Strongly Convex Divergences

James Melbourne

Department of Electrical and Computer Engineering, University of Minnesota-Twin Cities, Minneapolis, MN 55455, USA; melbo013@umn.edu

Received: 2 September 2020; Accepted: 9 November 2020; Published: 21 November 2020

Abstract: We consider a sub-class of the $f$-divergences satisfying a stronger convexity property, which we refer to as strongly convex, or $\kappa$-convex divergences. We derive new and old relationships, based on convexity arguments, between popular $f$-divergences.

Keywords: information measures; $f$-divergence; hypothesis testing; total variation; skew-divergence; convexity; Pinsker’s inequality; Bayes risk; Jensen–Shannon divergence

1. Introduction

The concept of an $f$-divergence, introduced independently by Ali-Silvey [1], Morimoto [2], and Csiszár [3], unifies several important information measures between probability distributions, as integrals of a convex function $f$, composed with the Radon–Nikodym of the two probability distributions. (An additional assumption can be made that $f$ is strictly convex at 1, to ensure that $D_f(\mu||\nu) > 0$ for $\mu \neq \nu$. This obviously holds for any $f''(1) > 0$, and can hold for some $f$-divergences without classical derivatives at 0, for instance the total variation is strictly convex at 1. An example of an $f$-divergence not strictly convex is provided by the so-called “hockey-stick” divergence, where $f(x) = (x - \gamma)_+$, see [4–6].) For a convex function $f: (0, \infty) \to \mathbb{R}$ such that $f(1) = 0$, and measures $P$ and $Q$ such that $P \ll Q$, the $f$-divergence from $P$ to $Q$ is given by $D_f(P||Q) := \int f\left(\frac{dP}{dQ}\right) dQ$. The canonical example of an $f$-divergence, realized by taking $f(x) = x \log x$, is the relative entropy (often called the KL-divergence), which we denote with the subscript $f$ omitted. $f$-divergences inherit many properties enjoyed by this special case; non-negativity, joint convexity of arguments, and a data processing inequality. Other important examples include the total variation, the $\chi^2$-divergence, and the squared Hellinger distance. The reader is directed to Chapter 6 and 7 of [7] for more background.

We are interested in how stronger convexity properties of $f$ give improvements of classical $f$-divergence inequalities. More explicitly, we consider consequences of $f$ being $\kappa$-convex, in the sense that the map $x \mapsto f(x) - kx^2/2$ is convex. This is in part inspired by the work of Sason [8], who demonstrated that divergences that are $\kappa$-convex satisfy “stronger than $\chi^2$” data-processing inequalities.

Perhaps the most well known example of an $f$-divergence inequality is Pinsker’s inequality, which bounds the square of the total variation above by a constant multiple of the relative entropy. That is for probability measures $P$ and $Q$, $|P - Q|^2_{TV} \leq c \ D(P||Q)$. The optimal constant is achieved for Bernoulli measures, and under our conventions for total variation, $c = 1/2 \log e$. Many extensions and sharpenings of Pinsker’s inequality exist (for examples, see [9–11]). Building on the work of Guntuboyina [9] and Topsøe [11], we achieve a further sharpening of Pinsker’s inequality in Theorem 9.

Aside from the total variation, most divergences of interest have stronger than affine convexity, at least when $f$ is restricted to a sub-interval of the real line. This observation is especially relevant to the situation in which one wishes to study $D_f(P||Q)$ in the existence of a bounded Radon–Nikodym derivative $\frac{dP}{dQ} \in (a, b) \subseteq (0, \infty)$. One naturally obtains such bounds for skew divergences. That is divergences of the form $(P, Q) \mapsto D_f((1-t)P + tQ||(1-s)P + sQ)$ for $t, s \in [0, 1]$, as in this case,
We also derive upper and lower total variation bounds for Nielsen’s generalized Jensen–Shannon Entropy, the Jensen–Shannon divergence introduced recently by Nielsen [17] (based on skewing the relative measures). Important examples of skew-divergences include the skew divergence defined on $\mu$ with respect to a common reference measure $X$ (functions defined on $(\mathbb{R}^d,\mathcal{B},\mu)$). Notation decomposition induced by the Bayes estimator. The notations give a table of examples of skew-divergences realized as linear combinations of skewed divergences.

Let us outline the paper. In Section 2, we derive elementary results of $\kappa$-convex divergences and give a table of examples of $\kappa$-convex divergences. We demonstrate that $\kappa$-convex divergences can be lower bounded by the $\chi^2$-divergence, and that the joint convexity of the map $(P,Q) \mapsto D_f(P||Q)$ can be sharpened under $\kappa$-convexity conditions on $f$. As a consequence, we obtain bounds between the mean square total variation distance of a set of distributions from their barycenter, and the average $f$-divergence from the set to the barycenter.

In Section 3, we investigate general skewing of $f$-divergences. In particular, we introduce the skew-symmetrization of an $f$-divergence, which recovers the Jensen–Shannon divergence and the Vincze–Le Cam divergences as special cases. We also show that a scaling of the Vincze–Le Cam divergence is minimal among skew-symmetrizations of $\kappa$-convex divergences on $(0,2)$. We then consider linear combinations of skew divergences and show that a generalized Vincze–Le Cam divergence (based on skewing the $\chi^2$-divergence) can be upper bounded by the generalized Jensen–Shannon divergence introduced recently by Nielsen [17] (based on skewing the relative entropy), reversing the classical convexity bounds $D(P||Q) \leq \log(1 + \chi^2(P||Q)) \leq \log \chi^2(P||Q)$. We also derive upper and lower total variation bounds for Nielsen’s generalized Jensen–Shannon divergence.

In Section 4, we consider a family of densities $\{p_i\}$ weighted by $\lambda_i$, and a density $q$. We use the Bayes estimator $T(x) = \arg\max_i \lambda_i p_i(x)$ to derive a convex decomposition of the barycenter $p = \sum \lambda_i p_i$ and of $q$, each into two auxiliary densities. (Recall, a Bayes estimator is one that minimizes the expected value of a loss function. By the assumptions of our model, that $P(\theta = i) = \lambda_i$, and $P(X \in A|\theta = i) = \int_A p_i(x) dx$, we have $\mathbb{E}(\ell(\hat{\theta}, \theta)) = 1 - \int \lambda_{\theta(x)} p_{\theta(x)}(x) dx$ for the loss function $\ell(i,j) = 1 - \delta_i(j)$ and any estimator $\hat{\theta}$. It follows that $\mathbb{E}(\ell(\hat{\theta}, \theta)) \geq \mathbb{E}(\ell(\theta, T))$ by $\lambda_{\hat{\theta}(x)} p_{\hat{\theta}(x)}(x) \leq \lambda_T(x)p_T(x)$.) Thus, $T$ is a Bayes estimator associated to $\ell$. We use this decomposition to sharpen, for $\kappa$-convex divergences, an elegant theorem of Guntuboyina [9] that generalizes Fano and Pinsker’s inequality to $f$-divergences. We then demonstrate explicitly, using an argument of Topsøe, how our sharpening of Guntuboyina’s inequality gives a new sharpening of Pinsker’s inequality in terms of the convex decomposition induced by the Bayes estimator.

**Notation**

Throughout, $f$ denotes a convex function $f : (0, \infty) \to \mathbb{R} \cup \{\infty\}$, such that $f(1) = 0$. For a convex function defined on $(0, \infty)$, we define $f(0) := \lim_{x \to 0} f(x)$. We denote by $f^*$, the convex function $f^* : (0, \infty) \to \mathbb{R} \cup \{\infty\}$ defined by $f^*(x) = xf(x^{-1})$. We consider Borel probability measures $P$ and $Q$ on a Polish space $X$ and define the $f$-divergence from $P$ to $Q$, via densities $p$ for $P$ and $q$ for $Q$ with respect to a common reference measure $\mu$ as

$$D_f(p||q) = \int_X f \left( \frac{p}{q} \right) q d\mu = \int_{\{pq > 0\}} qf \left( \frac{p}{q} \right) d\mu + f(0)Q(\{p = 0\}) + f^*(0)P(\{q = 0\}).$$

We note that this representation is independent of $\mu$, and such a reference measure always exists, take $\mu = P + Q$ for example.

For $t, s \in [0, 1]$, define the binary $f$-divergence

$$D_f(t|s) := sf \left( \frac{1}{s} \right) + (1 - s)f \left( \frac{1 - t}{1 - s} \right)$$

(2)
with the conventions, \( f(0) = \lim_{t \to 0^+} f(t) \), \( 0 f(0/0) = 0 \), and \( 0 f(a/0) = a \lim_{t \to \infty} f(t)/t \). For a random variable \( X \) and a set \( A \), we denote the probability that \( X \) takes a value in \( A \) by \( \mathbb{P}(X \in A) \), the expectation of the random variable by \( \mathbb{E}X \), and the variance by \( \text{Var}(X) := \mathbb{E}[X - \mathbb{E}X]^2 \). For a probability measure \( \mu \) satisfying \( \mu(A) = \mathbb{P}(X \in A) \) for all Borel \( A \), we write \( X \sim \mu \), and, when there exists a probability density function such that \( \mathbb{P}(X \in A) = \int_A f(x) d\gamma(x) \) for a reference measure \( \gamma \), we write \( X \sim f \). For a probability measure \( \mu \) on \( \mathcal{X} \), and an \( L^2 \) function \( f : \mathcal{X} \to \mathbb{R} \), we denote \( \text{Var}_\mu(f) := \mathbb{V}(f(X)) \) for \( X \sim \mu \).

