Optimal Bounds between f-Divergences and Integral Probability Metrics

Rohit Agrawal [*1] **Thibaut Horel** [*2]

Abstract

The families of f-divergences (e.g. the Kullback–Leibler divergence) and Integral Probability Metrics (e.g. total variation distance or maximum mean discrepancies) are commonly used in optimization and estimation. In this work, we systematically study the relationship between these two families from the perspective of convex duality. Starting from a tight variational representation of the f-divergence, we derive a generalization of the moment generating function, which we show exactly characterizes the best lower bound of the f-divergence as a function of a given IPM. Using this characterization, we obtain new bounds on IPMs defined by classes of unbounded functions, while also recovering in a unified manner well-known results for bounded and subgaussian functions (e.g. Pinsker's inequality and Hoeffding's lemma).

1. Introduction

Quantifying the extent to which two probability distributions differ from one another is central in most, if not all, problems and methods in machine learning and statistics. For example, maximum likelihood estimation is equivalent to minimizing the Kullback–Leibler divergence between the empirical distribution—or the ground truth distribution in the limit of infinitely large sample—and a distribution chosen from a parametric family.

A natural generalization of the Kullback–Leibler divergence is provided by the family of ϕ-divergences[1] (Csiszár, 1963; 1967) also known in statistics as Ali–Silvey distances (Ali & Silvey, 1966). Informally, a ϕ-divergence quantifies the divergence between two distributions μ and ν as an average cost of the likelihood ratio, that is, $I_\phi(\mu \,\|\, \nu) := \int \phi(d\mu/d\nu)\, d\nu$ for a convex cost function $\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Notable examples of ϕ-divergences include the Hellinger distance, the α-divergences (a convex transformation of the Rényi divergences), and the χ^2-divergence.

Crucial in applications of ϕ-divergences are their so-called *variational representations*. For example, the Donsker–Varadhan representation (Donsker & Varadhan, 1976, Theorem 5.2) expresses the Kullback–Leibler divergence $\mathrm{D}(\mu \,\|\, \nu)$ between probability distributions μ and ν as

$$\mathrm{D}(\mu \,\|\, \nu) = \sup_{g \in \mathcal{L}_\nu^\infty} \left\{ \int g\, d\mu - \log \int e^g\, d\nu \right\}, \qquad (1)$$

where $\mathcal{L}_\nu^\infty$ is the space of functions essentially bounded with respect to ν. Similar variational representations were for example used by Nguyen et al. (2008; 2010); Ruderman et al. (2012); Belghazi et al. (2018) to construct estimates of ϕ-divergences by restricting the optimization problem in (1) to a class of functions $\mathcal{G} \subseteq \mathcal{L}_\nu^\infty$ for which the problem becomes tractable (for example when $\mathcal{G}$ is a RKHS or representable by a given neural network architecture). In recent work, Nowozin et al. (2016); Nock et al. (2017) conceptualized an extension of generative adversarial networks (GANs) in which the problem of minimizing a ϕ-divergence is expressed via variational representations such as (1) as a minimax game involving two neural networks, one minimizing over probability distributions μ, the other maximizing over g as in (1).

Another important class of distances between probability distributions is given by Integral Probability Metrics (IPMs) defined by Müller (1997) and taking the form

$$d_{\mathcal{G}}(\mu, \nu) = \sup_{g \in \mathcal{G}} \left\{ \left| \int g\, d\mu - \int g\, d\nu \right| \right\}, \qquad (2)$$

where $\mathcal{G}$ is a class of functions parametrizing the distance. Notable examples include the total variation distance ($\mathcal{G}$ is the class of all functions taking value in $[-1, 1]$), the Wasserstein metric ($\mathcal{G}$ is a class of Lipschitz functions) and Maximum Mean Discrepancies ($\mathcal{G}$ is the unit ball of a RKHS). Being already expressed as a variational problem, IPMs are amenable to estimation, as was exploited by Sriperumbudur et al. (2012); Gretton et al. (2012). MMDs have also been

[*]Equal contribution [1]Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, Massachusetts, USA [2]Institute for Data, Systems, and Society, MIT, Cambridge, Massachusetts, USA. Correspondence to: Rohit Agrawal <rohitagr@seas.harvard.edu>, Thibaut Horel <thibauth@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]In the rest of this paper, we use ϕ-divergence instead of f-divergence and reserve the letter f for a generic function.

used in lieu of ϕ-divergences to train GANs as was first done by Dziugaite et al. (2015).

Rewriting the optimization problem (1) as

$$\sup_{g \in \mathcal{L}_\nu^\infty} \left\{ \int g \, d\mu - \int g \, d\nu - \log \int e^{(g - \int g \, d\nu)} \, d\nu \right\} \tag{3}$$

suggests an important connection between ϕ-divergences and IPMs. Indeed, (3) expresses the divergence as the solution to a regularized optimization problem in which one attempts to maximize the mean deviation $\int g \, d\mu - \int g \, d\nu$, as in (2), while also penalizing functions g which are too "complex" as measured by the centered log moment generating function of g. In this work, we further explore the connection between ϕ-divergences and IPMs, guided by the following question:

what is the best lower bound of a given ϕ-divergence as a function of a given integral probability metric?

Some specific instances of this question are already well understood. For example, the best lower bound of the Kullback–Leibler divergence by a quadratic function of the total variation distance is known as Pinsker's inequality. More generally, describing the best lower bound of a ϕ-divergence as a function of the total variation distance (without being restricted to being a quadratic), is known as Vajda's problem, to which an answer was given by Fedotov et al. (2003) for the Kullback–Leibler divergence and by Gilardoni (2006) for an arbitrary ϕ-divergence.

In the case of the total variation distance, this question can also be seen as a specific instance of the problem of determining the *joint range* of values taken by an arbitrary pair of ϕ-divergences (Harremoës & Vajda, 2011; Guntuboyina et al., 2014). In contrast, in this work, we generalize the question in different manner by replacing the the total variation distance by an arbitrary IPM, as opposed to an arbitrary ϕ-divergence. This is incomparable since the total variation distance is the only ϕ-divergence which is also an IPM (Sriperumbudur et al., 2009; 2012).

Beyond the total variation distance—in particular, when the class $\mathcal{G}$ in (2) contains unbounded functions—few results are known. Using (3), Boucheron et al. (2013, Section 4.9) show that Pinsker's inequality holds as long as the log moment generating function grows at most quadratically. Since this is the case for bounded functions (via Hoeffding's lemma), this recovers Pinsker's inequality and extends it to the class of so-called *sub-Gaussian* functions. This was recently used by Russo & Zou (2020) to control bias in adaptive data analysis.

In this work, we systematize the convex analytic perspective underlying many of these results, thereby developing the necessary tools to resolve the above guiding question. As an application, we recover in a unified manner the known bounds between ϕ-divergences and IPMs, and extend them along several dimensions. Specifically, starting from the observation of Ruderman et al. (2012) that the variational representation of ϕ-divergences commonly used in the literature is not "tight" for probability measures (in a sense which will be made formal in the paper), we make the following contributions:

- we derive a tight representation of ϕ-divergences for probability measures, exactly generalizing the Donsker–Varadhan representation of the Kullback–Leibler divergence.
- we define a generalization of the log moment generating function and show that it exactly characterizes the best lower bound of a ϕ-divergence by an IPM.
- after proving a generalization of Hoeffding's lemma, we obtain as an application a generalization of Pinsker's inequality to a large class of ϕ-divergences.
- the answer to Vajda's problem is re-derived in a principled manner, providing a new geometric interpretation on the optimal lower bound of the ϕ-divergence by the total variation distance.
- finally, we introduce a generalization of sub-Gaussian functions to arbitrary ϕ-divergences and show that it exactly characterizes the class of function for which Pinsker's type inequalities can be obtained.

The rest of this paper is organized as follows: Section 2 gives a brief overview of concepts and tools used in this paper, Section 3 presents our general theorem characterizing the best lower bound, Section 4 focuses, as an application, on the total variation distance (bounded functions), and Section 5 explores applications to unbounded functions. In the full version of this work (Agrawal & Horel, 2020), we provide additional background, discuss more related work, and include some extra results and proofs omitted from this version due to space constraints.

