Journal of Machine Learning Research 22 (2021) 1-59 Submitted 8/20; Revised 5/21; Published 5/21

Optimal Bounds between f-Divergences and Integral Probability Metrics[*]

Rohit Agrawal ROHITAGR@SEAS.HARVARD.EDU
Harvard John A. Paulson School of Engineering and Applied Sciences
Cambridge, MA 02138, USA

Thibaut Horel THIBAUTH@MIT.EDU
Institute for Data, Systems, and Society
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Gábor Lugosi

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 widely used to quantify the similarity between probability distributions. 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 while also recovering in a unified manner well-known results, such as Hoeffding's lemma, Pinsker's inequality and its extension to subgaussian functions, and the Hammersley–Chapman–Robbins bound. This characterization also allows us to prove new results on topological properties of the divergence which may be of independent interest.

Keywords: f-Divergence, Integral Probability Metrics, Probability Inequalities, Convex Analysis, Convergence of Measures

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. In a line of research going back at least to the work of Kullback (1959), information theoretic measures of dissimilarity between probability distributions have provided a fruitful and unifying perspective on a wide range of statistical procedures. A prototypical example of this perspective is the interpretation of maximum likelihood estimation as 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 and

[*] An extended abstract of this work appeared as Agrawal and Horel (2020).

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

Silvey, 1966).[2] Informally, a ϕ-divergence quantifies the divergence between two distributions μ and ν as an average cost of the likelihood ratio, that is, $\mathrm{D}_\phi(\mu \parallel \nu) := \int \phi(d\mu/d\nu)\,\mathrm{d}\nu$ 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 and Varadhan, 1976, Theorem 5.2) expresses the Kullback–Leibler divergence $\mathrm{D}(\mu \parallel \nu)$ between probability distributions μ and ν as

$$\mathrm{D}(\mu \parallel \nu) = \sup_{g \in \mathcal{L}^b} \left\{ \int g \,\mathrm{d}\mu - \log \int e^g \,\mathrm{d}\nu \right\}, \tag{1}$$

where $\mathcal{L}^b$ is the space of bounded measurable functions. 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 (1) to a class of functions $\mathcal{G} \subseteq \mathcal{L}^b$ 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 representations such as (1) as a two-player game between 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

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

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

Rewriting the optimization problem (1) as

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

reveals 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 $\int g \,\mathrm{d}\mu - \int g \,\mathrm{d}\nu$, 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:

[2] ϕ-divergences had previously been considered (Rényi, 1961; Morimoto, 1963), though not as an independent object of study.

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.

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, §4.10) shows 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 *subgaussian* functions. This was recently used by Russo and 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 a given IPM. As an application, we show that this function grows quadratically if and only if the ϕ-divergence can be lower bounded by a quadratic function of the given IPM and recover in a unified manner the extension of Pinsker's inequality to subgaussian functions and the Hammersley–Chapman–Robbins bound.
- we characterize the existence of *any* non-trivial lower bound on an IPM in terms of the generalized log moment-generating function, and give implications for topological properties of the divergence, for example regarding compactness of sets of measures with bounded ϕ-divergence and the relationship between convergence in ϕ-divergence and weak convergence.
- the answer to Vajda's problem for bounded functions 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. From this, we derive a refinement of Hoeffding's lemma and generalizations of Pinsker's inequality to a large class of ϕ-divergences.

The rest of this paper is organized as follows: Section 2 discusses related work, Section 3 gives a brief overview of concepts and tools used in this paper, Section 4 derives the tight variational representation of the ϕ-divergence, Section 5 focuses on the case of an IPM given by a single function g with respect to a reference measure ν, deriving the optimal bound in this case and discussing topological applications, and Section 6 extends this to arbitrary IPMs and sets of measures, with applications to subgaussian functions and Vajda's problem.

2. Related work

The question studied in the present paper is an instance of the broader problem of the constrained minimization of a ϕ-divergence, which has been extensively studied in works spanning information theory, statistics and convex analysis.

Kullback–Leibler divergence. The problem of minimizing the Kullback–Leibler divergence (Kullback and Leibler, 1951) subject to a convex constraint can be traced back at least to Sanov (1957) in the context of large deviation theory and to Kullback (1959) for the purpose of formulating an information theoretic approach to statistics. In information theory, this problem is known as an I-projection (Csiszár, 1975; Csiszár and Matúš, 2003). The case where the convex set is defined by finitely many affine equality constraints, which is closest to our work, was specifically studied in Ben-Tal and Charnes (1977, 1979) via a convex duality approach. This special case is of particular relevance to the field of statistics, since the exponential family arises as the optimizer of this problem.

Convex integral functionals and general ϕ. With the advent of the theory of convex integral functionals, initiated in convex analysis by Rockafellar (1966, 1968), the problem is generalized to arbitrary ϕ-divergences, sometimes referred to as ϕ-entropies, especially when seen as functionals over spaces of functions, and increasingly studied via a systematic application of convex duality (Teboulle and Vajda, 1993). In the case of affine constraints, the main technical challenge is to identify constraint qualifications guaranteeing that strong duality holds: Borwein and Lewis (1991, 1993); Broniatowski and Keziou (2006) investigate the notion of quasi-relative interior for this purpose, and Léonard (2001b,a) consider integrability conditions on the functions defining the affine constraints. A comprehensive account of this case can be found in Csiszár and Matúš (2012). We also note the work Altun and Smola (2006), which shows a duality between *approximate* divergence minimization—where the affine constraints are only required to hold up to a certain accuracy—and maximum a posteriori estimation in statistics.

At a high level, in our work we show in Section 6 that one can essentially reduce the problem of minimizing the divergence on probability measures subject to a constraint on an IPM to the problem of minimizing the divergence on finite measures subject to two affine constraints: the first restricting to probability measures, and the second constraining the mean deviation of a single function in the class defining the IPM. For the restriction to probability measures, we prove that constraint qualification always holds, a fact which was not observed in the aforecited works, to the best of our knowledge. For the second constraint, we show in Section 5.3 that by focusing on a single function, we can relate strong duality of the minimization problem to compactness properties of the divergence. In particular, we obtain strong duality under similar assumptions as those considered in Léonard (2001b), even when the usual interiority conditions for constraint qualification do not hold.

