shioya and C. Park, "Statistical Inference: The Minimum Distance Approach," *Chapman & Hall/ CRC Monographs on Statistics and Applied Probability*, vol. 120, CRC Press, Boca Raton, Florida, USA, 2011.

[3] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, 2013.

[4] I. Csiszár, "Eine Informationstheoretische Ungleichung und ihre Anwendung auf den Bewis der Ergodizität von Markhoffschen Ketten," *Publ. Math. Inst. Hungar. Acad. Sci.*, vol. 8, pp. 85–108, January 1963.

[5] I. Csiszár, "On topological properties of $f$-divergences," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 329–339, 1967.

[6] I. Csiszár, "A class of measures of informativity of observation channels," *Periodica Mathematicarum Hungarica*, vol. 2, no. 1, pp. 191–23, March 1972.

[7] I. Csiszár and P. C. Shields, "Information Theory and Statistics: A Tutorial", *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, 2004.

[8] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Trans. on Information Theory*, vol. 52, no. 3, pp. 1007–1016, March 2006.

[9] P. Diaconis and L. Saloff-Coste, "Logarithmic Sobolev inequalities for finite Markov chains," *Annals of Applied Probability*, vol. 6, pp. 695–750, 1996.

[10] S. S. Dragomir, "Upper and lower bounds for Csiszár $f$-divergence in terms of the Kullback-Leibler distance and applications," *Inequalities for Csiszár f-Divergence in Information Theory, RGMIA Monographs*, 2000.

[11] S. S. Dragomir, "Bounds for the normalized Jensen functional," *Bulletin of the Australian Mathematical Society*, vol. 74, no. 3, pp. 471–478, 2006.

[12] A. Guntuboyina, S. Saha and G. Schiebinger, "Sharp inequalities for $f$-divergences," *IEEE Trans. on Information Theory*, vol. 60, no. 1, pp. 104–121, Jan. 2014.

[13] F. Liese and I. Vajda, *Convex Statistical Distances*, Teubner-Texte Zur Mathematik, vol. 95, Germany, 1987.

[14] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. on Information Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[15] K. Marton, "A measure concentration inequality for contracting Markov chains," *Geometric and Functional Analysis*, vol. 6, pp. 556–571, 1996.

[16] P. M. Samson, "Concentration of measure inequalities for Markov chains and $\phi$-mixing processes," *Annals of Probability*, vol. 28, no. 1, pp. 416–461, January 2000.

[17] I. Sason and S. Verdú, "$f$-divergence inequalities," to appear in the *IEEE Trans. on Information Theory*, vol. 62, 2016. Available at `http://arxiv.org/abs/1508.00335`.

[18] I. Vajda, "On $f$-divergence and singularity of probability measures," *Periodica Mathematica Hungarica*, vol. 2, no. 1–4, pp. 223–234, 1972.

[19] I. Vajda, "On metric divergences of probability measures," *Kybernetika*, vol. 45, no. 6, pp. 885–900, 2009.

[20] S. Verdú, *Information Theory*, in preparation.