2. Preliminaries

Notations. Unless otherwise noted, all the probability measures in this paper are defined on a common measurable space $(\Omega, \mathcal{A})$. We denote by $\mathcal{M}(\Omega, \mathcal{A})$, $\mathcal{M}_f(\Omega, \mathcal{A})$, and $\mathcal{M}_1(\Omega, \mathcal{A})$ the spaces of all (non-negative) measures, finite measures, and probability measures respectively. For $\nu \in \mathcal{M}(\Omega, \mathcal{A})$, and $1 \leq p \leq \infty$, $\mathcal{L}_\nu^p(\Omega, \mathcal{A})$ denotes the space of measurable functions with finite p-norm. $\mathcal{L}^0(\Omega)$ denotes the space of all measurable functions from Ω to $\mathbb{R}$. When there is no ambiguity, we drop the indication $(\Omega, \mathcal{A})$.

For two measures μ and ν, $\mu \ll \nu$ denotes that μ is absolutely continuous with respect to ν and in this case we denote by $\frac{d\mu}{d\nu} : \Omega \to [0, \infty)$ the Radon–Nikodym derivative of μ with respect to ν. For a measurable function $f : \Omega \to \mathbb{R}$, $\mu(f) \coloneqq \int f \, d\mu$ denotes the integral of f with respect to μ.

Convex analysis. We consider a pair (X, Y) of topological vector spaces and a bilinear form $\langle \cdot, \cdot \rangle \to \mathbb{R}$ such that $(X, Y, \langle \cdot, \cdot \rangle)$ form a dual pair (see e.g. Berg et al., 1984). For a convex function $f : \mathbb{R} \to \overline{\mathbb{R}}$, $\operatorname{dom} f \coloneqq \{x \in \mathbb{R} : f(x) < \infty\}$ is the effective domain of f and $\partial f(\cdot)$ denotes its subdifferential. For a set C, δ_C denotes the characteristic function of C ($\delta_C(x)$ is 0 if $x \in C$ and $+\infty$ elsewhere).

Definition 2.1 (Convex conjugate)**.** The *convex conjugate* (also called Fenchel dual or Fenchel–Legendre transform) of $f : X \to \overline{\mathbb{R}}$ is the function $f^\star : Y \to \overline{\mathbb{R}}$ defined by

$$f^\star(y) \coloneqq \sup_{x \in X} \big\{ \langle x, y \rangle - f(x) \big\}, \quad y \in Y \, .$$

ϕ-divergences. In this paper, $\phi : \mathbb{R} \to \overline{\mathbb{R}}$ is a convex lower semicontinuous (lsc) function, finite on an open interval around 1, strictly convex at 1 and such that $\phi(1) = 0$.

Definition 2.2 (ϕ-divergence)**.** Let μ and ν be two finite measures. Decomposing $\mu = \mu_c + \mu_s$ with $\mu_c \ll \nu$ and $\mu_s \perp \nu$, the *ϕ-divergence* I_ϕ of μ with respect to ν, is defined by

$$I_\phi(\mu \,\|\, \nu) \coloneqq \int \phi\left(\frac{d\mu_c}{d\nu}\right) d\nu + \mu_s(\Omega) \cdot \lim_{x \to \infty} \frac{\phi(x)}{x} \, ,$$

with the convention $0 \cdot \infty = 0$.

Remark. If $\mu \ll \nu$, the definition simplifies to $I_\phi(\mu \,\|\, \nu) = \nu\big(\phi \circ \frac{d\mu}{d\nu}\big)$. Furthermore, if $\lim_{x\to\infty} \phi(x)/x = +\infty$, then $I_\phi(\mu \,\|\, \nu) = +\infty$ whenever $\mu \not\ll \nu$.

If $\mu \in \mathcal{M}_1$, the value of $I_\phi(\mu \,\|\, \nu)$ is invariant when replacing ϕ with $\tilde{\phi} : x \mapsto \phi(x) + c \cdot (x - 1)$ for $c \in \mathbb{R}$. Hence, we "normalize" ϕ so that $0 \in \partial\phi(1)$. As a consequence, ϕ is a non-negative function reaching its minimum at 1. Furthermore, since we are interested only in non-negative measures, we redefine $\phi(x) = +\infty$ for $x < 0$, thereby hard-coding the non-negativity constraint in the definition of ϕ itself.

Integral Probability Metrics.

Definition 2.3. For $\mathcal{G} \subseteq \mathcal{L}^0$, the *integral probability metric* associated with $\mathcal{G}$ is given by

$$d_{\mathcal{G}}(\mu, \nu) \coloneqq \sup_{g \in \mathcal{G}} \left\{ \left| \int g \, d\mu - \int g \, d\nu \right| \right\} , \; (\mu, \nu) \in \mathcal{M}^2 \, .$$

Remark. When the class $\mathcal{G}$ is closed under negation, one can drop the absolute value in the definition.

The total variation distance $\mathrm{TV}(\mu, \nu)$ is obtained when $\mathcal{G}$ is the class of measurable functions taking value in $[-1, 1]$.

3. Dual representations and optimal bounds

In this section, we apply the convex duality framework to the analysis of ϕ-divergences. Starting from the usual variational representation for I_ϕ, we give an explicit and tighter representation for its restriction to the set of probability measures. This new representation lets us identify a natural generalization of the log moment generating function (a.k.a. the cumulant generating function) which we then show characterizes the best lower bound of the ϕ-divergence as a function of an IPM.

We fix a probability measure $\nu \in \mathcal{M}_1$ and consider the space of finite (signed) measures μ with $\mu \ll \nu$ which is identified with $\mathcal{L}^1_\nu$ by the Radon–Nikodym theorem[2]. The topological dual—for the norm topology—of $\mathcal{L}^1_\nu$ is $\big(\mathcal{L}^1_\nu\big)^* = \mathcal{L}^\infty_\nu$, the space of functions which are essentially bounded with respect to ν. More generally, we consider a pair of vector spaces $\mathcal{F}_\nu \subseteq \mathcal{L}^1_\nu$ and $\mathcal{G}_\nu \supseteq \mathcal{L}^\infty_\nu$ decomposable in the sense of (Rockafellar, 1976, §3) and such that $\big(\mathcal{F}_\nu, \mathcal{G}_\nu, \langle \cdot, \cdot \rangle\big)$ form a dual pair for the pairing $\langle \mu, g \rangle = \mu(g)$, $(\mu, g) \in \mathcal{F}_\nu \times \mathcal{G}_\nu$. We endow $\mathcal{F}_\nu$ and $\mathcal{G}_\nu$ with topologies compatible with the pair in the sense that $(\mathcal{F}_\nu)^* = \mathcal{G}_\nu$ and $(\mathcal{G}_\nu)^* = \mathcal{F}_\nu$. This formalism will be useful when considering unbounded functions in Section 5, but one can mentally substitute $\mathcal{F}_\nu$ (resp. $\mathcal{G}_\nu$) with $\mathcal{L}^1_\nu$ (resp. $\mathcal{L}^\infty_\nu$) at first reading.

3.1. Variational representation

Consider the functional $I_{\phi,\nu} : \mu \mapsto I_\phi(\mu \,\|\, \nu)$ over[3] $\mathcal{F}_\nu$, with convex conjugate

$$I^\star_{\phi,\nu}(g) \coloneqq \sup_{\mu \in \mathcal{F}_\nu} \big\{ \mu(g) - I_\phi(\mu \,\|\, \nu) \big\} \, , \quad g \in \mathcal{G}_\nu \, . \tag{4}$$

The following proposition, first stated in Rockafellar (1968), shows that $I^\star_{\phi,\nu}$ can itself be written as the convex integral functional $I_{\phi^\star,\nu}$ associated with $\phi^\star$. In other words, the integration and conjugacy operations commute. This implies that $I_{\phi,\nu}$ on $\mathcal{F}_\nu$ and $I_{\phi^\star,\nu}$ on $\mathcal{G}_\nu$ are conjugate to each other, "lifting" the fact that $(\phi, \phi^\star)$ form a conjugate pair to the associated integral functionals. As an immediate consequence, we obtain the usual variational representation of the ϕ-divergence.