Relationship between ϕ-divergences. A specific case of the minimization question which has seen significant work is when the feasible set is defined by other ϕ-divergences, and most notably is a level set the total variation distance. The best-known result in this line is Pinsker's inequality, first proved in a weaker form in Pinsker (1960, 1964) and then strengthened independently in Kullback (1967); Kemperman (1969); Csiszár (1967), which gives the best possible quadratic lower bound on the Kullback–Leibler divergence by the total variation distance. More recently, for ϕ-divergences other than the Kullback–Leibler divergence, Gilardoni (2010) identified conditions on ϕ under which quadratic "Pinsker-type" lower bounds can be obtained.

More generally, the problem of finding the best lower bound of the Kullback–Leibler divergence as a (possibly non-quadratic) function of the total variation distance was introduced by Vajda in Vajda (1970) and generalized to arbitrary ϕ-divergences in Vajda (1972), and is therefore sometimes referred to as *Vajda's problem*. Approximations of the best lower bound were obtained in Bretagnolle and Huber (1979); Vajda (1970) for the Kullback–Leibler divergence and in Vajda (1972); Gilardoni (2008, 2010) for ϕ-divergences under various assumptions on ϕ. The optimal lower bound was derived in Fedotov et al. (2003) for the Kullback–Leibler divergence and in Gilardoni (2006) for any ϕ-divergence. As an example application of Section 6, in Section 6.3 we rederive the optimal lower bound as well as its quadratic relaxations in a unified manner.

In Reid and Williamson (2009, 2011), the authors consider the generalization of Vajda's problem of obtaining a tight lower bound on an arbitrary ϕ-divergence given multiple values of *generalized total variation distances*; their result contains Gilardoni (2006) as a special case. Beyond the total variation distance, Harremoës and Vajda (2011) introduced the general question of studying the *joint range* of values taken by an arbitrary pair of ϕ-divergences, which has its boundary given by the best lower bounds of one divergence as a function of the other. Guntuboyina et al. (2014) generalize this further and consider the general problem of understanding the joint range of multiple ϕ-divergences, i.e. minimizing a ϕ-divergence subject to a finite number of constraints on other ϕ-divergences. A key conceptual contribution in this line of work is to show that these optimization problems, which are defined over (infinitely dimensional) spaces of measures, can be reduced to finite dimensional optimization problems. A related line of work (Sason and Verdú, 2016; Sason, 2018) deriving relations between ϕ-divergences instead approaches the problem by defining integral representations of ϕ-divergences in terms of simple ones.

Our work differs from results of this type since we are primarily concerned with IPMs other than the total variation distance, and in particular with those containing unbounded functions. It was shown in Khosravifard et al. (2006, 2007); Sriperumbudur et al. (2009, 2012) that the class of ϕ-divergences and the class of pseudometrics (including IPMs) intersect *only* at the total variation distance. As such, the problem studied in the present paper cannot be phrased as the one of a joint range between two ϕ-divergences, and to the best of our knowledge cannot be handled by the techniques used in studying the joint range.

Transport inequalities. Starting with the work of Marton (1986), transportation inequalities upper bounding the Wasserstein distance by a function of the relative entropy have been instrumental in the study of the concentration of measure phenomenon (see e.g. Gozlan and Léonard (2010) for a survey). These inequalities are related to the question studied in this work since the 1-Wasserstein distance is an IPM when the probability space is a Polish space and coincides with the total variation distance when the probability space is discrete and endowed with the discrete metric. In an influential paper, Bobkov and Götze (1999) proved that upper bounding the 1-Wasserstein distance by a square root of the relative entropy is equivalent to upper bounding the log moment-generating function of all 1-Lipschitz functions by a quadratic function. The extension of Pinsker's inequality in Boucheron et al. (2013, §4.10), which was inspired by Bobkov and Götze (1999), is also based on quadratic upper bounds of the log moment-generating function and we in turn follow similar ideas in Sections 4.3 and 5.1 of the present work.

3. Preliminaries

3.1 Measure Theory

Notation. Unless otherwise noted, all the probability measures in this paper are defined on a common measurable space $(\Omega, \mathcal{F})$, which we assume is non-trivial in the sense that $\{\emptyset, \Omega\} \subsetneq \mathcal{F}$, as otherwise all questions considered in this paper become trivial. We denote by $\mathcal{M}(\Omega, \mathcal{F})$, $\mathcal{M}^+(\Omega, \mathcal{F})$ and $\mathcal{M}^1(\Omega, \mathcal{F})$ the sets of finite signed measures, finite non-negative measures, and probability measures respectively. $\mathcal{L}^0(\Omega, \mathcal{F})$ denotes the space of all measurable functions from Ω to $\mathbb{R}$, and $\mathcal{L}^b(\Omega, \mathcal{F}) \subseteq \mathcal{L}^0(\Omega, \mathcal{F})$ is the set of all bounded measurable functions. For $\nu \in \mathcal{M}(\Omega, \mathcal{F})$, and $1 \leq p \leq \infty$, $\mathcal{L}^p(\nu, \Omega, \mathcal{F})$ denotes the space of measurable functions with finite p-norm with respect to ν, and $L^p(\nu, \Omega, \mathcal{F})$ denotes the space obtained by taking the quotient with respect to the space of functions which are 0 ν-almost everywhere. Similarly, $L^0(\nu, \Omega, \mathcal{F})$ is the space of all measurable functions Ω to $\mathbb{R}$ up to equality ν-almost everywhere. When there is no ambiguity, we drop the indication $(\Omega, \mathcal{F})$. For a measurable function $f \in \mathcal{L}^0$ and measure $\nu \in \mathcal{M}$, $\nu(f) \coloneqq \int f \,\mathrm{d}\nu$ denotes the integral of f with respect to ν.

For two measures μ and ν, $\mu \ll \nu$ (resp. $\mu \perp \nu$) denotes that μ is absolutely continuous (resp. singular) with respect to ν and we define $\mathcal{M}_c(\nu) \coloneqq \{\mu \in \mathcal{M} \mid \mu \ll \nu\}$ and $\mathcal{M}_s(\nu) \coloneqq \{\mu \in \mathcal{M} \mid \mu \perp \nu\}$, so that by the Lebesgue decomposition theorem we have the direct sum $\mathcal{M} = \mathcal{M}_c(\nu) \oplus \mathcal{M}_s(\nu)$. For $\mu \in \mathcal{M}_c(\nu)$, $\frac{d\mu}{d\nu} \in L^1(\nu)$ denotes the Radon–Nikodym derivative of μ with respect to ν. For a signed measure $\nu \in \mathcal{M}$, we write the Hahn–Jordan decomposition $\nu = \nu^+ - \nu^-$ where $\nu^+, \nu^- \in \mathcal{M}^+$, and denote by $|\nu| = \nu^+ + \nu^-$ the total variation measure.