2. Strongly Convex Divergences

**Definition 1.** A \( \mathbb{R} \cup \{\infty\} \)-valued function \( f \) on a convex set \( K \subseteq \mathbb{R} \) is \( \kappa \)-convex when \( x, y \in K \) and \( t \in [0, 1] \) implies

\[
f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y) - \kappa t(1 - t)(x - y)^2/2. \tag{3}
\]

For example, when \( f \) is twice differentiable, (3) is equivalent to \( f''(x) \geq \kappa \) for \( x \in K \). Note that the case \( \kappa = 0 \) is just usual convexity.

**Proposition 1.** For \( f : K \to \mathbb{R} \cup \{\infty\} \) and \( \kappa \in [0, \infty) \), the following are equivalent:

1. \( f \) is \( \kappa \)-convex.
2. The function \( f - \kappa(t - a)^2/2 \) is convex for any \( a \in \mathbb{R} \).
3. The right handed derivative, defined as \( f'_+(t) := \lim_{h \to 0} \frac{f(t+h) - f(t)}{h} \) satisfies,

\[
f'_+(t) \geq f'_+(s) + \kappa(t - s)
\]

for \( t \geq s \).

**Proof.** Observe that it is enough to prove the result when \( \kappa = 0 \), where the proposition is reduced to the classical result for convex functions. \( \square \)

**Definition 2.** An \( f \)-divergence \( D_f \) is \( \kappa \)-convex on an interval \( K \) for \( \kappa \geq 0 \) when the function \( f \) is \( \kappa \)-convex on \( K \).

Table 1 lists some \( \kappa \)-convex \( f \)-divergences of interest to this article.

| Divergence          | \( f \)                      | \( \kappa \) | Domain       |
|---------------------|-----------------------------|-------------|--------------|
| relative entropy (KL) | \( t \log t \)              | \( \frac{1}{M} \) | \( (0, M] \) |
| total variation     | \( \frac{t-1}{\sqrt{t}} \)  | 0           | \( (0, \infty) \) |
| Pearson’s \( \chi^2 \) | \( (t-1)^2 \)               | 2           | \( (0, \infty) \) |
| squared Hellinger   | \( 2(1 - \sqrt{t}) \)       | \( M^{-1}/2 \) | \( (0, M] \) |
| reverse relative entropy | \( -\log t \)             | \( 1/M^2 \) | \( (0, M] \) |
| Vincze-Le Cam       | \( \frac{t-1}{t+1} \)       | \( \frac{M}{M+1} \) | \( (0, M] \) |
| Jensen–Shannon      | \( (t+1)\log\frac{t+1}{t} + t\log t \) | \( M/(M+1) \) | \( (0, M] \) |
| Neyman’s \( \chi^2 \) | \( \frac{t}{1+1} - 1 \)     | \( 2/M^2 \) | \( (0, M] \) |
| Sason’s \( s \)     | \( \log(s+t)(s+1)^{s+t} - \log(s+1)(s+1)^{s+1} \) | \( 2\log(s+M)+3 \) | \( [M, \infty), s > e^{-3/2} \) |
| \( a \)-divergence   | \( \frac{a(t^{a+1})}{1-t^a} \) | \( M^{a+2} \) | \( \begin{cases} [M, \infty), & a > 3 \\ (0, M], & a < 3 \end{cases} \) |

Observe that we have taken the normalization convention on the total variation (the total variation for a signed measure \( \mu \) on a space \( X \) can be defined through the Hahn-Jordan decomposition of the measure into non-negative measures \( \mu^+ \) and \( \mu^- \) such that \( \mu = \mu^+-\mu^- \), as \( \|\mu\| = \mu^+(X) + \mu^-(X) \).
(see [18]); in our notation, \( |\mu|_{TV} = \|\mu\|/2 \) which we denote by \( |P - Q|_{TV} \), such that \( |P - Q|_{TV} = \sup_A |P(A) - Q(A)| \leq 1 \). In addition, note that the \( \alpha \)-divergence interpolates Pearson’s \( \chi^2 \)-divergence when \( \alpha = 3 \), one half Neyman’s \( \chi^2 \)-divergence when \( \alpha = -3 \), the squared Hellinger divergence when \( \alpha = 0 \), and has limiting cases, the relative entropy when \( \alpha = 1 \) and the reverse relative entropy when \( \alpha = -1 \). If \( f \) is \( \kappa \)-convex on \([a, b]\), then recalling its dual divergence \( f^*(x) := xf(x^{-1}) \) is \( \kappa a^3 \)-convex on \([\frac{1}{a}, \frac{1}{b}]\). Recall that \( f^* \) satisfies the equality \( D_{f^*}(P||Q) = D_f(Q||P) \). For brevity, we use \( \chi^2 \)-divergence to refer to the Pearson \( \chi^2 \)-divergence, and we articulate Neyman’s \( \chi^2 \) explicitly when necessary.

The next lemma is a restatement of Jensen’s inequality.

**Lemma 1.** If \( f \) is \( \kappa \)-convex on the range of \( X \),

\[
\mathbb{E}f(X) \geq f(\mathbb{E}(X)) + \frac{\kappa}{2} \text{Var}(X).
\]

**Proof.** Apply Jensen’s inequality to \( f(x) - \kappa x^2/2 \). \( \square \)

For a convex function \( f \) such that \( f(1) = 0 \) and \( c \in \mathbb{R} \), the function \( \tilde{f}(t) = f(t) + c(t - 1) \) remains a convex function, and what is more satisfies

\[
D_f(P||Q) = D_{\tilde{f}}(P||Q)
\]

since \( \int c(p/q - 1)qd\mu = 0 \).

**Definition 3 (\( \chi^2 \)-divergence).** For \( f(t) = (t - 1)^2 \), we write

\[
\chi^2(P||Q) := D_f(P||Q).
\]

We pursue a generalization of the following bound on the total variation by the \( \chi^2 \)-divergence [19–21].

**Theorem 1 ([19–21]).** For measures \( P \) and \( Q \),

\[
|P - Q|^2_{TV} \leq \frac{\chi^2(P||Q)}{2}.
\]  

(4)

We mention the work of Harremos and Vadja [20], in which it is shown, through a characterization of the extreme points of the joint range associated to a pair of \( f \)-divergences (valid in general), that the inequality characterizes the “joint range”, that is, the range of the function \( (P, Q) \mapsto (|P - Q|_{TV}, \chi^2(P||Q)) \). We use the following lemma, which shows that every strongly convex divergence can be lower bounded, up to its convexity constant \( \kappa > 0 \), by the \( \chi^2 \)-divergence,

**Lemma 2.** For a \( \kappa \)-convex \( f \),

\[
D_f(P||Q) \geq \frac{\kappa}{2} \chi^2(P||Q).
\]

**Proof.** Define a \( \tilde{f}(t) = f(t) - f'_+(1)(t - 1) \) and note that \( \tilde{f} \) defines the same \( \kappa \)-convex divergence as \( f \). Thus, we may assume without loss of generality that \( f'_+ \) is uniquely zero when \( t = 1 \). Since \( f \) is
κ-convex \( \phi : t \mapsto f(t) - \kappa(t-1)^2/2 \) is convex, and, by \( f'_+(1) = 0, \phi'_+(1) = 0 \) as well. Thus, \( \phi \) takes its minimum when \( t = 1 \) and hence \( \phi \geq 0 \) so that \( f(t) \geq \kappa(t-1)^2/2 \). Computing,

\[
D_f(P\|Q) = \int f \left( \frac{dP}{dQ} \right) dQ \\
\geq \frac{\kappa}{2} \int \left( \frac{dP}{dQ} - 1 \right)^2 dQ \\
= \frac{\kappa}{2} \chi^2(P\|Q).
\]

□

Based on a Taylor series expansion of \( f \) about 1, Nielsen and Nock ([22], [Corollary 1]) gave the estimate

\[
D_f(P\|Q) \approx \frac{f''(1)}{2} \chi^2(P\|Q)
\]

for divergences with a non-zero second derivative and \( P \) close to \( Q \). Lemma 2 complements this estimate with a lower bound, when \( f \) is \( \kappa \)-concave. In particular, if \( f''(1) = \kappa \), it shows that the approximation in (5) is an underestimate.