Proposition 3.1. *The functionals $I_{\phi,\nu}$ and $I_{\phi^\star,\nu}$*[4] *are conju-*

[2] We show in the full version of this work (Agrawal & Horel, 2020) how to remove the requirement $\mu \ll \nu$.

[3] Recall that by definition, $\phi(x) = +\infty$ for $x < 0$, so $\operatorname{dom} I_{\phi,\nu}$ is contained in the positive cone $\mathcal{F}_\nu \cap \mathcal{M}_f$ of *non-negative* measures in $\mathcal{F}_\nu$. Defining $I_{\phi,\nu}$ as an extended real-valued function on the entire vector space $\mathcal{F}_\nu$ makes it more amenable to a convex duality treatment.

[4] Although the notation $I_{\phi^\star,\nu}$ denotes the partial function $\mu \mapsto I_{\phi^\star}(\mu \,\|\, \nu)$, we treat it here as a functional defined on the space of *functions* $\mathcal{G}_\nu$ by identifying $\mathcal{G}_\nu$ with a subspace of measures

gate to each other, that is,

$$I^\star_{\phi,\nu}(g) = I_{\phi^*,\nu}(g) \coloneqq \nu(\phi^\star \circ g)\,, \quad g \in \mathcal{G}_\nu, \tag{5}$$

$$I_{\phi,\nu}(\mu) = \sup_{g\in\mathcal{G}_\nu} \left\{\mu(g) - \nu(\phi^\star \circ g)\right\}, \quad \mu \in \mathcal{F}_\nu\,. \tag{6}$$

Proof. The fact that $I_{\phi,\nu}$ and $I_{\phi^\star,\nu}$ are conjugate to each other is an application of (Rockafellar, 1968, Corollary to Theorem 2) since $\mathcal{F}_\nu$ and $\mathcal{G}_\nu$ are decomposable by assumption. (5) explicitates the identity $I^\star_{\phi,\nu}(g) = I_{\phi^\star,\nu}(g)$, and (6) uses that $I_{\phi,\nu} = I_{\phi^{\star\star},\nu} = I^\star_{\phi^\star,\nu} = I^{\star\star}_{\phi,\nu}$. □

Example 3.2. Consider the case of the Kullback–Leibler divergence, corresponding to the function $\phi : x \mapsto x\log x - x + 1$. A simple computation gives $\phi^\star(x) = e^x - 1$ and (6) yields as a variational representation,

$$\mathrm{D}(\mu\|\nu) = \sup_{g\in\mathcal{G}_\nu} \left\{1 + \mu(g) - \int e^g\, d\nu\right\}, \tag{7}$$

for all measures $\mu \in \mathcal{F}_\nu$. Note that this representation differs from the Donsker–Varadhan representation (1). This discrepancy will be explained in the next section.

3.2. Restriction to probability measures

The variational representation given by Proposition 3.1—which holds for an arbitrary *finite* measure in $\mathcal{F}_\nu$—is loose when applied to *probability measures* as was first observed in Ruderman et al. (2012). In this section, we derive a tighter representation by "specializing" the derivation to *probability measures*.

Specifically, denote by $\tilde{I}_{\phi,\nu}$ the restriction of $I_{\phi,\nu}$ to the convex set of measures $\mu \in \mathcal{F}_\nu$ such that $\mu(\Omega) = 1$. Since the effective domain of $I_{\phi,\nu}$ is contained in $\mathcal{M}_f$ (cf. footnote 3), this is equivalent to restricting $I_{\phi,\nu}$ to the positive cone $\mathcal{M}_1 \cap \mathcal{F}_\nu$ of *probability* measures in $\mathcal{F}_\nu$, that is, $\tilde{I}_{\phi,\nu}(\mu) \coloneqq I_{\phi,\nu}(\mu) + \delta_{\mathcal{M}_1}(\mu)$, $\mu \in \mathcal{F}_\nu$. Consider now the convex conjugate $\tilde{I}^\star_{\phi,\nu}$ of $\tilde{I}_{\phi,\nu}$,

$$\tilde{I}^\star_{\phi,\nu}(g) \coloneqq \sup_{\substack{\mu\in\mathcal{F}_\nu \\ \mu(\Omega)=1}} \left\{\mu(g) - I_\phi\left(\mu \,\|\, \nu\right)\right\}, \quad g \in \mathcal{G}_\nu\,. \tag{8}$$

Compared to (4), the supremum is taken over a smaller set of measures, and hence $\tilde{I}^\star_{\phi,\nu}(g) \le I^\star_{\phi,\nu}(g)$, $g \in \mathcal{F}_\nu$. The following proposition gives a simpler expression for $\tilde{I}^\star_{\phi,\nu}$ showing that it is the solution of a *single dimensional* convex optimization problem, efficiently solvable in practice.

Proposition 3.3. *The functional $\tilde{I}^\star_{\phi,\nu}$ can be written*

$$\tilde{I}^\star_{\phi,\nu}(g) = \inf_{\lambda\in\mathbb{R}} \left\{\int \phi^\star(g+\lambda)\, d\nu - \lambda\right\}, \quad g \in \mathcal{G}_\nu\,.$$

Furthermore, if $\tilde{I}^\star_{\phi,\nu}(g) < \infty$, in particular whenever $g \in \mathcal{L}^\infty_\nu$, the infimum is reached.

Proof. Fix $g \in \mathcal{G}_\nu$. From (8), we obtain $-\tilde{I}^\star_{\phi,\nu}(g) = \inf_{\mu\in\mathcal{F}_\nu}\{I_{\phi,\nu}(\mu) - \mu(g) + \delta_{\{1\}}(\mu(1))\}$. We apply Fenchel's duality theorem (see e.g. Zălinescu (2002, Theorem 2.8.3)) with $X = \mathcal{F}_\nu$, $Y = \mathbb{R}$, $f : \mu \mapsto I_{\phi,\nu}(\mu) - \mu(g)$, $h = \delta_{\{1\}}$ and $A : \mu \mapsto \mu(1)$. Observe that f is convex with $f^\star(x^\star) = I^\star_{\phi,\nu}(x^\star + g)$, $x^\star \in \mathcal{G}_\nu$; h is convex with $h^\star(y^\star) = -y^\star$, $y^\star \in \mathbb{R}$, and A is linear and its adjoint operator $A^\star : \mathbb{R} \to \mathcal{G}_\nu$ is given by $A^\star(\lambda) = \lambda$, $\lambda \in \mathbb{R}$, where we also use λ to denote the constant function equal to λ everywhere. Hence the dual problem is $-\inf_{\lambda\in\mathbb{R}}\{I^\star_{\phi,\nu}(g+\lambda) - \lambda\}$. We now verify that constraint qualification holds. Observe that $\operatorname{dom} h = \{1\}$, so it is sufficient to show that $1 \in \operatorname{int}(A \operatorname{dom} f)$. By assumption, there exists $\varepsilon > 0$ such that $(1-\varepsilon, 1+\varepsilon) \subseteq \operatorname{dom}\phi$. This implies that the set of measures $\mathcal{C} = \{\alpha\cdot\nu : |\alpha - 1| < \varepsilon\}$ is contained in $\operatorname{dom} I_{\phi,\nu}$. Since $A\mathcal{C} = (1-\varepsilon, 1+\varepsilon)$, this implies that $1 \in \operatorname{int}(A \operatorname{dom} f)$ and strong duality holds. Using the expression of $I^\star_{\phi,\nu}$ found in Proposition 3.1 gives the desired expression for $\tilde{I}^\star_{\phi,\nu}$ and that the infimum is reached whenever it is not finite as claimed.