More generally, given a σ-ideal $\Sigma \subseteq \mathcal{F}$ we write $\mu \ll \Sigma$ to express that $|\mu|(A) = 0$ for all $A \in \Sigma$ and define $\mathcal{M}_c(\Sigma) \coloneqq \{\mu \in \mathcal{M} \mid \mu \ll \Sigma\}$. Similarly, $L^0(\Sigma)$ denotes the quotient of $\mathcal{L}^0$ by the space of functions equal to 0 except on an element of Σ. For a measurable function $f \in \mathcal{L}^0$, and σ-ideal Σ, $\operatorname{ess\,im}_\Sigma(f) \coloneqq \bigcap_{\varepsilon>0} \big\{x \in \mathbb{R} \bigm| f^{-1}\big((x-\varepsilon, x+\varepsilon)\big) \notin \Sigma\big\}$ is the essential range of f with respect to Σ, and $\operatorname{ess\,sup}_\Sigma f \coloneqq \sup \operatorname{ess\,im}_\Sigma(f)$ and $\operatorname{ess\,inf}_\Sigma f \coloneqq \inf \operatorname{ess\,im}_\Sigma(f)$ denote the Σ-essential supremum and infimum respectively. Finally $L^\infty(\Sigma)$ denotes the space of a functions whose Σ-essential range is bounded, up to equality except on an element of Σ. When Σ is the σ-ideal of null sets of a measure ν, we abuse notations and write $\operatorname{ess\,im}_\nu(f)$ for $\operatorname{ess\,im}_\Sigma(f)$ and similarly for the essential supremum and infimum.

Finally, for brevity, we define for a subspace $X \subseteq \mathcal{M}$ of finite signed measures the subsets $X^+ \coloneqq X \cap \mathcal{M}^+$ and $X^1 \coloneqq X \cap \mathcal{M}^1$, and for $\nu \in \mathcal{M}$ we also define $X_c(\nu) \coloneqq X \cap \mathcal{M}_c(\nu)$ and $X_s(\nu) \coloneqq X \cap \mathcal{M}_s(\nu)$.

Integral Probability Metrics.

Definition 1 *For a non-empty set of measurable functions $\mathcal{G} \subseteq \mathcal{L}^0$, the* integral probability metric *associated with $\mathcal{G}$ is defined by*

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

for all pairs of measures $(\mu, \nu) \in \mathcal{M}^2$ such that all functions in $\mathcal{G}$ are absolutely μ- and ν-integrable. We extend this definition to all pairs of measures $(\mu, \nu) \in \mathcal{M}^2$ by $d_{\mathcal{G}}(\mu, \nu) = +\infty$ in cases where there exists a function in $\mathcal{G}$ which is not μ- or ν- integrable.

Remark 2 *When the class $\mathcal{G}$ is closed under negation, one can drop the absolute value in the definition.*

Example 1 *The total variation distance $\mathrm{TV}(\mu, \nu)$ is obtained when $\mathcal{G}$ is the class of measurable functions taking values in $[-1, 1]$.*[3]

Example 2 *Note that the integrals $\int g \,\mathrm{d}\mu$ and $\int g \,\mathrm{d}\nu$ depend only on the pushforward measures $g_*\mu$ and $g_*\nu$ on $\mathbb{R}$. Equivalently, when μ and ν are the probability distributions of random variables X and Y taking values in Ω, we have that $\int g \,\mathrm{d}\mu = \int \mathrm{Id}_{\mathbb{R}} \,\mathrm{d}g_*\mu = \mathbb{E}[g(X)]$, the expectation of the random variable $g(X)$, and similarly $\int g \,\mathrm{d}\nu = \mathbb{E}[g(Y)]$. The integral probability metric $d_{\mathcal{G}}$ thus defines the distance between random variables X and Y as the largest difference in expectation achievable by "observing" X and Y through a function from the class $\mathcal{G}$.*

3.2 Convex analysis

Most of the convex functions considered in this paper will be defined over spaces of measures or functions. Consequently, we will apply tools from convex analysis in its general formulation for locally convex topological vector spaces. References on this subject include Berg et al. (1984) and Bourbaki (1987, II. and IV.§1) for the topological background, and Ekeland and Témam (1999, Part I) and Zălinescu (2002, Chapters 1 & 2) for convex analysis. We now briefly review the main concepts appearing in the present paper.

Definition 3 (Dual pair) *A* dual pair *is a triplet $(X, Y, \langle\,\cdot\,,\,\cdot\,\rangle)$ where X and Y are real vector spaces, and $\langle\,\cdot\,,\,\cdot\,\rangle : X \times Y \to \mathbb{R}$ is a bilinear form satisfying the following properties:*

(i) for every $x \in X \setminus \{0\}$, there exists $y \in Y$ such that $\langle x, y\rangle \neq 0$.

(ii) for every $y \in Y \setminus \{0\}$, there exists $x \in X$ such that $\langle x, y\rangle \neq 0$.

We say that the pairing $\langle\,\cdot\,,\,\cdot\,\rangle$ puts X and Y in (separating) duality. *Furthermore, a topology τ on X is said to be* compatible *with the pairing if it is locally convex and if the topological dual $X^\star$ of X with respect to τ is isomorphic to Y. Topologies on Y compatible with the pairing are defined similarly.*

[3] Note that total variation distance is sometimes defined as half of this quantity, corresponding to functions taking values in $[0, 1]$.

Example 3 *For an arbitrary dual pair $(X, Y, \langle\,\cdot\,,\,\cdot\,\rangle)$, the* weak topology $\sigma(X, Y)$ *induced by Y on X is defined to be the coarsest topology such that for each $y \in Y$, $x \mapsto \langle x, y\rangle$ is a continuous linear form on X. It is a locally convex Hausdorff topology induced by the family of seminorms $p_y : x \mapsto |\langle x, y\rangle|$ for $y \in Y$ and is thereby compatible with the duality between X and Y.*

Note that in finite dimension, all Hausdorff vector space topologies coincide with the standard topology.