**Theorem 2.** For measures \( P \) and \( Q \), and a \( \kappa \)-convex divergence \( D_f \),

\[
|P - Q|^2_{TV} \leq \frac{D_f(P\|Q)}{\kappa}.
\]

**Proof.** By Lemma 2 and then Theorem 1,

\[
\frac{D_f(P\|Q)}{\kappa} \geq \frac{\chi^2(P\|Q)}{2} \geq |P - Q|_{TV}.
\]

□

The proof of Lemma 2 uses a pointwise inequality between convex functions to derive an inequality between their respective divergences. This simple technique was shown to have useful implications by Sason and Verdu in [6], where it appears as Theorem 1 and is used to give sharp comparisons in several \( f \)-divergence inequalities.

**Theorem 3** (Sason–Verdu [6]). For divergences defined by \( g \) and \( f \) with \( c f(t) \geq g(t) \) for all \( t \), then

\[
D_g(P\|Q) \leq c D_f(P\|Q).
\]

Moreover, if \( f'(1) = g'(1) = 0 \), then

\[
\sup_{P \neq Q} \frac{D_g(P\|Q)}{D_f(P\|Q)} = \sup_{t \neq 1} \frac{g(t)}{f(t)}.
\]

**Corollary 1.** For a smooth \( \kappa \)-convex divergence \( f \), the inequality

\[
D_f(P\|Q) \geq \frac{\kappa}{2} \chi^2(P\|Q)
\]

is sharp multiplicatively in the sense that

\[
\inf_{P \neq Q} \frac{D_f(P\|Q)}{\chi^2(P\|Q)} = \frac{\kappa}{2}.
\]
Thus, (9) follows.

In information geometry, a standard f-divergence is defined as an f-divergence satisfying the normalization \( f(1) = f'(1) = 0, f''(1) = 1 \) (see [23]). Thus, Corollary 1 shows that \( \frac{1}{2} \kappa^2 \) provides a sharp lower bound on every standard f-divergence that is 1-convex. In particular, the lower bound in Lemma 2 complimenting the estimate (5) is shown to be sharp.

**Proof.** Without loss of generality, we assume that \( f'(1) = 0 \). If \( f''(1) = \kappa + 2\epsilon \) for some \( \epsilon > 0 \), then taking \( g(t) = (t - 1)^2 \) and applying Theorem 3 and Lemma 2

\[
\sup_{\rho \neq \chi} \frac{D_\rho f(P) - D_\rho f(Q)}{D_f(P) - D_f(Q)} \leq \frac{2}{\kappa},
\]

(10)

Observe that, after two applications of L’Hospital,

\[
\lim_{\epsilon \to 0} \frac{g(1 + \epsilon)}{f(1 + \epsilon)} = \lim_{\epsilon \to 0} \frac{g'(1 + \epsilon)}{f'(1 + \epsilon)} = \frac{g''(1)}{f''(1)} = \frac{2}{\kappa} \leq \sup_{t \neq 1} \frac{g(t)}{f(t)}.
\]

Thus, (9) follows. \( \square \)

**Proposition 2.** When \( D_f \) is an f divergence such that \( f \) is \( \kappa \)-convex on \([a, b]\) and that \( P_\theta \) and \( Q_\theta \) are probability measures indexed by a set \( \Theta \) such that \( a \leq \frac{dP_\theta}{dQ_\theta}(x) \leq b \), holds for all \( \theta \) and \( P := \int_{\Theta} P_\theta d\mu(\theta) \) and \( Q := \int_{\Theta} Q_\theta d\mu(\theta) \) for a probability measure \( \mu \) on \( \Theta \), then

\[
D_f(P || Q) \leq \int_{\Theta} D_f(P_\theta || Q_\theta) d\mu(\theta) - \frac{\kappa}{2} \int_{\Theta} \int_X \left( \frac{dP_\theta}{dQ_\theta} - \frac{dP}{dQ} \right)^2 dQ d\mu,
\]

(11)

In particular, when \( Q_\theta = Q \) for all \( \theta \)

\[
D_f(P || Q) \leq \int_{\Theta} D_f(P_\theta || Q_\theta) d\mu(\theta) - \frac{\kappa}{2} \int_{\Theta} \int_X \left( \frac{dP_\theta}{dQ_\theta} - \frac{dP}{dQ} \right)^2 dQ d\mu(\theta)
\]

(12)

**Proof.** Let \( d\theta \) denote a reference measure dominating \( \mu \) so that \( d\mu = \varphi(\theta) d\theta \) then write \( v_\theta = v(\theta, x) = \frac{dQ_\theta}{dQ}(x) \varphi(\theta) \).

\[
D_f(P || Q) = \int_X f \left( \frac{dP}{dQ} \right) dQ
\]

\[
= \int_X f \left( \int_{\Theta} \frac{dP_\theta}{dQ_\theta} d\mu(\theta) \right) dQ
\]

(13)

By Jensen’s inequality, as in Lemma 1

\[
f \left( \int_{\Theta} \frac{dP_\theta}{dQ_\theta} v_\theta d\theta \right) \leq \int_{\Theta} f \left( \frac{dP_\theta}{dQ_\theta} \right) v_\theta d\theta - \frac{\kappa}{2} \int_{\Theta} \left( \frac{dP_\theta}{dQ_\theta} - \int_{\Theta} \frac{dP_\theta}{dQ_\theta} v_\theta d\theta \right)^2 v_\theta d\theta
\]
Integrating this inequality gives

\[
D_f(P||Q) \leq \int_X \left( \int_{\Theta} f\left( \frac{dP_\theta}{dQ_\theta} \right) v_\theta d\theta - \frac{\kappa}{2} \int_{\Theta} \left( \frac{dP_\theta}{dQ_\theta} - \int_{\Theta} \frac{dP_\theta}{dQ_\theta} v_\theta d\theta \right)^2 v_\theta d\theta \right) dQ \tag{14}
\]

Note that

\[
\int_X \int_{\Theta} \left( \frac{dP_\theta}{dQ_\theta} \right)^2 dQ \int_{\Theta} \frac{dP_\theta}{dQ_\theta} v_\theta d\theta = \int_{\Theta} \int_X \left( \frac{dP_\theta}{dQ_\theta} - \frac{dP}{dQ} \right)^2 dQ d\mu_\theta,
\]

and

\[
\int_X \int_{\Theta} f\left( \frac{dP_\theta}{dQ_\theta} \right) v(\theta,x) d\theta dQ = \int_{\Theta} \int_X f\left( \frac{dP_\theta}{dQ_\theta} \right) v(\theta,x) dQ d\theta
\]

\[
= \int_{\Theta} \int_X f\left( \frac{dP_\theta}{dQ_\theta} \right) dQ d\mu(\theta)
\]

\[
= \int_{\Theta} D(P_\theta||Q_\theta) d\mu(\theta)
\]

Inserting these equalities into (14) gives the result.

To obtain the total variation bound, one needs only to apply Jensen’s inequality,

\[
\int_X \left( \frac{dP_\theta}{dQ} - \frac{dP}{dQ} \right)^2 dQ \geq \left( \int_X \left| \frac{dP_\theta}{dQ} - \frac{dP}{dQ} \right| dQ \right)^2
\]

\[
= |P_\theta - P|_{TV}^2.
\]

\[
\square
\]

Observe that, taking \( Q = P = \int_{\Theta} P_\theta d\mu(\theta) \) in Proposition 2, one obtains a lower bound for the average \( f \)-divergence from the set of distribution to their barycenter, by the mean square total variation of the set of distributions to the barycenter,

\[
\kappa \int_{\Theta} |P_\theta - P|_{TV}^2 d\mu(\theta) \leq \int_{\Theta} D_f(P_\theta||P) d\mu(\theta).
\]

An alternative proof of this can be obtained by applying \( |P_\theta - P|_{TV}^2 \leq D_f(P_\theta||P)/\kappa \) from Theorem 2 pointwise.

The next result shows that, for \( f \) strongly convex, Pinsker type inequalities can never be reversed,

**Proposition 3.** Given \( f \) strongly convex and \( M > 0 \), there exists \( P, Q \) measures such that

\[
D_f(P||Q) \geq M|P - Q|_{TV}.
\]

**Proof.** By \( \kappa \)-convexity \( \phi(t) = f(t) - \kappa t^2/2 \) is a convex function. Thus, \( \phi(t) \geq \phi(1) + \phi'(1)(t - 1) \) and hence \( \lim_{t \to \infty} \frac{\phi(t)}{t^2} \geq \lim_{t \to \infty} \kappa t/2 + (f'_+(1) - \kappa) \left( 1 - \frac{1}{2} \right) = \infty \). Taking measures on the two points space \( P = \{1/2, 1/2\} \) and \( Q = \{1/2t, 1 - 1/2t\} \) gives \( D_f(P||Q) \geq \frac{1}{2} f(t) \) which tends to infinity with \( t \to \infty \), while \( |P - Q|_{TV} \leq 1 \). \( \square \)

In fact, building on the work of Basu-Shioya-Park [24] and Vadja [25], Sason and Verdú proved [6] that, for any \( f \) divergence, \( \sup_{P \neq Q} \frac{D_f(P||Q)}{|P - Q|_{TV}} = f(0) + f'(0) \). Thus, an \( f \)-divergence can be bounded above by a constant multiple of a the total variation, if and only if \( f(0) + f'(0) < \infty \). From this perspective, Proposition 3 is simply the obvious fact that strongly convex functions have super linear (at least quadratic) growth at infinity.
3. Skew Divergences

If we denote $C\text{ov}(0, \infty)$ to be quotient of the cone of convex functions $f$ on $(0, \infty)$ such that $f(1) = 0$ under the equivalence relation $f_1 \sim f_2$ when $f_1 - f_2 = c(x - 1)$ for $c \in \mathbb{R}$, then the map $f \mapsto D_f$ gives a linear isomorphism between $C\text{ov}(0, \infty)$ and the space of all $f$-divergences. The mapping $T : C\text{ov}(0, \infty) \to C\text{ov}(0, \infty)$ defined by $Tf = f^*$, which we recall $f^*(t) = tf(t^{-1})$, gives an involution of $C\text{ov}(0, \infty)$. Indeed, $D_{Tf}(P\|Q) = D_f(Q\|P)$, so that $D_{T(T(f))}(P\|Q) = D_f(P\|Q)$. Mathematically, skew divergences give an interpolation of this involution as

$$(P, Q) \mapsto D_f((1-t)P + tQ||(1-s)P + sQ)$$

gives $D_f(P\|Q)$ by taking $s = 1$ and $t = 0$ or yields $D_{f^*}(P\|Q)$ by taking $s = 0$ and $t = 1$.