It remains to show that when $g \in \mathcal{L}^\infty_\nu$ that the infimum has a finite upper bound. Since $\phi(x) = \infty$ for all $x < 0$ and $\phi(x) \ge 0$ for all $x \ge 0$, we have for $t < 0$ that $\phi^\star(t) = \sup_x tx - \phi(x) = \sup_{x\ge 0} tx - \phi(x) \le \sup_{x\ge 0} tx = 0$. In particular, we get that by choosing $\lambda = -M$ for M the essential supremum of g that $\inf_{\lambda\in\mathbb{R}}\left\{\int \phi^\star(g+\lambda)\, d\nu - \lambda\right\} \le \int \phi^\star(g-M)\, d\nu + M \le \int 0\, d\nu + M = M < \infty$ since $g - M \le 0$ ν-a.s. by definition of essential supremum. □

Example 3.4. The squared Hellinger distance is the ϕ-divergence given by $\phi(x) = (\sqrt{x}-1)^2$, which has $\phi^\star(x) = \frac{2x}{2-x}$ for $x < 2$ and $\phi^\star(x) = \infty$ for $x \ge 2$. In particular, any random variable g such that $\nu(\{g > 2\}) > 0$ has $I^\star_{\phi,\nu}(g) = \nu(\phi^\star(g)) = \infty$, but because of the additive shift in Proposition 3.3, any bounded g has $\tilde{I}^\star_{\phi,\nu}(g) < \infty$.

As a corollary, we obtain a different variational representation of the ϕ-divergence, valid for probability measures and containing as a special case the Donsker–Varadhan representation of the Kullback–Leibler divergence.

Corollary 3.5. *For all probability measure $\mu \in \mathcal{F}_\nu$,*

$$I_\phi\left(\mu \,\|\, \nu\right) = \sup_{g\in\mathcal{G}_\nu} \left\{\mu(g) - \left(\inf_{\lambda\in\mathbb{R}} \int \phi^\star(g+\lambda)\, d\nu - \lambda\right)\right\}.$$

Proof. Define $C \coloneqq \left\{\mu \in \mathcal{F}_\nu : \mu(\Omega) = 1\right\}$ and observe that C is the preimage of $\{1\}$ under the map $\mu \mapsto \int 1\, d\mu$. Since $1 \in \mathcal{L}^\infty_\nu \subseteq \mathcal{G}_\nu$, this map is continuous for any topology compatible with the dual pair $(\mathcal{F}_\nu, \mathcal{G}_\nu)$, hence

absolutely continuous with respect to ν by the Radon–Nikodym theorem.

C is closed. This implies that δ_C is lsc., hence so is $\tilde{I}_{\phi,\nu} = I_{\phi,\nu} + \delta_C$ as the sum of two lsc functions. By the Fenchel–Moreau theorem, we can thus write $\tilde{I}_{\phi,\nu}$ as its biconjugate, which immediately gives the the statement of the corollary since $I_\phi(\mu \,\|\, \nu) = \tilde{I}_{\phi,\nu}(\mu)$ on $\mathcal{M}_1 \cap \mathcal{F}_\nu$. □

Example 3.6. As in Example 3.2, we consider the case of the Kullback–Leibler divergence, given by $\phi(x) = x \log x - x + 1$. Since $\phi^\star(x) = e^x - 1$ we get

$$\tilde{I}^\star_{\phi,\nu}(g) = \inf_{\lambda\in\mathbb{R}} \int e^{g+\lambda} - 1 \, d\nu - \lambda = \log \int e^g \, d\nu \,, \; g \in \mathcal{G}_\nu \,,$$

where the last equality comes from the optimal choice of $\lambda = -\log \int e^g \, d\nu$. Writing the divergence as its biconjugate then gives for all probability measure $\mu \in \mathcal{F}_\nu$

$$\begin{aligned} \mathrm{D}(\mu \,\|\, \nu) &= \sup_{g\in\mathcal{G}_\nu} \left\{ \mu(g) - \log \int e^g \, d\nu \right\} \\ &= \sup_{g\in\mathcal{G}_\nu} \left\{ \mu(g) - \nu(g) - \log \int e^{\left(g - \nu(g)\right)} \, d\nu \right\} , \end{aligned}$$

which is the Donsker–Varadhan representation of the Kullback–Leibler divergence. Using the inequality $\log(x) \le x - 1$, $x > 0$, we see that the optimand in the previous supremum is pointwise (for all g) greater than the optimand in (7).

Example 3.7. For the family of divergences $\phi(x) = |x-1|^\alpha/\alpha$ for $\alpha \ge 1$, Jiao et al. (2017) used the variational representation given by $I_\phi(\mu \,\|\, \nu) = \sup_g \mu(g) - \nu(g) - \nu(\frac{|x|^\beta}{\beta})$ where $\beta \ge 1$ is such that $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, whereas the tight representation is $I_\phi(\mu \,\|\, \nu) = \sup_g \mu(g) - \nu(g) - \inf_\lambda \int \begin{cases} -x - \lambda - 1/\alpha & x + \lambda < -1 \\ \frac{|x+\lambda|^\beta}{\beta} & x + \lambda \ge -1 \end{cases} d\nu$ (where the $-x - \lambda - 1/\alpha \le \frac{|x+\lambda|^\beta}{\beta}$ for $x < -1$ case comes from the fact that $\phi(x) = \infty$ for $x < 0$). Note that the shift, in e.g. the case $\alpha = 2$, reduces the second term from the raw second moment $\nu(g^2)$ to something no larger than the variance $\nu((g - \nu(g))^2)$, which is potentially much smaller.

3.3. Optimal bounds relating ϕ-divergences and IPMs

For a fixed function g, the optimal lower bound of the ϕ-divergence as a function of the mean deviation $\mu(g) - \nu(g)$ is given by the constrained optimization problem $\inf_{\mu \ll \nu : \mu(g) - \nu(g) = \varepsilon} I_\phi(\mu \,\|\, \nu)$ where the infimum is taken over $\mu \in \mathcal{M}_1$. Fenchel's duality theorem then implies that one should consider functions of the form $t \cdot g$ for a multiplier (or scaling parameter) $t \ge 0$, which motivates the following definition.

Definition 3.8. For $g \in \mathcal{G}_\nu$, we define the ϕ-cumulant generating function $K_{g,\nu} : \mathbb{R} \to \overline{\mathbb{R}}$ by

$$\begin{aligned} K_{g,\nu}(t) &\coloneqq \inf_{\lambda\in\mathbb{R}} \int \psi^\star(tg + \lambda) \, d\nu \qquad (9) \\ &= \tilde{I}^\star_{\phi,\nu}(tg) - t \cdot \nu(g) \,, \qquad (10) \end{aligned}$$

where $\psi : \mathbb{R} \to \overline{\mathbb{R}}$ is defined by $\psi(x) \coloneqq \phi(x+1)$.

Proof. Equation (10) follows from the fact that $\psi^\star(x) = \phi^\star(x) - x$ and Proposition 3.3. □

Example 3.9. For the Kullback–Leibler divergence, we see by Example 3.6 that $K_{g,\nu}(t) = \log \nu(e^{t(g-\nu(g))})$, which is the standard (centered) cumulant generating function, thereby justifying the name.

Remark. We show in the full version of this work (Agrawal & Horel, 2020) that the ϕ-cumulant generating function retains many of the standard properties of the true cumulant generating function.

Using the above definition, we reformulate the strong duality result obtained in Section 3.2 and show that there is an exact equivalence between upper-bounding the ϕ-cumulant generating function of a function g and lower bounding the ϕ-divergence in terms of the mean deviation $\mu(g) - \nu(g)$. Although the proof is simple, this result will be central in obtaining lower bounds of the divergence in terms of IPMs in the next sections.