In the remainder of this section, we fix a dual pair $(X, Y, \langle\,\cdot\,,\,\cdot\,\rangle)$ and endow X and Y with topologies compatible with the pairing. As is customary in convex analysis, convex functions take values in the set of extended reals $\overline{\mathbb{R}} \coloneqq \mathbb{R} \cup \{-\infty, +\infty\}$ to which the addition over $\mathbb{R}$ is extended using the usual conventions, including $(+\infty) + (-\infty) = +\infty$. In this manner, convex functions can always be extended to be defined on the entirety of their domain by assuming the value $+\infty$ when they are not defined. For a convex function $f : X \to \overline{\mathbb{R}}$, $\operatorname{dom} f \coloneqq \{x \in X \mid f(x) < +\infty\}$ is the *effective domain* of f and $\partial f(x) \coloneqq \{y \in Y \mid \forall x' \in X, f(x') \geq f(x) + \langle x' - x, y\rangle\}$ denotes its subdifferential at $x \in X$.

Definition 4 (Lower semicontinuity, inf-compactness) *The function $f : X \to \overline{\mathbb{R}}$ is* lower semicontinuous (lsc) *(resp.* inf-compact*) if for every $t \in \mathbb{R}$ the sublevel set $f^{-1}(-\infty, t] \coloneqq \{x \in X \mid f(x) \leq t\}$ is closed (resp. compact).*

Lemma 5 *If $f : X \times C \to \overline{\mathbb{R}}$ is a convex function for C a convex subset of some linear space, then $g : X \to \overline{\mathbb{R}}$ defined as $g(x) \coloneqq \inf_{c \in C} f(x, c)$ is convex. Furthermore, if for some topology on C the function f is inf-compact with respect to the product topology, then g is also inf-compact.*

Definition 6 (Properness) *A convex function $f : X \to \overline{\mathbb{R}}$ is* proper *if $\operatorname{dom} f \neq \emptyset$ and $f(x) > -\infty$ for all $x \in X$.*

Definition 7 (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 for $y \in Y$ by*

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

For a set $C \subseteq X$, $\delta_C : X \to \overline{\mathbb{R}}_{\geq 0}$ denotes the characteristic function of C, that is $\delta_C(x)$ is 0 if $x \in C$ and $+\infty$ elsewhere. The support function of C is $h_C : Y \to \mathbb{R} \cup \{+\infty\}$ defined by $h_C(y) = \sup_{x \in C}\langle x, y\rangle$. If C is closed and convex then (δ_C, h_C) form a pair of convex conjugate functions.

Proposition 8 *Let $f : X \to \overline{\mathbb{R}}$ be a function. Then:*

1. *$f^\star : Y \to \overline{\mathbb{R}}$ is convex and lower semicontinuous.*
2. *for all $x \in X$ and $y \in Y$, $f(x) + f^\star(y) \geq \langle x, y\rangle$ with equality iff $y \in \partial f(x)$.*
3. *$f^{\star\star} \leq f$ with equality iff f is proper convex lower semicontinuous, $f \equiv +\infty$ or $f \equiv -\infty$.*
4. *if $f \leq g$ for some $g : X \to \overline{\mathbb{R}}$, then $g^\star \geq f^\star$.*

Remark 9 *In Proposition 8, Item 2 is known as the Fenchel–Young inequality and Item 3 as the Fenchel–Moreau theorem.*

In the special case of $X = \mathbb{R} = Y$ and a proper convex function $f : \mathbb{R} \to \overline{\mathbb{R}}$, we can be more explicit about some properties of $f^\star$ and $f^{\star\star}$.

Lemma 10 *If $f : \mathbb{R} \to \overline{\mathbb{R}}$ is a proper convex function, then $x \in \mathbb{R}$ is such that $f(x) \neq f^{\star\star}(x)$ only if $\operatorname{dom} f$ has non-empty interior and x is one of the (at most two) points on its boundary, in which case $f^{\star\star}(x)$ is the limit of $f(x')$ as $x' \to x$ within $\operatorname{dom} f$.*

Definition 11 *For $f : \mathbb{R} \to \overline{\mathbb{R}}$ a proper convex function, we define for $\ell \in \{-\infty, +\infty\}$ the quantity $f'(\ell) \coloneqq \lim_{x \to \ell} f(x)/x \in \mathbb{R} \cup \{+\infty\}$.*

Remark 12 *The limit is always well-defined in $\mathbb{R} \cup \{+\infty\}$ for proper convex functions. The name $f'(\ell)$ is motivated by the fact that when f is differentiable, we have $f'(\ell) = \lim_{x \to \ell} f'(x)$.*

Lemma 13 *If $f : \mathbb{R} \to \overline{\mathbb{R}}$ is a proper convex function, then the domain of $f^\star : \mathbb{R} \to \overline{\mathbb{R}}$ satisfies $\operatorname{int}(\operatorname{dom} f^\star) = \big(f'(-\infty), f'(+\infty)\big)$.*

Lemma 14 *Let $(f_i)_{i \in I}$ be a collection of convex functions from $\mathbb{R}$ to $\overline{\mathbb{R}}$ which are non-decreasing over some convex set $C \subseteq \mathbb{R}$. Then for all $x \in \operatorname{int} C$*

$$\lim_{x' \to x^-} \inf_{i \in I} f_i(x') \leq \inf_{i \in I} f_i^{\star\star}(x) \leq \inf_{i \in I} f_i(x)\,.$$

Proof For each $i \in I$ we have by Lemma 10 that $f_i^{\star\star}(x) \in \{f_i(x), \lim_{x' \to x^-} f_i(x')\}$, so since f_i is non-decreasing over C and $f_i^{\star\star} \leq f_i$ by Proposition 8, the result follows by taking the infimum over $i \in I$ as $\lim_{x' \to x^-} \inf_{i \in I} f_i(x') \leq \inf_{i \in I} \lim_{x' \to x^-} f_i(x')$. ■

Fenchel duality theorem is arguably the most fundamental result in convex analysis, and we will use it in this paper to compute the convex conjugate and minimum of a convex function subject to a linear constraint. The following proposition summarizes the conclusions obtained by instantiating the duality theorem to this specific case.

Proposition 15 *Let $f : X \to (-\infty, +\infty]$ be a convex function. For $y \in Y$ and $\varepsilon \in \mathbb{R}$, define $f_{y,\varepsilon} : X \to (-\infty, +\infty]$ by*

$$f_{y,\varepsilon}(x) \coloneqq f(x) + \delta_{\{\varepsilon\}}\big(\langle x, y\rangle\big) = \begin{cases} f(x) & \text{if } \langle x, y\rangle = \varepsilon \\ +\infty & \text{otherwise} \end{cases}$$

for all $x \in X$.