Moreover, as mentioned in the Introduction, skewing imposes boundedness of the Radon–Nikodym derivative $\frac{dP}{dQ}$, which allows us to constrain the domain of $f$-divergences and leverage $\kappa$-convexity to obtain $f$-divergence inequalities in this section.

The following appears as Theorem III.1 in the preprint [26]. It states that skewing an $f$-divergence preserves its status as such. This guarantees that the generalized skew divergences of this section are indeed $f$-divergences. A proof is given in the Appendix A for the convenience of the reader.

**Theorem 4** (Melbourne et al [26]). For $t, s \in [0, 1]$ and a divergence $D_f$, then

$$S_f(P\|Q) := D_f((1-t)P + tQ||(1-s)P + sQ)$$

is an $f$-divergence as well.

**Definition 4.** For an $f$-divergence, its skew symmetrization,

$$\Delta_f(P\|Q) := \frac{1}{2}D_f\left(P\left|\|\frac{P+Q}{2}\right.\right) + \frac{1}{2}D_f\left(Q\left|\|\frac{P+Q}{2}\right.\right).$$

$\Delta_f$ is determined by the convex function

$$x \mapsto \frac{1 + x}{2} \left( f\left(\frac{2x}{1+x}\right) + f\left(\frac{2}{1+x}\right) \right).$$

(20)

Observe that $\Delta_f(P\|Q) = \Delta_f(Q\|P)$, and when $f(0) < \infty$, $\Delta_f(P\|Q) \leq \sup_{x \in [0, 2]} f(x) < \infty$ for all $P, Q$ since $\frac{dP}{d\frac{P+Q}{2}} = \frac{dQ}{d\frac{P+Q}{2}} \leq 2$. When $f(x) = x \log x$, the relative entropy’s skew symmetrization is the Jensen–Shannon divergence. When $f(x) = (x - 1)^2$ up to a normalization constant the $\chi^2$-divergence’s skew symmetrization is the Vincze–Le Cam divergence which we state below for emphasis. The work of Topsoe [11] provides more background on this divergence, where it is referred to as the triangular discrimination.

**Definition 5.** When $f(t) = \frac{(t-1)^2}{t+1}$, denote the Vincze–Le Cam divergence by

$$\Delta(P\|Q) := D_f(P\|Q).$$

If one denotes the skew symmetrization of the $\chi^2$-divergence by $\Delta_{\chi^2}$, one can compute easily from (20) that $\Delta_{\chi^2}(P\|Q) = \Delta(P\|Q)/2$. We note that although skewing preserves 0-convexity, by the above example, it does not preserve $\kappa$-convexity in general. The skew symmetrization of the $\chi^2$-divergence a 2-convex divergence while $f(t) = (t - 1)^2/(t + 1)$ corresponding to the Vincze–Le Cam divergence satisfies $f''(t) = \frac{8}{(t+1)^4}$, which cannot be bounded away from zero on $(0, \infty)$. 


Corollary 2. For an $f$-divergence such that $f$ is a $\kappa$-convex on $(0, 2)$,

$$\Delta_f(P||Q) \geq \frac{K}{4} \Delta(P||Q) = \frac{K}{2} \Delta_\chi^2(P||Q),$$

with equality when the $f(t) = (t - 1)^2$ corresponding the the $\chi^2$-divergence, where $\Delta_f$ denotes the skew symmetrized divergence associated to $f$ and $\Delta$ is the Vincze- Le Cam divergence.

Proof. Applying Proposition 2

$$0 = D_f\left(\frac{P + Q}{2}||\frac{Q + P}{2}\right)$$

$$\leq \frac{1}{2} D_f\left(P||\frac{Q + P}{2}\right) + \frac{1}{2} D_f\left(Q||\frac{Q + P}{2}\right) - \frac{K}{8} \int \left(\frac{2P}{P + Q} - \frac{2Q}{P + Q}\right)^2 d(P + Q)/2$$

$$= \Delta_f(P||Q) - \frac{K}{4} \Delta(P||Q).$$

□

When $f(x) = x \log x$, we have $f''(x) \geq \log \frac{e}{x}$ on $[0, 2]$, which demonstrates that up to a constant $\log \frac{e}{x}$ the Jensen–Shannon divergence bounds the Vincze–Le Cam divergence (see [11] for improvement of the inequality in the case of the Jensen–Shannon divergence, called the “capacitory discrimination” in the reference, by a factor of 2).

We now investigate more general, non-symmetric skewing in what follows.

Proposition 4. For $\alpha, \beta \in [0, 1]$, define

$$C(\alpha) := \begin{cases} 1 - \alpha & \text{when } \alpha \leq \beta, \\ \alpha & \text{when } \alpha > \beta, \end{cases}$$

and

$$S_{\alpha, \beta}(P||Q) := D((1 - \alpha)P + \alpha Q||(1 - \beta)P + \beta Q).$$

Then,

$$S_{\alpha, \beta}(P||Q) \leq C(\alpha) D_\infty(\alpha||\beta)|P - Q|_{TV},$$

where $D_\infty(\alpha||\beta) := \log \left(\max \left\{ \frac{\alpha}{\beta}, \frac{1 - \alpha}{1 - \beta}\right\}\right)$ is the binary $\infty$-Rényi divergence [27].

We need the following lemma originally proved by Audenart in the quantum setting [28]. It is based on a differential relationship between the skew divergence [12] and the [15] (see [29,30]).

Lemma 3 (Theorem III.1 [26]). For $P$ and $Q$ probability measures and $t \in [0, 1],$

$$S_{\alpha,t}(P||Q) \leq - \log t|P - Q|_{TV}.$$  (25)

Proof of Theorem 4. If $\alpha \leq \beta$, then $D_\infty(\alpha||\beta) = \log \frac{1 - \alpha}{1 - \beta}$ and $C(\alpha) = 1 - \alpha$. In addition,

$$(1 - \beta)P + \beta Q = t((1 - \alpha)P + \alpha Q) + (1 - t)Q$$

with $t = \frac{1 - \beta}{1 - \alpha}$, thus

$$S_{\alpha, t}(P||Q) = S_{\alpha,t}((1 - \alpha)P + \alpha Q||Q)$$

$$\leq (- \log t) |((1 - \alpha)P + \alpha Q) - Q|_{TV}$$

$$= C(\alpha) D_\infty(\alpha||\beta) |P - Q|_{TV},$$

22
where the inequality follows from Lemma 3. Following the same argument for $\alpha > \beta$, so that $C(\alpha) = \alpha$, $D_\infty(\alpha||\beta) = \log \frac{\beta}{\alpha}$, and
\[
(1 - \beta)P + \beta Q = t ((1 - \alpha)P + \alpha Q) + (1 - t)P
\]
for $t = \frac{\beta}{\alpha}$ completes the proof. Indeed,
\[
S_{\alpha,\beta}(P||Q) = S_{0,1}((1 - \alpha)P + \alpha Q||P) \\
\leq - \log t \left|\left|((1 - \alpha)P + \alpha Q) - P\right|\right|_{TV} \\
= C(\alpha) \, D_\infty(\alpha||\beta) \left|\left|P - Q\right|\right|_{TV}.
\]

We recover the classical bound [11,16] of the Jensen–Shannon divergence by the total variation.

**Corollary 3.** For probability measure $P$ and $Q$,
\[
\text{JSD}(P||Q) \leq \log 2 \, |P - Q|_{TV}
\]

**Proof.** Since $\text{JSD}(P||Q) = \frac{1}{2} S_{0,1}(P||Q) + \frac{1}{2} S_{1,\frac{1}{2}}(P||Q)$. \qed

Proposition 4 gives a sharpening of Lemma 1 of Nielsen [17], who proved $S_{\alpha,\beta}(P||Q) \leq D_\infty(\alpha||\beta)$, and used the result to establish the boundedness of a generalization of the Jensen–Shannon Divergence.