Theorem 3.10. *Let $B : \mathbb{R} \to \overline{\mathbb{R}}$ be a function and let $K_{g,\nu}$ be as in Definition 3.8 for some $g \in \mathcal{G}_\nu$ and $\nu \in \mathcal{M}_1$. Then the following two properties are equivalent:*

1. *for all $t \in \mathbb{R}$, $K_{g,\nu}(t) \le B(t)$.*
2. *for all $\mu \in \mathcal{M}_1 \cap \mathcal{F}_\nu$, $I_\phi(\mu \,\|\, \nu) \ge B^\star\big(\mu(g) - \nu(g)\big)$.*

Proof. We prove the theorem by a sequence of equivalences, starting from 2., where every line is quantified over all $\mu \in \mathcal{M}_1 \cap \mathcal{F}_\nu$ and $t \in \mathbb{R}$:

$$\begin{aligned} & I_\phi(\mu \,\|\, \nu) \ge B^\star\big(\mu(g) - \nu(g)\big) \,, \\ \iff & I_\phi(\mu \,\|\, \nu) \ge t\big(\mu(g) - \nu(g)\big) - B(t) \,, \\ \iff & B(t) \ge \mu(tg) - I_\phi(\mu \,\|\, \nu) - t \cdot \nu(g) \,, \\ \iff & B(t) \ge \tilde{I}^\star_{\phi,\nu}(tg) - t \cdot \nu(g) = K_{g,\nu}(t) \,, \end{aligned}$$

where the first equivalence is by definition of $B^\star$, the second is by rearranging the terms and the last one is by definition of $\tilde{I}^\star_{\phi,\nu}$ (recall that $\tilde{I}_{\phi,\nu}(\mu) = I_\phi(\mu \,\|\, \nu)$ if $\mu \in \mathcal{M}_1 \cap \mathcal{F}_\nu$ and $+\infty$ otherwise). □

Remark. Another way to state the Theorem is that $K^\star_{g,\nu}$ is the best lower bound on the divergence by a lsc function of the deviation, and in particular is the best bound possible except for possibly the (at most

2) discontinuity points of the best bound. Indeed, the best bound is $L(\varepsilon) = \inf_{\zeta \ll \nu : \zeta(g)-\nu(g)=\varepsilon} I_\phi(\zeta \,\|\, \nu) = \inf_{\zeta \ll \nu} I_\phi(\zeta \,\|\, \nu) + \delta_{\{0\}}(\zeta(g)-\nu(g)-\varepsilon)$ which is a convex function of ε, and since $I_\phi(\mu \,\|\, \nu) \geq L(\mu(g)-\nu(g)) \geq L^{\star\star}(\mu(g)-\nu(g))$ by definition, Theorem 3.10 implies $K_{g,\nu} \leq L^\star$ and $K^\star_{g,\nu} \geq L^{\star\star}$ by the order-reversing property of the conjugate. We conclude that $K^\star_{g,\nu} = L$ is the best possible bound except at the most 2 points where $L \neq L^{\star\star}$ (which can occur only at the boundary of the domain of L).

Note that in Theorem 3.10, the function B can depend on ν and g (as just discussed, the best choice among lsc functions is $B = K_{g,\nu}$). When the bound B holds uniformly over a class of functions or measures one obtains the following useful corollaries, proven in the full version of this paper (Agrawal & Horel, 2020).

First, keeping ν fixed, the best convex lsc lower bound on the divergence in terms of an IPM results from taking the supremum of the functions $K_{g,\nu}$ for g in the class.

Corollary 3.11. *Let $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}_{\geq 0}$ be a convex lsc function with $L(0) = 0$, $\nu \in \mathcal{M}_1$ be a probability measure, and $\mathcal{G} \subseteq \mathcal{G}_\nu$ be a non-empty set of measurable functions closed under negation. Then defining $K_{\mathcal{G},\nu}(t) = \sup_{g \in \mathcal{G}} K_{g,\nu}(t)$, the following are equivalent:*

1. *for all $t \in \mathbb{R}$, $K_{\mathcal{G},\nu}(t) \leq L^\star(|t|)$.*
2. *for all $\mu \in \mathcal{M}_1 \cap \mathcal{F}_\nu$, $I_\phi(\mu \,\|\, \nu) \geq L(d_{\mathcal{G}}(\mu,\nu))$.*

In particular, $I_\phi(\mu \,\|\, \nu) \geq K^\star_{\mathcal{G},\nu}(d_{\mathcal{G}}(\mu,\nu))$, and if $I_\phi(\mu \,\|\, \nu) \geq L(d_{\mathcal{G}}(\mu,\nu))$ then $K^\star_{\mathcal{G},\nu} \geq L$, i.e. $K^\star_{\mathcal{G},\nu}$ is the best convex lsc lower bound on $I_\phi(\mu \,\|\, \nu)$ in terms of $d_{\mathcal{G}}(\mu,\nu)$.

Remark. Given an arbitrary function $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$, the function L_{fix} defined by $L_{\text{fix}}(x) = \max\{0, L(x)\}$ for $x > 0$ and $L_{\text{fix}}(0) = 0$ is a lower bound on $I_\phi(\mu \,\|\, \nu)$ in terms of $d_{\mathcal{G}}(\mu,\nu)$ if and only if L is, and furthermore $L^{\star\star}_{\text{fix}} \leq L_{\text{fix}}$ satisfies the conditions of the corollary.

In another common case, where we wish to consider an IPM consisting of bounded functions with respect to all pairs of measures, we get a similar result, this time without the requirement that $\mu \ll \nu$.

Corollary 3.12. *Let $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}_{\geq 0}$ be a convex lsc function with $L(0) = 0$, and let $\mathcal{G}$ a non-empty of set of measurable functions $g : \Omega \to \mathbb{R}$ such that $\sup_{\omega \in \Omega} |g(\omega)| < \infty$. Then defining $K_{\mathcal{G}}(t) = \sup_{\nu \in \mathcal{M}_1} K_{\mathcal{G},\nu}(t)$ using the definition of $K_{\mathcal{G},\nu}$ in Corollary 3.11, the following are equivalent:*

1. *for all $t \in \mathbb{R}$, $K_{\mathcal{G}}(t) \leq L^\star(|t|)$.*
2. *for all $(\mu,\nu) \in \mathcal{M}_1^2$, $I_\phi(\mu \,\|\, \nu) \geq L(d_{\mathcal{G}}(\mu,\nu))$.*

In particular, $I_\phi(\mu \,\|\, \nu) \geq K^\star_{\mathcal{G}}(d_{\mathcal{G}}(\mu,\nu))$, and if $I_\phi(\mu \,\|\, \nu) \geq L(d_{\mathcal{G}}(\mu,\nu))$ then $K^\star_{\mathcal{G}} \geq L$, i.e. $K^\star_{\mathcal{G}}$ is the best convex lsc lower bound on $I_\phi(\mu \,\|\, \nu)$ in terms of $d_{\mathcal{G}}(\mu,\nu)$.

4. Bounded functions

In this section, we show how to apply the results of Section 3 to obtain lower bounds of the ϕ-divergence by the total variation distance. In Section 4.1, we derive quadratic upper bounds on the ϕ-cumulant generating function, implying Pinsker's type inequalities. In Section 4.2, we study Vajda's problem: obtaining the *best* lower bound of the ϕ-divergence by a function of the total variation distance.

4.1. Pinsker-type inequalities

Corollary 3.12 implies that Pinsker-type inequalities, which are quadratic lower bounds on the divergence $I_\phi(\mu \,\|\, \nu)$ in terms of the total variation $\mathrm{TV}(\mu,\nu)$, are equivalent to quadratic upper bounds on the function $K_{g,\nu}(t)$, as in Hoeffding's lemma. Under certain mild assumptions on the function ϕ, one can uniformly bound the second derivate $K''_{g,\nu}(t)$, thereby obtaining these standard results and generalizing them to a broad class of ϕ-divergences.

Proposition 4.1. *Let $g \in \mathcal{L}^\infty_\nu$ and denote by m (resp. M) the essential infimum (resp. supremum) of g with respect to ν. Assume that ϕ is strictly convex and continuously twice differentiable on its domain and that $1/\phi''$ is concave. Let $\ell = \lim_{x\to\infty} \phi(x)/x$ and let us further assume that $\lim_{x \to \ell^-} \psi^\star(x) = +\infty$ (this is true in particular when $\ell = +\infty$). Then*

$$K_{g,\nu}(t) \leq \frac{(M-m)^2}{8 \cdot \phi''(1)} t^2\,, \quad t \in \mathbb{R}\,.$$

Example 4.2. For the KL divergence, we have that $\phi(x) = x\log(x) - x + 1$ and $\phi''(x) = 1/x$, so that $1/\phi''(x) = x$ is concave, and thus the above Proposition implies $\log \nu(e^{t(g-\nu(g))}) \leq t^2(M-m)^2/8$, which is exactly the standard statement of Hoeffding's lemma.