1. *Assume that f is lower semicontinuous and define $\langle \operatorname{dom} f, y\rangle \coloneqq \{\langle x, y\rangle \mid x \in \operatorname{dom} f\}$. If $\varepsilon \in \operatorname{int}\big(\langle \operatorname{dom} f, y\rangle\big)$, then $f^\star_{y,\varepsilon}(x^\star) = \inf_{\lambda \in \mathbb{R}} f^\star(x^\star + \lambda y) - \lambda \cdot \varepsilon$ for all $x^\star \in Y$, where the infimum is reached whenever $f^\star_{y,\varepsilon}(x^\star)$ is finite.*

2. *Assume that f is non-negative and satisfies $f(0) = 0$. Define the* marginal value function
$$\mathscr{L}_{y,f}(\varepsilon) := \inf_{x\in X} f_{y,\varepsilon}(x) = \inf\{f(x) \mid x \in X \,\wedge\, \langle x, y\rangle = \varepsilon\}\,. \tag{4}$$
Then $\mathscr{L}_{y,f}$ is a non-negative convex function satisfying $\mathscr{L}_{y,f}(0) = 0$ and its convex conjugate is given by $\mathscr{L}^\star_{y,f}(t) = f^\star(ty)$. Furthermore, $\mathscr{L}_{y,f}$ is lower semicontinuous at ε, that is $\mathscr{L}_{y,f}(\varepsilon) = \mathscr{L}^{\star\star}_{y,f}(\varepsilon)$, if and only if strong duality holds for problem (4)*, i.e. if and only if*
$$\inf\{f(x) \mid x \in X \,\wedge\, \langle x, y\rangle = \varepsilon\} = \sup\{t\cdot\varepsilon - f^\star(t\cdot y) \mid t \in \mathbb{R}\}\,.$$

Proof

1. This follows from a direct application of Fenchel's duality theorem (see e.g. Zălinescu (2002, Corollary 2.6.4, Theorem 2.8.1)).

2. Define the *perturbation function* $F : X \times \mathbb{R} \to \overline{\mathbb{R}}$ by $F(x, \varepsilon) := f_{y,\varepsilon}(x) = f(x) + \delta_{\{0\}}\big(\langle x, y\rangle - \varepsilon\big)$ so that $\mathscr{L}_{y,f}(\varepsilon) = \inf_{x\in X} F(x, \varepsilon)$. Since F is non-negative, jointly convex over the convex set $X \times \mathbb{R}$ and $F(0, 0) = 0$, we get that $\mathscr{L}_{y,f}$ is itself convex, non-negative, and satisfies $\mathscr{L}_{y,f}(0) = 0$. Furthermore, $F^\star(x^\star, t) = f^\star(x^\star + ty)$ and $\mathscr{L}^\star_{y,f}(t) = F^\star(0, t) = f^\star(ty)$ by e.g. Zălinescu (2002, Theorem 2.6.1, Corollary 2.6.4).

■

Finally, we will use the following result giving a sufficient condition for a convex function to be bounded below. Most such results in convex analysis assume that the function is either lower semicontinuous or bounded above on an open set. In contrast, the following lemma assumes that the function is upper bounded on a closed, convex, bounded set of a Banach space, or more generally on a *cs-compact* subset of a real Hausdorff topological vector space.

Lemma 16 (cf. König (1986, Example 1.6(0), Remark 1.9)) *Let C be a cs-compact subset of a real Hausdorff topological vector space. If $f : C \to \mathbb{R}$ is a convex function such that* $\sup_{x\in C} f(x) < +\infty$*, then* $\inf_{x\in C} f(x) > -\infty$*. In particular, if $f : C \to \mathbb{R}$ is linear, then* $\sup_{x\in C} f(x) < +\infty$ *if and only if* $\inf_{x\in C} f(x) > -\infty$*.*

The notion of cs-compactness (called σ-convexity in König (1986)) was introduced and defined in Jameson (1972), and Proposition 2 of the same paper states that closed, convex, bounded sets of Banach spaces are cs-compact. For completeness, we include a proof of Lemma 16 in Appendix A.1.

3.3 Orlicz spaces

We will use elementary facts from the theory of Orlicz spaces which we now briefly review (see for example Léonard (2007) for a concise exposition or Rao and Ren (1991) for a more complete reference). A function $\theta : \mathbb{R} \to [0, +\infty]$ is a *Young function* if it is a convex, lower semicontinuous, and even function with $\theta(0) = 0$ and $0 < \theta(s) < +\infty$ for some $s > 0$. Then writing $I_{\theta,\nu} : f \mapsto \int \theta(f)\,\mathrm{d}\nu$ for $\nu \in \mathcal{M}$, one defines[4] two spaces associated with θ:

- the Orlicz space $L^\theta(\nu) \coloneqq \big\{f \in L^0(\nu) \,\big|\, \exists\alpha > 0, I_{\theta,\nu}(\alpha f) < \infty\big\}$,
- the Orlicz heart (Edgar and Sucheston, 1989) $L^\theta_\heartsuit(\nu) \coloneqq \big\{f \in L^0(\nu) \,\big|\, \forall\alpha > 0, I_{\theta,\nu}(\alpha f) < \infty\big\}$, also known as the Morse–Transue space (Morse and Transue, 1950),

which are both Banach spaces when equipped with the Luxemburg norm $\|f\|_\theta \coloneqq \inf\{t > 0 \mid I_{\theta,\nu}(f/t) \leq 1\}$. Furthermore, $L^\theta_\heartsuit(\nu) \subseteq L^\theta(\nu) \subseteq L^1(\nu)$ and $L^\infty(\nu) \subseteq L^\theta(\nu)$ for all θ, and $L^\infty(\nu) \subseteq L^\theta_\heartsuit(\nu)$ when $\operatorname{dom}\theta = \mathbb{R}$. If $\theta^\star$ is the convex conjugate of θ, we have the following analogue of Hölder's inequality: $\int f_1 f_2\,\mathrm{d}\nu \leq 2\|f_1\|_\theta\|f_2\|_{\theta^\star}$, for all $f_1 \in L^\theta(\nu)$ and $f_2 \in L^{\theta^\star}(\nu)$, implying that $(L^\theta, L^{\theta^\star})$ are in dual pairing. Furthermore, if $\operatorname{dom}\theta = \mathbb{R}$, we have that the dual Banach space $(L^\theta_\heartsuit, \|\cdot\|_\theta)^\star$ is isomorphic to $(L^{\theta^\star}, \|\cdot\|_{\theta^\star})$.