**Definition 6** (Nielsen [17]). For $p$ and $q$ densities with respect to a reference measure $\mu$, $w_i > 0$, such that $\sum_{i=1}^{n} w_i = 1$ and $\alpha_i \in [0,1]$, define
\[
\text{JS}^{\alpha,w}(p : q) = \sum_{i=1}^{n} w_i \, D((1 - \alpha_i)p + \alpha_i q || (1 - \bar{\alpha})p + \bar{\alpha}q)
\]
where $\sum_{i=1}^{n} w_i \alpha_i = \bar{\alpha}$.

Note that, when $n = 2$, $\alpha_1 = 1$, $\alpha_2 = 0$ and $w_1 = \frac{1}{2}$, $\text{JS}^{\alpha,w}(p : q) = \text{JSD}(p||q)$, the usual Jensen–Shannon divergence. We now demonstrate that Nielsen’s generalized Jensen–Shannon Divergence can be bounded by the total variation distance just as the ordinary Jensen–Shannon Divergence.

**Theorem 5.** For $p$ and $q$ densities with respect to a reference measure $\mu$, $w_i > 0$, such that $\sum_{i=1}^{n} w_i = 1$ and $\alpha_i \in (0,1)$,
\[
\log e \, \text{Var}_{w}(\alpha) \, |p - q|^2_{TV} \leq \text{JS}^{\alpha,w}(p : q) \leq H(\mu) \, |p - q|_{TV}
\]
where $H(\mu) := - \sum_i w_i \log w_i \geq 0$ and $\Delta = \max_i |\alpha_i - \bar{\alpha}_i|$ with $\bar{\alpha}_i = \sum_{j \neq i} \frac{w_j}{w_i} \alpha_j$.

Note that, since $\bar{\alpha}_i$ is the $w$ average of the $\alpha_j$ terms with $\alpha_i$ removed, $\bar{\alpha}_i \in [0,1]$ and thus $\Delta \leq 1$.

We need the following Theorem from Melbourne et al. [26] for the upper bound.

**Theorem 6** ([26] Theorem 1.1). For $f_i$ densities with respect to a common reference measure $\gamma$ and $\lambda_i > 0$ such that $\sum_{i=1}^{n} \lambda_i = 1$,
\[
h_\gamma(\sum_{i} \lambda_i f_i) - \sum_{i} \lambda_i h_\gamma(f_i) \leq T H(\lambda),
\]
where $h_\gamma(f_i) := - \int f_i(x) \log f_i(x)d\gamma(x)$ and $T = \sup_i |f_i - \hat{f}_i|_{TV}$ with $\hat{f}_i = \sum_{j \neq i} \frac{\lambda_j}{\lambda_i} f_j$.\]
Proof of Theorem 5. We apply Theorem 6 with \( f_i = (1 - \alpha_i)p + \alpha_iq \), \( \lambda_i = w_i \), and noticing that in general
\[
    h_T(\sum_i \lambda_i f_i) - \sum_i \lambda_i h_T(f_i) = \sum_i \lambda_i D(f_i || f),
\]
we have
\[
    JS^{\alpha,w}(p : q) = \sum_{i=1}^{n} w_i D((1 - \alpha_i)p + \alpha_iq || (1 - \tilde{\alpha})p + \tilde{\alpha}q)
\]
\[
    \leq TH(w).
\]
It remains to determine \( T = \max_i |f_i - \tilde{f}_i|_{TV} \),
\[
    \tilde{f}_i - f_i = \frac{f - f_i}{1 - \lambda_i}
\]
\[
    = \frac{((1 - \tilde{\alpha})p + \tilde{\alpha}q) - ((1 - \alpha_i)p + \alpha_iq)}{1 - w_i}
\]
\[
    = \frac{(\alpha_i - \tilde{\alpha})(p - q)}{1 - w_i}
\]
\[
    = (\alpha_i - \tilde{\alpha})(p - q).
\]
Thus, \( T = \max_i (\alpha_i - \tilde{\alpha})|p - q|_{TV} = A|p - q|_{TV} \), and the proof of the upper bound is complete.

To prove the lower bound, we apply Pinsker’s inequality, \( 2 \log e |P - Q|_{TV}^2 \leq D(P||Q) \),
\[
    JS^{\alpha,w}(p : q) = \sum_{i=1}^{n} w_i D((1 - \alpha_i)p + \alpha_iq || (1 - \tilde{\alpha})p + \tilde{\alpha}q)
\]
\[
    \geq \frac{1}{2} \sum_{i=1}^{n} w_i 2 \log e |((1 - \alpha_i)p + \alpha_iq) - ((1 - \tilde{\alpha})p + \tilde{\alpha}q)|_{TV}^2
\]
\[
    = \log e \sum_{i=1}^{n} w_i (\alpha_i - \tilde{\alpha})^2 |p - q|_{TV}^2
\]
\[
    = \log e \text{Var}_w(\alpha) |p - q|_{TV}^2.
\]
\( \square \)

Definition 7. Given an \( f \)-divergence, densities \( p \) and \( q \) with respect to common reference measure, \( \alpha \in [0, 1]^n \) and \( w \in (0, 1)^n \) such that \( \sum_i w_i = 1 \) define its generalized skew divergence
\[
    D_f^{\alpha,w}(p : q) = \sum_{i=1}^{n} w_i D_f((1 - \alpha_i)p + \alpha_iq || (1 - \tilde{\alpha})p + \tilde{\alpha}q).
\]
where \( \tilde{\alpha} = \sum_i w_i \alpha_i \).

Note that, by Theorem 4, \( D_f^{\alpha,w} \) is an \( f \)-divergence. The generalized skew divergence of the relative entropy is the generalized Jensen–Shannon divergence \( JS^{\alpha,w} \). We denote the generalized skew divergence of the \( \chi^2 \)-divergence from \( p \) to \( q \) by
\[
    \chi^2_{\alpha,w}(p : q) := \sum_i w_i \chi^2((1 - \alpha_i)p + \alpha_iq || (1 - \tilde{\alpha}p + \tilde{\alpha}q)
\]
\[
    \chi^2_{\alpha,w} (p : q) := \sum_{i=1}^{n} w_i \chi^2((1 - \alpha_i)p + \alpha_iq || (1 - \tilde{\alpha}p + \tilde{\alpha}q)
\]
\[
    \chi^2_{\alpha,w} (p : q) := \sum_{i=1}^{n} w_i \chi^2((1 - \alpha_i)p + \alpha_iq || (1 - \tilde{\alpha}p + \tilde{\alpha}q)).
\]
Note that, when \( n = 2 \) and \( \alpha_1 = 0, \alpha_2 = 1 \) and \( w_i = \frac{1}{2} \), we recover the skew symmetrized divergence in Definition 4

\[
D_{f, f(t_1, t_2/2)}(p : q) = \Delta_f(p||q)
\]

(40)

The following theorem shows that the usual upper bound for the relative entropy by the \( \chi^2 \)-divergence can be reversed up to a factor in the skewed case.

**Theorem 7.** For \( p \) and \( q \) with a common dominating measure \( \mu \),

\[
\chi^2_{\alpha, w}(p : q) \leq N_\infty(\alpha, w) JS^{\alpha, w}(p : q).
\]

Writing \( N_\infty(\alpha, w) = \max_{\alpha} \max \left\{ \frac{1 - \alpha_i}{1 - \alpha}, \frac{\alpha_i}{\alpha} \right\} \). For \( \alpha \in [0, 1]^n \) and \( w \in (0, 1)^n \) such that \( \sum w_i = 1 \), we use the notation \( N_\infty(\alpha, w) := \max_i e^{D_\infty(\alpha_i||\alpha)} \) where \( \alpha := \sum w_i \alpha_i \).