The condition that $1/\phi''$ concave is not satisfied by all common ϕ-divergences, but we give a similar result (with a slightly worse constant) when ϕ'' is monotone.

Proposition 4.3. *Let $g \in \mathcal{L}^\infty_\nu$ and denote by m (resp. M) the essential infimum (resp. supremum) of g with respect to ν. Assume that ϕ is twice differentiable on its domain and that ϕ'' monotone, then*

$$K_{g,\nu}(t) \leq \frac{(M-m)^2}{2 \cdot \phi''(1)} t^2\,, \quad t \in \mathbb{R}\,.$$

Corollary 4.4. *If ϕ satisfies the conditions of Proposition 4.1, then for every $(\mu,\nu) \in \mathcal{M}_1^2$ we have $I_\phi(\mu \,\|\, \nu) \geq \frac{\phi''(1)}{2}\mathrm{TV}(\mu,\nu)^2$, and if ϕ satisfies the conditions of Proposition 4.3 we have $I_\phi(\mu \,\|\, \nu) \geq \frac{\phi''(1)}{8}\mathrm{TV}(\mu,\nu)^2$.*

Proof. Recall that the total variation distance is the IPM associated with the class of bounded functions taking values in the interval $[-1, 1]$. For such a function g, the upper bound on $K_{g,\nu}(t)$ takes the form $t^2/(2\phi''(1))$ in Proposition 4.1 and $2t^2/\phi''(1)$ in Proposition 4.3. We conclude by Corollary 3.12 since the convex conjugate of $t \mapsto at^2$ is $t \mapsto t^2/(4a)$ for any $a > 0$. □

Example 4.5. Since by Example 4.2 the KL divergence satisfies the conditions of Proposition 4.1 and $\phi''(1) = 1$, Corollary 4.4 implies that $I_\phi(\mu \,\|\, \nu) \geq \frac{1}{2}\mathrm{TV}(\mu, \nu)^2$, which is the standard statement of Pinsker's inequality.

Example 4.6. The α-divergence given by $\phi_\alpha(x) = \frac{x^\alpha - \alpha(x-1) - 1}{\alpha(\alpha-1)}$ has $\phi''_\alpha(x) = x^{\alpha-2}$ which is monotone for all α and hence Corollary 4.4 implies that $I_{\phi_\alpha}(\mu \,\|\, \nu) \geq \frac{1}{8} \cdot \mathrm{TV}(\mu, \nu)^2$ for all α, which appears to be previously unknown for $\alpha > 2$. For $\alpha \in [1, 2]$, we also have that $1/\phi''_\alpha(x) = x^{2-\alpha}$ is concave and Corollary 4.4 then implies the tighter bound $I_{\phi_\alpha}(\mu \,\|\, \nu) \geq \frac{1}{2} \cdot \mathrm{TV}(\mu, \nu)^2$. This tighter bound was already observed in Gilardoni (2010) and in fact shown to hold for any $\alpha \in [-1, 2]$. We recover this more general result via a different technique in the full version of this work (Agrawal & Horel, 2020).

4.2. Vajda's problem

The quadratic Pinsker-type bounds derived in the previous subsection are easy to understand and apply, but are useful only in the regime when the divergence $I_\phi(\mu \,\|\, \nu)$ is less than some absolute constant, since $0 \leq \frac{1}{2}\mathrm{TV}(\mu, \nu)^2 \leq 2$ always, but $\mathrm{TV}(\mu, \nu) \to 2$ implies $\mathrm{D}(\mu \,\|\, \nu) \to \infty$. The Vajda problem (Vajda, 1972) is to quantify the optimal relationship between the two, that is to compute the function

$$L_\phi(\varepsilon) = \inf_{\substack{\mu,\nu \in \mathcal{M}_1 \\ \mathrm{TV}(\mu,\nu) = \varepsilon}} I_\phi(\mu \,\|\, \nu) \,.$$

This problem was solved in the case of the KL divergence by Fedotov et al. (2003), and for the general case of an arbitrary ϕ by Gilardoni (2010), but we rederive these bounds here with a new geometric interpretation to show the applicability of our techniques. In particular, the duality result of Corollary 3.12 reduces the problem of computing L_ϕ to the problem of computing the best upper bound on $K_{g,\nu}$ for all probability measures ν and functions g in the set $\mathcal{B}$ of all measurable functions from Ω to $[-1, 1]$.

Lemma 4.7. *Let* $K_{\mathcal{B}}(t) = \sup_{\nu \in \mathcal{M}_1, g \in \mathcal{B}} K_{\nu,g}(t)$. *Then* $L_\phi(\varepsilon) = K^\star_{\mathcal{B}}(\varepsilon)$ *for all* $\varepsilon \geq 0$ *and* $L^\star_\phi(t) = K_{\mathcal{B}}(t)$.

Proof. The Vajda function is convex and lower semi-continuous (Vajda, 1972), so by Corollary 3.12 we get that $I_\phi(\mu \,\|\, \nu) \geq L_\phi(\mathrm{TV}(\mu, \nu))$ implies $I_\phi(\mu \,\|\, \nu) \geq K^\star_{\mathcal{B}}(\mathrm{TV}(\mu, \nu)) \geq L_\phi(\mathrm{TV}(\mu, \nu))$, so since L_ϕ is by definition the largest function lower bounding $I_\phi(\mu \,\|\, \nu)$ we get $L_\phi = K^\star_{\mathcal{B}}$ as desired. Finally, $K_{\mathcal{B}}$ is convex and lsc as the supremum of convex lsc functions, and so $L^\star_\phi = K^{\star\star}_{\mathcal{B}} = K_{\mathcal{B}}$. □

The main result of this section establishes that L_ϕ is the convex conjugate of the following function to which we associate a natural geometric interpretation below.

Definition 4.8. The *height-for-width* function $\mathrm{hgt}_{\psi^\star} : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$ associated with $\psi^\star$ is given by the following equivalent definitions:

1. $\mathrm{hgt}_{\psi^\star}(t) = \inf_\lambda \max\{\psi^\star(\lambda + t/2), \psi^\star(\lambda - t/2)\}$. Furthermore, if there exists $\lambda(t)$ such that $\psi^\star(\lambda(t) + t/2) = \psi^\star(\lambda(t) - t/2)$, then $\mathrm{hgt}_{\psi^\star}(t) = \psi^\star(\lambda(t) + t/2) = \psi^\star(\lambda(t) - t/2)$.
2. $\mathrm{hgt}_{\psi^\star}(t) = \inf\{y \in \overline{\mathbb{R}} : \ell(\{x : \psi^\star(x) \leq y\}) \geq t\}$ where $\ell([a, b]) = b - a$ is the length of the interval $\{x : \psi^\star(x) \leq y\}$ (recall that $\psi^\star$ is convex, coercive, and lsc so that this set is a compact interval).

Note that the second definition expresses $\mathrm{hgt}_{\psi^\star}(t)$ as the (right) inverse of the *sublevel set volume* function $y \mapsto \ell(\{x : \psi^\star(x) \leq y\})$ mapping a level y to the Lebesgue measure of the associated sublevel set of $\psi^\star$. Equivalently, the sublevel set volume function maps a level y to the length of the longest horizontal segment which can be placed in the epigraph of $\psi^\star$ subject to being no higher than y. Hence, its inverse, the height-for-width function, asks for the minimal height at which one can place a horizontal segment of length t in the epigraph of $\psi^\star$. This is shown to be equivalent to the first definition in the full version of this work (Agrawal & Horel, 2020) and Figure 1 illustrates this definition in the case of the Kullback–Leibler divergence.

Example 4.9. For the case of the KL divergence (e.g. $\psi^*(t) = e^t - t - 1$), one can compute that $\psi^*(\lambda(t) + t/2) = \psi^*(\lambda(t) - t/2)$ for $\lambda(t) = -\log \frac{e^{t/2} - e^{-t/2}}{t} = -\log \frac{2\sinh(t/2)}{t}$, so that $\mathrm{hgt}_{\psi^*}(t) = -1 + \frac{t}{2}\coth\frac{t}{2} + \log \frac{2\sinh(t/2)}{t}$.