4. Variational representations of ϕ-divergences

In the rest of this paper, we fix a convex and lower semicontinuous function $\phi : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$ such that $\phi(1) = 0$. After defining ϕ-divergences in Section 4.1, we start with the usual variational representation of the ϕ-divergence in Section 4.2, which we then strengthen in the case of probability measures in Section 4.3. A reader interested primarily in optimal bounds between ϕ-divergences and IPMs can skip Sections 4.2 and 4.3 at a first reading.

4.1 Convex integral functionals and ϕ-divergences

The notion of a ϕ-divergence is closely related to the one of a convex integral functional that we define first.

Definition 17 (Integral functional) *For $\nu \in \mathcal{M}^+$ and $f : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ a proper convex function, the convex integral functional associated with f and ν is the function $I_{f,\nu} : L^1(\nu) \to \mathbb{R} \cup \{\infty\}$ defined for $g \in L^1(\nu)$ by*

$$I_{f,\nu}(g) = \int f \circ g\,\mathrm{d}\nu\,.$$

The systematic study of convex integral functionals from the perspective of convex analysis was initiated by Rockafellar (1968, 1971), who considered more generally functionals of the form $g \mapsto \int f(\omega, g(\omega))\,\mathrm{d}\nu$ for $g : \Omega \to \mathbb{R}^n$ and $f : \Omega \times \mathbb{R}^n \to \mathbb{R}$ such that $f(\omega, \cdot)$ is convex ν-almost everywhere. A good introduction to the theory of such functionals can be found in Rockafellar (1976); Rockafellar and Wets (1998b). The specific case of Definition 17 is known as an *autonomous* integral functional, but we drop this qualifier since it applies to all functionals studied in this paper.

Definition 18 (ϕ-divergence) *For $\mu \in \mathcal{M}$ and $\nu \in \mathcal{M}^+$, write $\mu = \mu_c + \mu_s$ with $\mu_c \ll \nu$ and $\mu_s \perp \nu$, the Lebesgue decomposition of μ with respect to ν, and $\mu_s = \mu_s^+ - \mu_s^-$ with*

[4] The definition and theory of Orlicz spaces holds more generally for σ-finite measures. The case of finite measures already covers all the applications considered in this paper whose focus is primarily on probability measures.

Name	ϕ	$\phi'(\infty) < \infty$?	$\phi(0) < \infty$?	Notes
α-divergences	$\frac{x^\alpha - 1}{\alpha(\alpha-1)}$	when $\alpha < 1$	when $\alpha > 0$	$\phi_\alpha^\dagger = \phi_{1-\alpha}$
KL	$x \log x$	No	Yes	Limit of $\alpha \to 1^-$
reverse KL	$-\log x$	Yes	No	Limit of $\alpha \to 0^+$
squared Hellinger	$(\sqrt{x} - 1)^2$	Yes	Yes	Scaling of $\alpha = \frac{1}{2}$
χ^2-divergence	$(x-1)^2$	No	Yes	Scaling of $\alpha = 2$
Jeffreys	$(x-1)\log x$	No	No	KL + reverse KL
χ^α-divergences	$\lvert x-1 \rvert^\alpha$	when $\alpha = 1$	Yes	For $\alpha \geq 1$ (Vajda, 1973)
Total variation	$\lvert x - 1 \rvert$	Yes	Yes	χ^1-divergence
Jensen–Shannon	$x \log x - (1+x)\log\big(\frac{1+x}{2}\big)$	Yes	Yes	a.k.a. total divergence to the average
Triangular discrimination	$\frac{(x-1)^2}{x+1}$	Yes	Yes	a.k.a. Vincze–Le Cam distance

Table 1: Common ϕ-divergences (see e.g. Sason and Verdú (2016))

$\mu_s^+, \mu_s^- \in \mathcal{M}^+$*, the Hahn–Jordan decomposition of* μ_s*. The* ϕ-divergence *of* μ *with respect to* ν *is the quantity* $\mathrm{D}_\phi(\mu \parallel \nu) \in \mathbb{R} \cup \{\infty\}$ *defined by*

$$\mathrm{D}_\phi(\mu \parallel \nu) \coloneqq \int \phi\bigg(\frac{d\mu_c}{d\nu}\bigg)\,\mathrm{d}\nu + \mu_s^+(\Omega) \cdot \phi'(\infty) - \mu_s^-(\Omega) \cdot \phi'(-\infty)\,,$$

with the convention $0 \cdot (\pm\infty) = 0$.

Remark 19 *An equivalent definition of* $\mathrm{D}_\phi(\mu \parallel \nu)$ *which does not require decomposing* μ *is obtained by choosing* $\lambda \in \mathcal{M}^+$ *dominating both* μ *and* ν *(e.g.* $\lambda = |\mu| + \nu$*) and defining*

$$\mathrm{D}_\phi(\mu \parallel \nu) = \int \frac{d\nu}{d\lambda} \cdot \phi\bigg(\frac{d\mu/d\lambda}{d\nu/d\lambda}\bigg)\,\mathrm{d}\lambda,$$

with the conventions coming from continuous extension that $0 \cdot \phi(a/0) = a \cdot \phi'(\infty)$ *if* $a \geq 0$ *and* $0 \cdot \phi(a/0) = a \cdot \phi'(-\infty)$ *if* $a \leq 0$ *(see Definition 11). It is easy to check that this definition does not depend on the choice of* λ *and coincides with Definition 18.*

The notion of ϕ-divergence between probability measures was introduced by Csiszár (1963, 1967) in information theory and independently by Ali and Silvey (1966) in statistics. The generalization to finite signed measures is from Csiszár et al. (1999). Some useful properties of the ϕ-divergence include: it is jointly convex in both its arguments, if $\mu(\Omega) = \nu(\Omega)$ then $\mathrm{D}_\phi(\mu \parallel \nu) \geq 0$, with equality if and only if $\mu = \nu$ assuming that ϕ is strictly convex at 1.