**Proof.** By definition,

\[
JS^{\alpha, w}(p : q) = \sum_{i=1}^n w_i D((1 - \alpha_i)p + \alpha_i q || (1 - \alpha) p + \alpha q).
\]

Taking \( P_i \) to be the measure associated to \( (1 - \alpha_i)p + \alpha_i q \) and \( Q \) given by \( (1 - \alpha)p + \alpha q \), then

\[
\frac{dP_i}{dQ} = \frac{(1 - \alpha_i)p + \alpha_i q}{(1 - \alpha)p + \alpha q} \leq \max \left\{ \frac{1 - \alpha_i}{1 - \alpha}, \frac{\alpha_i}{\alpha} \right\} = e^{D_\infty(\alpha_i||\alpha)} \leq N_\infty(\alpha, w).
\]

(41)

Since \( f(x) = x \log x \), the convex function associated to the usual KL divergence, satisfies \( f''(x) = \frac{1}{x} \), \( f \) is \( e^{-D_\infty(\alpha)} \)-convex on \( [0, \sup \rho_i, \frac{dP_i}{dQ}(x)] \), applying Proposition 2, we obtain

\[
D \left( \sum_i w_i P_i \mid\mid Q \right) \leq \sum_i w_i D(P_i \mid\mid Q) - \frac{\sum_i w_i \int_x \left( \frac{dP_i}{dQ} - \frac{dP}{dQ} \right)^2 dQ}{2N_\infty(\alpha, w)}.
\]

(42)

Since \( Q = \sum_i w_i P_i \), the left hand side of (42) is zero, while

\[
\sum_i w_i \int_x \left( \frac{dP_i}{dQ} - \frac{dP}{dQ} \right)^2 dQ = \sum_i w_i \int_x \left( \frac{dP_i}{dP} - 1 \right)^2 dP
\]

(43)

\[
= \sum_i w_i \chi^2(P_i \mid\mid P)
\]

(44)

\[
= \chi^2_{\alpha, w}(p : q).
\]

Rearranging gives,

\[
\frac{\chi^2_{\alpha, w}(p : q)}{2N_\infty(\alpha, w)} \leq JS^{\alpha, w}(p : q),
\]

which is our conclusion. \( \Box \)

**4. Total Variation Bounds and Bayes Risk**

In this section, we derive bounds on the Bayes risk associated to a family of probability measures with a prior distribution \( \lambda \). Let us state definitions and recall basic relationships. Given probability densities \( \{ p_i \}_{i=1}^n \) on a space \( \mathcal{X} \) with respect a reference measure \( \mu \) and \( \lambda_i \geq 0 \) such that \( \sum_{i=1}^n \lambda_i = 1 \), define the Bayes risk,

\[
R := R_\lambda(p) := 1 - \int_{\mathcal{X}} \max_i \{ \lambda_i p_i(x) \} d\mu(x)
\]

(45)
If $\ell(x, y) = 1 - \delta_z(y)$, and we define $T(x) := \arg\max_i \lambda_i p_i(x)$ then observe that this definition is consistent with, the usual definition of the Bayes risk associated to the loss function $\ell$. Below, we consider $\theta$ to be a random variable on $\{1, 2, \ldots, n\}$ such that $\mathbb{P}(\theta = i) = \lambda_i$, and $x$ to be a variable with conditional distribution $\mathbb{P}(X \in A | \theta = i) = \int_A p_i(x) d\mu(x)$. The following result shows that the Bayes risk gives the probability of the categorization error, under an optimal estimator.

**Proposition 5.** The Bayes risk satisfies

$$R = \min_{\hat{\theta}} \mathbb{E}\ell(\theta, \hat{\theta}(X)) = \mathbb{E}\ell(\theta, T(X))$$

where the minimum is defined over $\hat{\theta} : \mathcal{X} \to \{1, 2, \ldots, n\}$.

**Proof.** Observe that $R = 1 - \int_{\mathcal{X}} \lambda_{T(x)} p_{T(x)}(x) d\mu(x) = \mathbb{E}\ell(\theta, T(X))$. Similarly,

$$\mathbb{E}\ell(\theta, \hat{\theta}(X)) = 1 - \int_{\mathcal{X}} \lambda_{\hat{\theta}(x)} p_{\hat{\theta}(x)}(x) d\mu(x) \geq 1 - \int_{\mathcal{X}} \lambda_{T(x)} p_{T(x)}(x) d\mu(x) = R,$$

which gives our conclusion. \(\square\)

It is known (see, for example, [9,31]) that the Bayes risk can also be tied directly to the total variation in the following special case, whose proof we include for completeness.

**Proposition 6.** When $n = 2$ and $\lambda_1 = \lambda_2 = \frac{1}{2}$, the Bayes risk associated to the densities $p_1$ and $p_2$ satisfies

$$2R = 1 - |p_1 - p_2|_{TV}$$

**Proof.** Since $p_T = \frac{|p_1 - p_2|^2 + p_1 + p_2}{2}$, integrating gives $\int_{\mathcal{X}} p_T(x) d\mu(x) = |p_1 - p_2|_{TV} + 1$ from which the equality follows. \(\square\)

Information theoretic bounds to control the Bayes and minimax risk have an extensive literature (see, for example, [9,32–35]). Fano’s inequality is the seminal result in this direction, and we direct the reader to a survey of such techniques in statistical estimation (see [36]). What follows can be understood as a sharpening of the work of Guntuboyina [9] under the assumption of a $\kappa$-convexity.

The function $T(x) = \arg\max_i \{\lambda_i p_i(x)\}$ induces the following convex decompositions of our densities. The density $q$ can be realized as a convex combination of $q_1 = \frac{\lambda_2 Q}{1-Q}$ where $Q = 1 - \int \lambda_T q d\mu$ and $q_2 = \frac{(1-\lambda_1)Q}{Q}$,

$$q = (1 - Q)q_1 + Qq_2.$$

If we take $p := \sum_i \lambda_i p_i$, then $p$ can be decomposed as $p_1 = \frac{\lambda_T p_T}{1 - R}$ and $p_2 = \frac{p - \lambda_T p_T}{R}$ so that

$$p = (1 - R)p_1 + Rp_2.$$

**Theorem 8.** When $f$ is $\kappa$-convex, on $(a, b)$ with $a = \inf_{x, \lambda_i} \frac{p\lambda_i(x)}{\lambda_i(x)}$ and $b = \sup_{x, \lambda_i} \frac{p\lambda_i(x)}{\lambda_i(x)}$

$$\sum_i \lambda_i D_f(p_i || q) \geq D_f(R || Q) + \frac{\kappa W}{2}$$

where

$$W := W(\lambda_i, p_i, q) := \frac{(1 - R)^2}{1 - Q} \chi^2(p_1 || q_1) + \frac{R^2}{Q} \chi^2(p_2 || q_2) + W_0$$

for $W_0 \geq 0$. 

$W_0$ can be expressed explicitly as

$$W_0 = \int (1 - \lambda T) \text{Var}_{\lambda_i \neq T} \left( \frac{p_i}{q} \right) \, d\mu = \int \sum \lambda_i \left| p_i - \sum_{j \neq T} \frac{\lambda_j}{1 - \lambda T} p_j \right|^2 \, d\mu,$$

where for fixed $x$, we consider the variance $\text{Var}_{\lambda_i \neq T} \left( \frac{p_i}{q} \right)$ to be the variance of a random variable taking values $p_i(x)/q(x)$ with probability $\lambda_i / (1 - \lambda_T(x))$ for $i \neq T(x)$. Note this term is a non-zero term only when $n > 2$.

**Proof.** For a fixed $x$, we apply Lemma 1

$$\sum_i \lambda_i f \left( \frac{p_i}{q} \right) = \lambda_T f \left( \frac{p_T}{q} \right) + (1 - \lambda_T) \sum_i \frac{\lambda_i}{1 - \lambda_T} f \left( \frac{p_i}{q} \right) \geq \lambda_T f \left( \frac{p_T}{q} \right) + (1 - \lambda_T) \left[ f \left( \frac{p - \lambda_T p_T}{q(1 - \lambda_T)} \right) + \frac{\kappa}{2} \text{Var}_{\lambda_i \neq T} \left( \frac{p_i}{q} \right) \right]$$

Integrating,

$$\sum_i \lambda_i D_f(p_i||q) \geq \int \lambda_T f \left( \frac{p_T}{q} \right) q + \int (1 - \lambda_T) f \left( \frac{p - \lambda_T p_T + \sum \lambda_i p_i}{q(1 - \lambda_T)} \right) q + \frac{\kappa}{2} W_0,$$

where

$$W_0 = \int \sum_{i \neq T(x)} \lambda_i \left| p_i - \sum_{j \neq T} \frac{\lambda_j}{1 - \lambda_T(x)} p_j \right|^2 \, d\mu.$$  \hspace{1cm} (49)

Applying the $\kappa$-convexity of $f$,

$$\int \lambda_T f \left( \frac{p_T}{q} \right) q = (1 - Q) \int q_1 f \left( \frac{p_T}{q} \right)$$

$$\geq (1 - Q) \left( f \left( \frac{\lambda_T p_T}{1 - Q} \right) + \frac{\kappa}{2} \text{Var}_{q_1} \left( \frac{p_T}{q} \right) \right)$$

$$= (1 - Q) f((1 - R)/(1 - Q)) + \frac{Q \kappa}{2} W_1,$$

with

$$W_1 := \text{Var}_{q_1} \left( \frac{p_T}{q} \right)$$

$$= \left( \frac{1 - R}{1 - Q} \right)^2 \text{Var}_{q_1} \left( \frac{\lambda_T p_T}{1 - Q} \right)$$

$$= \left( \frac{1 - R}{1 - Q} \right)^2 \text{Var}_{q_1} \left( \frac{p_T}{q_1} \right)$$

$$= \left( \frac{1 - R}{1 - Q} \right)^2 \chi^2(p_1||q_1).$$  \hspace{1cm} (51)
Similarly,
\[
\int (1 - \lambda_T)f \left( \frac{p - \lambda_T p_T}{q(1 - \lambda_T)} \right) q = Q \int q_2 f \left( \frac{p - \lambda_T p_T}{q(1 - \lambda_T)} \right) \\
\geq Qf \left( \int q_2 \frac{p - \lambda_T p_T}{q(1 - \lambda_T)} \right) + \frac{Q\kappa}{2} W_2 \\
= Qf \left( \frac{R}{1 - Q} \right) + \frac{Q\kappa}{2} W_2 
\]
(52)

where
\[
W_2 := \text{Var}_{q_2} \left( \frac{p - \lambda_T p_T}{q(1 - \lambda_T)} \right) \\
= \left( \frac{R}{Q} \right)^2 \text{Var}_{q_2} \left( \frac{p - \lambda_T p_T}{q(1 - \lambda_T)} \right) \frac{Q}{R} \\
= \left( \frac{R}{Q} \right)^2 \text{Var}_{q_2} \left( \frac{p - \lambda_T p_T}{q(1 - \lambda_T)} \right) \frac{R}{Q} \\
= \left( \frac{R}{Q} \right)^2 \int q_2 \left( \frac{p_2}{q_2} - 1 \right)^2 \\
= \left( \frac{R}{Q} \right)^2 \chi^2(\rho_2 || q_2) 
\]
(53)