We are now ready to state the main result of this section.

Proposition 4.10. $K_{\mathcal{B}}(t) = \mathrm{hgt}_{\psi^\star}(2t)$.

Proof. Our goal is to use a minimax theorem to swap the supremum and infimum in the definition of $K_{\mathcal{B}}(t)$, since for fixed λ, by convexity of $\psi^\star$ the supremum $\sup_{\nu \in \mathcal{M}_1, g \in \mathcal{B}} \int \psi^\star(tg + \lambda)\, d\nu$ is achieved by taking g to be either the constant 1 or the constant -1. However, the set $\mathcal{M}_1 \times \mathcal{B}$ is not necessarily compact, which is required to apply the standard minimax theorem of Sion (1958). But as we show in the full version of this work (Agrawal & Horel, 2020), the optimization problem has an euqivalent formulation over a finite dimensional compact convex set, so that the minimax theorem can be applied. □

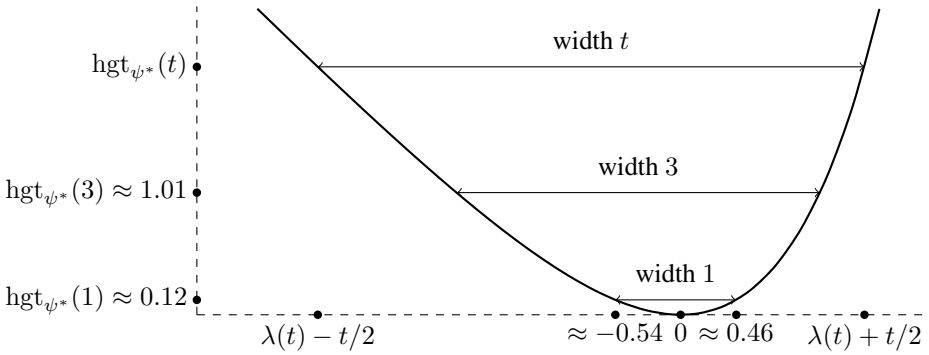

Figure 1. Illustration of height-for-width function for $\psi^\star(x) = e^x - x - 1$

Example 4.11. For the KL divergence, using the fact that $K_{g,\nu}(t) = \log \nu(e^{t(g-\nu(g))})$, Proposition 4.10 and Example 4.9 imply that the optimal bound on the cumulant generating function of a random variable g with $\nu(g) = 0$ and $m \le g \le M$ ν-a.s. is $\log \nu(e^{tg}) \le \mathrm{hgt}_{\psi^\star}[(M-m)t] = -1 + \frac{M-m}{2}\coth\frac{M-m}{2} + \log\frac{2\sinh((M-m)t/2)}{t}$, which strengthens the standard quadratic upper bound (Hoeffding's lemma) derived in Example 4.2.

Corollary 4.12. $L_\phi(\varepsilon) = \mathrm{hgt}^\star_{\psi^\star}(\varepsilon/2)$. *In particular, if* $\mathrm{hgt}_{\psi^\star}$ *is differentiable then* $L_\phi(2\,\mathrm{hgt}'_{\psi^\star}(x)) = x\,\mathrm{hgt}'_{\psi^\star}(x) - \mathrm{hgt}_{\psi^\star}(x)$.

Proof. $L_\phi(\varepsilon) = K^\star_{\mathcal{B}}(\varepsilon)$ and $K_{\mathcal{B}}(t) = \mathrm{hgt}_{\psi^\star}(2t)$, so that $L_\phi(\varepsilon) = \mathrm{hgt}^\star_{\psi^\star}(\varepsilon/2)$. The supplemental claim follows from the explicit expression for the convex conjugate. □

Example 4.13. For the KL divergence, using Example 4.9, Corollary 4.12 applied to $x = 2t$ gives $L_\phi(V(t)) = \log\frac{t}{\sinh t} + t\coth t - \frac{t^2}{\sinh^2 t}$ for $V(t) = 2\coth t - \frac{t}{\sinh^2 t} - 1/t$, which is exactly the formula derived by Fedotov et al. (2003).

5. Unbounded functions

We now apply the duality framework of Section 3 to spaces containing unbounded functions. We start with the case of a single unbounded function g absolutely integrable with respect to ν and obtain as a corollary of Theorem 3.10 an equivalence between lower bounding $I_\phi(\mu\,\|\,\nu)$ as a function of the mean deviation $\mu(g) - \nu(g)$, and upper bounding the cumulant function of g.

Corollary 5.1. *Let* $B : \mathbb{R} \to \overline{\mathbb{R}}$ *be a function and let* $g : \Omega \to \mathbb{R}$ *be a (possibly unbounded) function such that* $\nu(|g|) < \infty$. *Then the following two properties are equivalent:*

1. *for all* $t \in \mathbb{R}$, $K_{g,\nu}(t) \le B(t)$.
2. *for all* $\mu \in \mathcal{M}_1$ *with* $\mu \ll \nu$ *and* $\mu(|g|) < \infty$, $I_\phi(\mu\,\|\,\nu) \ge B^\star\big(\mu(g) - \nu(g)\big)$.

Proof. Writing $\overline{g} = \{g \cdot \mathbf{1}_A : A \in \mathcal{A}\}$, it is easy to see that $\mathcal{G}_\nu := \mathrm{Span}(\overline{g} \cup \mathcal{L}^\infty_\nu)$ is the smallest decomposable space containing g and that $\mathcal{L}^\infty_\nu \subset \mathcal{G}_\nu$. Furthermore, defining $\mathcal{F}_\nu = \{\mu \in \mathcal{L}^1_\nu : |\mu|(|g|) < \infty\}$, we verify that $\mathcal{F}_\nu$ is also a decomposable space and that $(\mathcal{F}_\nu, \mathcal{G}_\nu)$ form a dual pair. For decomposability, consider $\mu \in \mathcal{L}^1_\nu$ with $|\mu|(|g|) < \infty$, $A \in \mathcal{A}$ and $m \in \mathcal{L}^\infty_\nu$. Define $\xi \ll \nu$ whose derivative is given by $d\xi/d\nu = d\mu/d\nu \cdot \mathbf{1}_A + m \cdot \mathbf{1}_{\Omega\setminus A}$, then we want to show that $|\xi|(|g|) < \infty$. This is immediate since the triangle inequality gives $|\xi|(|g|) \le |\mu|(|g|) + \|m\|_{\nu,\infty}|\nu|(|g|) < \infty$. Similarly, the fact that for $\mu \in \mathcal{F}_\nu$, $h \in \mathcal{G}_\nu$, $\mu(h) < \infty$ easily follows from the triangle inequality. We can thus apply Theorem 3.10 to the pair $(\mathcal{F}_\nu, \mathcal{G}_\nu)$ and obtain the desired conclusion. □

We now turn to the question of obtaining Pinsker-type inequalities, that is giving a lower bound of the divergence by a quadratic function of the mean deviation. Corollary 5.1 implies that a necessary condition is that $K_{g,\nu}(t) \le c \cdot t^2/2$ for some $c > 0$. This motivates the following definition.

Definition 5.2 (ϕ-subgaussian)**.** Let $g \in \mathcal{L}^0$, we say that g is (c,ν) ϕ*-subgaussian* if $K_{g,\nu}(t) \le c \cdot t^2/2$, $t \in \mathbb{R}$. We denote by $\mathcal{G}_{c,\nu}$ the set of all (c,ν) ϕ-subgaussian functions.

As already seen, for $\phi : x \mapsto x\log x - (x-1)$, $K_{g,\nu}$ is the log moment generating function, and the previous definition generalizes the standard definition of subgaussian random variables. Furthermore, if ϕ satisfies the assumptions of either Proposition 4.1 or Proposition 4.3, the same propositions imply that all bounded functions with fixed-length range are (c,ν)-subgaussian for some constant c.

The following proposition shows that the class of subgaussian functions is the largest class of functions for which Pinsker's type inequalities can be obtained.