Remark 20 *If* $\mu \ll \nu$*, the definition simplifies to* $\mathrm{D}_\phi(\mu \parallel \nu) = \nu\Big(\phi \circ \frac{d\mu}{d\nu}\Big)$*. Furthermore, if* $\phi'(\pm\infty) = \pm\infty$*, then* $\mathrm{D}_\phi(\mu \parallel \nu) = +\infty$ *whenever* $\mu \not\ll \nu$*. When either* $\phi'(+\infty)$ *or* $\phi'(-\infty)$ *is* *finite, some authors implicitly or explicitly redefine* $\mathrm{D}_\phi(\mu \parallel \nu)$ *to be* $+\infty$ *whenever* $\mu \not\ll \nu$*, thus departing from Definition 18. This effectively defines* $\mathrm{D}_\phi(\cdot \parallel \nu)$ *as the integral functional* $I_{\phi,\nu}$ *and the rich theory of convex integral functionals can be readily applied. As we will see in this paper, this change of definition is unnecessary and the difficulties arising from the case* $\mu \not\ll \nu$ *in Definition 18 can be addressed by separately treating the component of* μ *singular with respect to* ν*.*

An important reason to prefer the general definition is the equality $\mathrm{D}_\phi(\nu \parallel \mu) = \mathrm{D}_{\phi^\dagger}(\mu \parallel \nu)$ *where* $\phi^\dagger : x \mapsto x\phi(1/x)$ *is the Csiszár dual of* ϕ*, which identifies the* reverse *ϕ-divergence—where the arguments are swapped—with the divergence associated with* $\phi^\dagger$*. Consequently, any result obtained for the partial function* $\mu \mapsto \mathrm{D}_\phi(\mu \parallel \nu)$ *can be translated into results for the partial function* $\nu \mapsto \mathrm{D}_\phi(\mu \parallel \nu)$ *by swapping the role of* μ *and* ν *and replacing* ϕ *with* $\phi^\dagger$*. Note that* $(\phi^\dagger)'(\infty) = \lim_{x\to 0^+} \phi(x)$ *and* $(\phi^\dagger)'(-\infty) = \lim_{x\to 0^-} \phi(x)$*, and for many divergences of interest (including the Kullback–Leibler divergence) at least one of* $\phi'(\infty)$ *and* $\phi(0)$ *is finite. See Table 1 for some examples.*

4.2 Variational representations: general measures

In this section, we fix a finite and non-negative measure $\nu \in \mathcal{M}^+ \setminus \{0\}$ and study the convex functional $\mathrm{D}_{\phi,\nu} : \mu \mapsto \mathrm{D}_\phi(\mu \parallel \nu)$ over a vector space X of finite measures containing ν. Our primary goal is to derive a *variational representation* of $\mathrm{D}_{\phi,\nu}$, expressing it as the solution of an optimization problem over Y, a vector space of functions put in dual pairing with X via $\langle \mu, h\rangle = \mu(h)$, for a measure $\mu \in X$ and a function $h \in Y$.

Care must be taken in specifying the dual pair (X, Y), since the variational representation we obtain depends on it, or more precisely, on the null σ-ideal Ξ of $(\Omega, \mathcal{F})$ consisting of the measurable sets that are null for all measures in X. Note that, as discussed in Remark 20, this ideal Ξ is irrelevant when $\phi'(\pm\infty) = \pm\infty$ (e.g. when D_ϕ is the KL divergence), since then $\mathrm{D}_\phi(\mu \parallel \nu) = +\infty$ whenever $\mu \not\ll \nu$, but when, $\phi'(+\infty)$ or $\phi'(-\infty)$ is finite, it is possible to have $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$ even when $\mu \not\ll \nu$, and we wish to obtain a variational representation for such discontinuous measures as well. Denoting by N the σ-ideal of ν-null sets, we always have $\Xi \subseteq N$ since we assume that $\nu \in X$. If $\Xi = N$, then we have $X \subseteq \mathcal{M}_c(\nu)$, corresponding to case of only absolutely continuous measures, but if $\Xi \subset N$ is a proper subset of N then $N \setminus \Xi$ quantifies the "amount of ν-singularities" of measures in X. The extreme case where $\Xi = \{\emptyset\}$ allows for arbitrary singularities, since this implies that for any measurable set $A \in \mathcal{F}$, there exists a measure in X with positive variation on A. Furthermore, for common measurable spaces Ω, there is usually an ambient measure λ for which it is natural to assume that $X \subseteq \mathcal{M}_c(\lambda)$ (e.g. λ could be the Lebesgue measure on $\mathbb{R}$ or more generally the Haar measure on a locally compact unimodular group). Denoting by L the null σ-ideal of this ambient measure, we could then take $L \subseteq \Xi$, thus restricting the singularities of measures in X.

Formally, we require that the pair (X, Y) satisfies a *decomposability* condition that we define next. It is closely related to Rockafellar's notion of a decomposable space (Rockafellar, 1976, Section 3) which plays an important role in the theory of convex integral functionals.

Definition 21 (Decomposability) *Let* $X \subseteq \mathcal{M}(\Omega, \mathcal{F})$ *be a vector space of finite measures and define* $\Xi \coloneqq \{A \in \mathcal{F} \mid \forall \mu \in X,\ |\mu|(A) = 0\}$ *the* σ*-ideal of measurable sets that are null* *for all measures in X. Let Y be a vector space of measurable functions and let $\nu \in X$ be a finite non-negative measure. We say that the pair (X, Y) is ν-decomposable if:*

1. *the pairing $(\mu, h) \mapsto \int h \,\mathrm{d}\mu$ puts X and Y in separating duality.*
2. $\left\{\mu \in \mathcal{M}_c(\nu) \,\middle|\, \frac{d\mu}{d\nu} \in L^\infty(\nu)\right\} \subseteq X$ *and* $L^\infty(\Xi) \subseteq Y \subseteq L^0(\Xi)$.
3. *for all $A \notin \Xi$, there exists $\mu \in X^+ \setminus \{0\}$ such that $\mu(\Omega \setminus A) = 0$.*