Writing \( W = W_0 + W_1 + W_2 \), we have our result. \( \square \)

**Corollary 4.** When \( \lambda_i = \frac{1}{n} \), and \( f \) is \( \kappa \)-convex on \((\text{inf}_{i,x} p_i/q, \text{sup}_{i,x} p_i/q)\)
\[
\frac{1}{n} \sum_i D_f(p_i || q) \\
\geq D_f(R|| (n - 1)/n) + \frac{\kappa}{2} \left( n^2 (1 - R)^2 \chi^2(\rho_1 || q) + \left( \frac{nR}{n - 1} \right)^2 \chi^2(\rho_2 || q) + W_0 \right) 
\]
(54)

further when \( n = 2 \),
\[
\frac{D_f(p_1 || q) + D_f(p_2 || q)}{2} \geq D_f \left( \frac{1 - |p_1 - p_2|_{TV}}{2} \right) \\
\geq D_f \left( \frac{1 - |p_1 - p_2|_{TV}}{2} \right) + \frac{\kappa}{2} \left( (1 + |p_1 - p_2|_{TV})^2 \chi^2(\rho_1 || q) + (1 - |p_1 - p_2|_{TV})^2 \chi^2(\rho_2 || q) \right). 
\]
(55)

**Proof.** Note that \( q_1 = q_2 = q \), since \( \lambda_i = \frac{1}{n} \) implies \( \lambda_T = \frac{1}{n} \) as well. In addition, \( Q = 1 - \int \lambda_T q d\mu = \frac{n - 1}{n} \) so that applying Theorem 8 gives
\[
\sum_{i=1}^n D_f(p_i || q) \geq n D_f(R|| (n - 1)/n) + \frac{\kappa n W(\lambda_i, p_i, q)}{2}. 
\]
(56)
The term $W$ can be simplified as well. In the notation of the proof of Theorem 8,

$$W_1 = n^2 (1 - R)^2 \chi^2(p_1, q)$$

$$W_2 = \left(\frac{nR}{n-1}\right)^2 \chi^2(p_2|q)$$

$$W_0 = \int \frac{1}{n-1} \sum_{i \neq T} \left(\frac{p_i - 1}{n-1} \sum_{j \neq T} p_j\right)^2 d\mu.$$  \hfill (57)

For the special case, one needs only to recall $R = \frac{1 - |p_1 - p_2|_{TV}}{2}$ while inserting 2 for $n$. \hfill □

**Corollary 5.** When $p_i \leq q/t^*$ for $t^* > 0$, and $f(x) = x \log x$

$$\sum_i \lambda_i D(p_i||q) \geq D(R||Q) + \frac{t^* W(\lambda_i, p_i, q)}{2}$$

for $D(p_i||q)$ the relative entropy. In particular,

$$\sum \lambda_i D(p_i||q) \geq D(p||q) + D(R||P) + \frac{t^* W(\lambda_i, p_i, p)}{2}$$

where $P = 1 - \int \lambda_T p d\mu$ for $p = \sum_i \lambda_i p_i$ and $t^* = \min \lambda_i$.

**Proof.** For the relative entropy, $f(x) = x \log x$ is $\frac{1}{x}$-convex on $[0, M]$ since $f''(x) = 1/x$. When $p_i \leq q/t^*$ holds for all $i$, then we can apply Theorem 8 with $M = \frac{1}{x}$. For the second inequality, recall the compensation identity, $\sum \lambda_i D(p_i||q) = \sum_i \lambda_i D(p_i||p) + D(p||q)$, and apply the first inequality to $\sum_i D(p_i||p)$ for the result. \hfill □

This gives an upper bound on the Jensen–Shannon divergence, defined as $\text{JSD}(\mu||v) = \frac{1}{2} D(\mu||\mu/2 + v/2) + \frac{1}{2} D(v||\mu/2 + v/2)$. Let us also note that through the compensation identity $\sum \lambda_i D(p_i||q) = \sum \lambda_i D(p_i||p) + D(p||q)$, $\sum \lambda_i D(p_i||q) \geq \sum \lambda_i D(p_i||p)$ where $p = \sum \lambda_i p_i$. In the case that $\lambda_i = \frac{1}{N}$

$$\sum \lambda_i D(p_i||q) \geq \sum_i \lambda_i D(p_i||p) \geq \text{Qf} \left(\frac{1 - R}{Q}\right) + (1 - Q) f \left(\frac{R}{1 - Q}\right) + \frac{t^* W}{2}$$  \hfill (58)

**Corollary 6.** For two densities $p_1$ and $p_2$, the Jensen–Shannon divergence satisfies the following,

$$\text{JSD}(p_1||p_2) \geq D \left(\frac{1 - |p_1 - p_2|_{TV}}{2}\right)^{1/2}$$

$$+ \frac{1}{4} \left((1 + |p_1 - p_2|_{TV})^2 \chi^2(p_1||p) + (1 - |p_1 - p_2|_{TV})^2 \chi^2(p_2||p)\right)$$

with $\rho(i)$ defined above and $p = p_1/2 + p_2/2$. \hfill (59)

**Proof.** Since $\frac{p_i}{|p_1 + p_2|/2} \leq 2$ and $f(x) = x \log x$ satisfies $f''(x) \geq \frac{1}{x}$ on $(0, 2)$. Taking $q = \frac{p_1 + p_2}{2}$, in the $n = 2$ example of Corollary 4 with $\kappa = \frac{1}{2}$ yields the result. \hfill □
Note that $2D((1 + V)/2||1/2) = (1 + V) \log(1 + V) + (1 - V) \log(1 - V) \geq V^2 \log e$, we see that a further bound,

$$\text{JSD}(p_1||p_2) \geq \frac{\log e}{2} V^2 + \frac{(1 + V)^2 \chi^2(p_1||p) + (1 - V)^2 \chi^2(p_2||p)}{4},$$

(60)
can be obtained for $V = |p_1 - p_2|_{TV}$.

**On Topsøe's Sharpening of Pinsker's Inequality**

For $P$, $Q$ probability measures with densities $p_i$ and $q$ with respect to a common reference measure, $\sum_{i=1}^n t_i = 1$, with $t_i > 0$, denote $P = \sum_i t_i P_i$ with density $p = \sum_i t_i p_i$, the compensation identity is

$$\sum_{i=1}^n t_i D(P_i||Q) = D(P||Q) + \sum_{i=1}^n t_i D(P_i||P).$$

(61)

**Theorem 9.** For $P_1$ and $P_2$, denote $M_k = 2^{-k} P_1 + (1 - 2^{-k}) P_2$, and define

$$M_1(k) = \frac{M_k \{ P_1 > P_2 \} + P_2 \{ P_1 < P_2 \}}{M_k \{ P_1 > P_2 \} + P_1 \{ P_1 < P_2 \}}$$

and

$$M_2(k) = \frac{M_k \{ P_1 < P_2 \} + P_2 \{ P_1 > P_2 \}}{M_k \{ P_1 < P_2 \} + P_1 \{ P_1 > P_2 \}},$$

then the following sharpening of Pinsker's inequality can be derived,

$$D(P_1||P_2) \geq (2 \log e) |P_1 - P_2|_{TV}^2 + \sum_{k=0}^{\infty} 2^k \left( \frac{\chi^2(M_1(k), M_{k+1})}{2} + \frac{\chi^2(M_2(k), M_{k+1})}{2} \right).$$

Proof. When $n = 2$ and $t_1 = t_2 = \frac{1}{2}$, if we denote $M = \frac{P_1 + P_2}{2}$, then (61) reads as

$$\frac{1}{2} D(P_1||Q) + \frac{1}{2} D(P_2||Q) = D(M||Q) + \text{JSD}(P_1||P_2).$$

(62)

Taking $Q = P_2$, we arrive at

$$D(P_1||P_2) = 2D(M||P_2) + 2\text{JSD}(P_1||P_2)$$

(63)