Proposition 5.3. *Let* $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ *be a class of measurable*

functions and $c \geq 0$, then the following two propositions are equivalent.

1. $\mathcal{G} \subseteq \mathcal{G}_{c,\nu}$.
2. *for all* $\mu \in \mathcal{M}_1 \cap \mathcal{L}^1_\nu$, $I_\phi(\mu \,\|\, \nu) \geq \frac{1}{2c^2} \cdot d_{\mathcal{G}}(\mu, \nu)^2$.

Proof. We proceed similarly to the proof of Corollary 3.11. First, assume 1. and consider $\mu \in \mathcal{M}_1$ with $\mu \ll \nu$. Either $I_\phi(\mu \,\|\, \nu) = +\infty$ and 2. is trivially true. Otherwise consider $g \in \mathcal{G}$. In the full version of this paper (Agrawal & Horel, 2020), we show that since $K_{g,\nu}$ is defined everywhere we have that $\nu(|g|) < \infty$ and $\mu(|g|) < \infty$. But then Corollary 5.1 implies $I_\phi(\mu \,\|\, \nu) \geq \frac{1}{2c^2}\big(\mu(g) - \nu(g)\big)^2$. Taking the supremum over $g \in \mathcal{G}$ gives 2. The reverse direction follows immediately from an application of Corollary 5.1 to each function $g \in \mathcal{G}$. □

Example 5.4. The χ^2-divergence, corresponding to $\phi(x) = (x-1)^2$ for $x \geq 0$, has

$$\psi^\star(x) = \begin{cases} x^2/4 & x \geq -2 \\ -1 - x & x < -2 \end{cases} \leq x^2/4\,,$$

so that $K_{g,\nu}(t) \leq \inf_\lambda \int (tg + \lambda)^2/4 \, d\nu = t^2 \operatorname{Var}_\nu(g)/4$, and in particular the class of "χ^2-subgaussian" random variables includes all those with bounded variance.

Divergences weaker than the KL. One notable feature of the Kullback–Leibler divergence is that it heavily penalizes large likelihood ratios $\frac{d\mu}{d\nu}$, and in particular, if $\mu \not\ll \nu$ then $I(\mu \,\|\, \nu) = \infty$. If this behavior is undesirable, one may wish to consider a weaker divergence, specifically one with $\lim_{x\to\infty} \phi(x)/x < \infty$. However, we show in the proposition below (whose proof is given in the full version (Agrawal & Horel, 2020)) that for such divergences *no* unbounded function satisfies *any* nontrivial bound of the form $I_\phi(\mu \,\|\, \nu) \geq L(|\mu(g) - \nu(g)|)$ for all μ. In particular, control on the absolute deviation of any unbounded function in terms of a ϕ-divergence I_ϕ requires $\lim_{x\to\infty} \phi(x)/x = \infty$.

Proposition 5.5. *Let $I_\phi(\mu \,\|\, \nu)$ be the ϕ-divergence associated to a function ϕ satisfying $\lim_{x\to\infty} \phi(x)/x = \ell < \infty$, and let $g \in \mathcal{L}^1_\nu \setminus \mathcal{L}^\infty_\nu$ be a ν-integrable but unbounded function. Then for any function $L : \mathbb{R}_{\geq 0} \to \bar{\mathbb{R}}_{\geq 0}$ such that $I_\phi(\mu \,\|\, \nu) \geq L(|\mu(g) - \nu(g)|)$ for all $\mu \in \mathcal{M}_1$ with $\mu(|g|) < \infty$, we have that in fact $L(x) = 0$ for all $x \geq 0$.*

Acknowledgements

We are grateful to Flavio du Pin Calmon, Julien Fageot, Salil Vadhan, and the anonymous reviewers for their helpful comments and suggestions on an earlier draft of this paper.

Rohit Agrawal was funded by the Department of Defense (DoD) through the National Science Defense & Engineering Graduate Fellowship (NDSEG) Program.

References

Agrawal, R. and Horel, T. Optimal Bounds between f-Divergences and Integral Probability Metrics. *arXiv:2006.05973*, June 2020.

Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B*, 28(1):131–142, 1966. ISSN 00359246. doi: 10.2307/2984279.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Berg, C., Christensen, J. P. R., and Ressel, P. *Introduction to Locally Convex Topological Vector Spaces and Dual Pairs*, pp. 1–15. Springer, New York, NY, 1984. ISBN 978-1-4612-1128-0. doi: 10.1007/978-1-4612-1128-0_1.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. ISBN 9780199535255. doi: 10.1093/acprof:oso/9780199535255.001.0001.

Csiszár, I. Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(1–2):85–108, 1963.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

Donsker, M. D. and Varadhan, S. R. S. Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on Pure and Applied Mathematics*, 29(4):389–461, 1976. doi: 10.1002/cpa.3160290405.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, pp. 258–267, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Fedotov, A. A., Harremoës, P., and Topsøe, F. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, June 2003. doi: 10.1109/TIT.2003.811927.

Gilardoni, G. L. On the minimum f-divergence for given total variation. *Comptes Rendus Mathematique*, 343(11): 763 – 766, 2006. ISSN 1631-073X. doi: 10.1016/j.crma.2006.10.027.

Gilardoni, G. L. On Pinsker's and Vajda's type inequalities for Csiszár's f-divergences. *IEEE Transactions on Information Theory*, 56(11):5377–5386, Nov 2010. doi: 10.1109/TIT.2010.2068710.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(25):723–773, March 2012. ISSN 1532-4435.

Guntuboyina, A., Saha, S., and Schiebinger, G. Sharp inequalities for f-divergences. *IEEE Transactions on Information Theory*, 60(1):104–121, Jan 2014. doi: 10.1109/TIT.2013.2288674.

Harremoës, P. and Vajda, I. On Pairs of f-Divergences and Their Joint Range. *IEEE Transactions on Information Theory*, 57(6):3230–3235, June 2011. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2011.2137353.

Jiao, J., Han, Y., and Weissman, T. Dependence measures bounding the exploration bias for general measurements. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 1475–1479, June 2017. doi: 10.1109/ISIT.2017.8006774.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. ISSN 00018678. doi: 10.2307/1428011.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T. (eds.), *Advances in Neural Information Processing Systems 20*, pp. 1089–1096. Curran Associates, Inc., 2008.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theor.*, 56 (11):5847–5861, November 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2068870.

Nock, R., Cranko, Z., Menon, A. K., Qu, L., and Williamson, R. C. f-GANs in an information geometric nutshell. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 456–464, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 271–279, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Rockafellar, R. T. Integrals which are convex functionals. *Pacific J. Math.*, 24(3):525–539, 1968.

Rockafellar, R. T. Integral functionals, normal integrands and measurable selections. In Gossez, J. P., Lami Dozo, E. J., Mawhin, J., and Waelbroeck, L. (eds.), *Nonlinear Operators and the Calculus of Variations*, pp. 157–207, Berlin, Heidelberg, 1976. Springer Berlin Heidelberg. ISBN 978-3-540-38075-7. doi: 10.1007/BFb0079944.

Ruderman, A., Reid, M., García-García, D., and Petterson, J. Tighter variational representations of f-divergences via restriction to probability measures. In Langford, J. and Pineau, J. (eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML '12)*, pp. 671–678, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1.

Russo, D. and Zou, J. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, Jan 2020. ISSN 1557-9654. doi: 10.1109/TIT.2019.2945779.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, March 1958. ISSN 0030-8730, 0030-8730. doi: 10.2140/pjm.1958.8.171.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. R. G., and Schölkopf, B. A note on integral probability metrics and ϕ-divergences. *CoRR*, abs/0901.2698v1, 2009.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. doi: 10.1214/12-EJS722.

Vajda, I. On the f-divergence and singularity of probability measures. *Periodica Mathematica Hungarica*, 2(1): 223–234, Mar 1972. ISSN 1588-2829. doi: 10.1007/BF02018663.

Zălinescu, C. *Convex Analysis in General Vector Spaces*. World Scientific, River Edge, N.J. ; London, 2002. ISBN 978-981-238-067-8.