Iterating and writing $M_k = 2^{-k} P_1 + (1 - 2^{-k}) P_2$, we have

$$D(P_1||P_2) = 2^n \left( D(M_n||P_2) + 2 \sum_{k=0}^{n} \text{JSD}(M_n||P_2) \right)$$

(64)

It can be shown (see [11]) that $2^n D(M_n||P_2) \to 0$ with $n \to \infty$, giving the following series representation,

$$D(P_1||P_2) = 2 \sum_{k=0}^{\infty} 2^k \text{JSD}(M_k||P_2).$$

(65)

Note that the $\rho$-decomposition of $M_k$ is exactly $\rho_i = M_k(i)$, thus, by Corollary 6,

$$D(P_1||P_2) = 2 \sum_{k=0}^{\infty} 2^k \text{JSD}(M_k||P_2)$$

$$\geq \sum_{k=0}^{\infty} 2^k \left( |M_k - P_2|_{TV}^2 \log e + \frac{\chi^2(M_1(k), M_{k+1})}{2} + \frac{\chi^2(M_2(k), M_{k+1})}{2} \right)$$

(66)

$$= (2 \log e) |P_1 - P_2|_{TV}^2 + \sum_{k=0}^{\infty} 2^k \left( \frac{\chi^2(M_1(k), M_{k+1})}{2} + \frac{\chi^2(M_2(k), M_{k+1})}{2} \right).$$

Thus, we arrive at the desired sharpening of Pinsker’s inequality. □
Observe that the $k = 0$ term in the above series is equivalent to
\[ 2^0 \left( \frac{\chi^2(M_1(0), M_{0,1})}{2} + \frac{\chi^2(M_2(0), M_{0,1})}{2} \right) = \frac{\chi^2(p_1, p)}{2} + \frac{\chi^2(p_2, p)}{2}, \]
where $\rho_i$ is the convex decomposition of $p = \frac{p_1 + p_2}{2}$ in terms of $T(x) = \arg \max \{ p_1(x), p_2(x) \}$.

5. Conclusions

In this article, we begin a systematic study of strongly convex divergences, and how the strength of convexity of a divergence generator $f$, quantified by the parameter $\kappa$, influences the behavior of the divergence $D_f$. We prove that every strongly convex divergence dominates the square of the total variation, extending the classical bound provided by the $\chi^2$-divergence. We also study a general notion of a skew divergence, providing new bounds, in particular for the generalized skew divergence of Nielsen. Finally, we show how $\kappa$-convexity can be leveraged to yield improvements of Bayes risk $f$-divergence inequalities, and as a consequence achieve a sharpening of Pinsker’s inequality.

**Funding:** This research was funded by NSF grant CNS 1809194.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Appendix A**

**Theorem A1.** The class of $f$-divergences is stable under skewing. That is, if $f$ is convex, satisfying $f(1) = 0$, then
\[ \hat{f}(x) := (tx + (1-t))f \left( \frac{rx + (1-r)}{tx + (1-t)} \right) \]
is convex with $\hat{f}(1) = 0$ as well.

**Proof.** If $\mu$ and $\nu$ have respective densities $u$ and $v$ with respect to a reference measure $\gamma$, then $r\mu + (1-r)\nu$ and $t\mu + 1-t\nu$ have densities $ru + (1-r)v$ and $tu + (1-t)v$
\[
S_{f,x,t}(\mu||\nu) = \int f \left( \frac{ru + (1-r)v}{tu + (1-t)v} \right) (tu + (1-t)v) d\gamma \tag{A2}
= \int f \left( \frac{ru + (1-r)}{tu + (1-t)} \right) (tu + (1-t)v) d\gamma \tag{A3}
= \int \hat{f} \left( \frac{u}{v} \right) vd\gamma. \tag{A4}
\]
Since $\hat{f}(1) = f(1) = 0$, we need only prove $f$ convex. For this, recall that the conic transform $g$ of a convex function $f$ defined by $g(x, y) = yf(x/y)$ for $y > 0$ is convex, since
\[ \frac{y_1 + y_2}{2} f \left( \frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \leq \frac{y_1}{2} f(x_1/y_1) + \frac{y_2}{2} f(x_2/y_2). \tag{A5}\]
Our result follows since $\hat{f}$ is the composition of the affine function $A(x) = (rx + (1-r), tx + (1-t))$ with the conic transform of $f$,
\[ \hat{f}(x) = g(A(x)). \tag{A7}\]
References

1. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B* 1966, 28, 131–142. [CrossRef]
2. Morimoto, T. Markov processes and the H-theorem. *J. Phys. Soc. Jpn.* 1963, 18, 328–331. [CrossRef]
3. Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Magy. Tud. Akad. Mat. Kutató Int. Közl.* 1963, 8, 85–108.
4. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* 1967, 2, 229–318.
5. Polyanskiy, Y.; Poor, H.V.; Verdú, S. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory* 2010, 56, 2307–2359. [CrossRef]
6. Sason, I.; Verdú, S. $f$-divergence inequalities. *IEEE Trans. Inf. Theory* 2016, 62, 5973–6006. [CrossRef]
7. Polyanskiy, Y.; Wu, Y. Lecture Notes on Information Theory. Available online: http://people.lids.mit.edu/yp/homepage/data/itlectures_v5.pdf (accessed on 13 November 2019).
8. Sason, I. On data-processing and majorization inequalities for $f$-divergences with applications. *Entropy* 2019, 21, 1022. [CrossRef]
9. Guntuboyina, A. Lower bounds for the minimax risk using $f$-divergences, and applications. *IEEE Trans. Inf. Theory* 2011, 57, 2386–2399. [CrossRef]
10. Reid, M.; Williamson, R. Generalised Pinsker inequalities. *arXiv* 2009, arXiv:0906.1244.
11. Topsøe, F. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory* 2000, 46, 1602–1609. [CrossRef]
12. Lee, L. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association For Computational Linguistics on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1999; pp. 25–32.
13. Le Cam, L. *Asymptotic Methods in Statistical Decision Theory*; Springer Series in Statistics; Springer: New York, NY, USA, 1986.
14. Vincze, I. On the concept and measure of information contained in an observation. *Contrib. Probab.* 1981, 207–214. [CrossRef]
15. Györfi, L.; Vajda, I. A class of modified Pearson and Neyman statistics. *Stat. Decis.* 2001, 19, 239–251. [CrossRef]
16. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* 1991, 37, 145–151. [CrossRef]
17. Nielsen, F. On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid. *Entropy* 2020, 22, 221. [CrossRef]
18. Folland, G. *Real Analysis: Modern Techniques and Their Applications*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
19. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* 2002, 70, 419–435. [CrossRef]
20. Harremoës, P.; Vajda, I. On pairs of $f$-divergences and their joint range. *IEEE Trans. Inf. Theory* 2011, 57, 3230–3235. [CrossRef]
21. Reiss, R. *Approximate Distributions of Order Statistics: With Applications to Nonparametric Statistics*; Springer: Berlin/Heidelberg, Germany, 2012.
22. Nielsen, F.; Nock, R. On the chi square and higher-order chi distances for approximating $f$-divergences. *IEEE Signal Process. Lett.* 2013, 21, 10–13. [CrossRef]
23. Amari, S. *Information Geometry and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2016; p. 194.
24. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; CRC Press: Boca Raton, FL, USA, 2011.
25. Vajda, I. On the $f$-divergence and singularity of probability measures. *Period. Math. Hung.* 1972, 2, 223–234. [CrossRef]
26. Melbourne, J.; Talukdar, S.; Bhaban, S.; Madiman, M.; Salapaka, M.V. The differential entropy of mixtures: new bounds and applications. *arXiv* 2020, arXiv:1805.11257.
27. Erven, T.V.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* 2014, 60, 3797–3820. [CrossRef]
28. Audenaert, K.M.R. Quantum skew divergence. *J. Math. Phys.* 2014, 55, 112202. [CrossRef]
29. Melbourne, J.; Madiman, M.; Salapaka, M.V. Relationships between certain f-divergences. In Proceedings of the 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 24–27 September 2019; pp. 1068–1073.
30. Nishiyama, T.; Sason, I. On relations between the relative entropy and χ2-divergence, generalizations and applications. *Entropy* 2020, 22, 563. [CrossRef]
31. Nielsen, F. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognit. Lett.* 2014, 42, 25–34. [CrossRef]
32. Birgé, L. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inf. Theory* 2005, 51, 1611–1615. [CrossRef]
33. Chen, X.; Guntuboyina, A.; Zhang, Y. On Bayes risk lower bounds. *J. Mach. Learn. Res.* 2016, 17, 7687–7744.
34. Xu, A.; Raginsky, M. Information-theoretic lower bounds on Bayes risk in decentralized estimation. *IEEE Trans. Inf. Theory* 2016, 63, 1580–1600. [CrossRef]
35. Yang, Y.; Barron, A. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* 1999, 27, 1564–1599.
36. Scarlett, J.; Cevher, V. An introductory guide to Fano’s inequality with applications in statistical estimation. *arXiv* 2019, arXiv:1901.00555.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).