Remark 22 *Note that items 2 and 3 together imply that the duality is necessarily separating, so the definition would remain identical by only requiring in item 1 that $\mu(h)$ be finite for each $\mu \in X$ and $h \in Y$. Furthermore, if we strengthen condition 2 by requiring that $\{\mu \in \mathcal{M}_c(\mu') \mid \frac{d\mu}{d\mu'} \in L^\infty(\mu')\} \subseteq X$ for all measures $\mu' \in X$, then 3 is implied. Thus, starting from an arbitrary dual pair (X, Y) in separating duality, one can extend X by taking its sum with the space of all measures of bounded derivative with respect to measures in X and extend Y by taking its sum with $L^\infty(\Xi)$. The resulting pair of extended spaces will then be decomposable with respect to any measure in X.*

Example 4 *If $X \subseteq \mathcal{M}_c(\nu)$, then item 2 implies item 3, and ν-decomposability then simply expresses that X and Y form a dual pair of decomposable spaces in the sense of Rockafellar (1976, Section 3), once $\mathcal{M}_c(\nu)$ is identified with $L^1(\nu)$ via the Radon–Nikodym theorem. An example of a ν-decomposable pair in this case is given by $\big(\mathcal{M}_c(\nu), L^\infty(\nu)\big)$. More generally, if Ξ is a proper subset of the σ-ideal of ν-null sets, then item 3 requires X to contain "sufficiently many" ν-singular measures. An example of a ν-decomposable pair for which $\Xi = \{\emptyset\}$ is given by $X = \mathcal{M}$ and $Y = \mathcal{L}^b(\Omega)$. An intermediate example which will be useful when considering IPMs can be obtained by constructing the largest dual pair (X, Y) such that Y contains a class of functions $\mathcal{G}$ of interest. The details of the construction are given in Definition 41 and decomposability is stated and proved in Lemma 43.*

With Definition 21 at hand, our approach to obtain variational representations of the divergence is simple. We first compute the convex conjugate $\mathrm{D}^\star_{\phi,\nu}$ of $\mathrm{D}_{\phi,\nu}$ defined for $h \in Y$ by

$$\mathrm{D}^\star_{\phi,\nu}(h) = \sup_{\mu \in X}\{\mu(h) - \mathrm{D}_{\phi,\nu}(\mu)\} \tag{5}$$

and prove that $\mathrm{D}_{\phi,\nu}$ is lower semicontinuous. By the Fenchel–Moreau theorem, we thus obtain the representation $\mathrm{D}_{\phi,\nu}(\mu) = \mathrm{D}^{\star\star}_{\phi,\nu}(\mu) = \sup_{h \in Y}\big\{\mu(h) - \mathrm{D}^\star_{\phi,\nu}(h)\big\}$.

We start with the simplest case where $X \subseteq \mathcal{M}_c(\nu)$, that is when all the measures in X are ν-absolutely continuous. Since $\mathrm{D}_{\phi,\nu}$ coincides with the integral functional $I_{\phi,\nu}$ in this case, this lets us exploit the well-known fact that under our decomposability condition, $(I_{\phi,\nu}, I_{\phi^\star,\nu})$ form a pair of convex conjugate functionals. This fact was first observed in Luxemburg and Zaanen (1956) in the context of Orlicz spaces, and then generalized in Rockafellar (1968, 1971).

Proposition 23 *Let $\nu \in \mathcal{M}^+$ be non-negative and finite, and let (X, Y) be ν-decomposable with $X \subseteq \mathcal{M}_c(\nu)$. Then the convex conjugate $\mathrm{D}^\star_{\phi,\nu}$ of $\mathrm{D}_{\phi,\nu}$ over X is given for all $h \in Y$ by*

$$\mathrm{D}^\star_{\phi,\nu}(h) = I_{\phi^\star,\nu}(h) = \int \phi^\star \circ h \,\mathrm{d}\nu.$$

Furthermore $\mathrm{D}_{\phi,\nu}$ *is lower semicontinuous, therefore for all* $\mu \in X$

$$\mathrm{D}_\phi(\mu \parallel \nu) = \sup_{h\in Y}\left\{\int h\,\mathrm{d}\mu - \int \phi^\star \circ h\,\mathrm{d}\nu\right\}. \tag{6}$$

Proof Since $\nu \in X$ by assumption, the function $\mathrm{D}_{\phi,\nu}$ is proper and convex over X. The proposition is then immediate consequence of Rockafellar (1976, Theorem 3C) after identifying $\mathcal{M}_c(\nu)$ with $L^1(\nu)$ by the Radon–Nikodym theorem and noting that X and Y are decomposable (Rockafellar, 1976, Section 3) by Assumption 21. ■

Example 5 *Consider the case of the Kullback–Leibler divergence, corresponding to the function* $\phi : x \mapsto x\log x$. *A simple computation gives* $\phi^\star(x) = e^{x-1}$ *and* (6) *yields as a variational representation, for all* $\mu \in X$

$$\mathrm{D}(\mu \parallel \nu) = \sup_{g\in Y}\left\{\mu(g) - \int e^{g-1}\,\mathrm{d}\nu\right\}, \tag{7}$$

Note that this representation differs from the Donsker–Varadhan representation (1) *discussed in the introduction. This discrepancy will be explained in the next section.*

The variational representation of the ϕ-divergence in Proposition 23 is well-known (see e.g. Ruderman et al. (2012)). However, as already discussed, the case where X contains ν-singular measures is also of interest and has been comparatively less studied in the literature. The following proposition generalizes the expression for $\mathrm{D}^\star_{\phi,\nu}$ obtained in Proposition 23 to the general case of an arbitrary ν-decomposable pair (X, Y), without requiring that $X \subseteq \mathcal{M}_c(\nu)$.

Proposition 24 *Let* $\nu \in \mathcal{M}^+$ *be a non-negative and finite measure and assume that* (X, Y) *is* ν*-decomposable. Then, the functional* $\mathrm{D}_{\phi,\nu}$ *over* X *has convex conjugate* $\mathrm{D}^\star_{\phi,\nu}$ *given for all* $g \in Y$ *by*

$$\mathrm{D}^\star_{\phi,\nu}(h) = \begin{cases} I_{\phi^\star,\nu}(h) & \textit{if } \operatorname{ess\,im}_\Xi(h) \subseteq [\phi'(-\infty), \phi'(\infty)] \\ +\infty & \textit{otherwise}\end{cases}, \tag{8}$$

where $\Xi \coloneqq \{A \in \mathcal{F} \mid \forall \mu \in X,\ |\mu|(A) = 0\}$ *is the null* σ*-ideal of* X.

Proof For $h \in Y$, let $C(h)$ be the right-hand side of Eq. (8), our claimed expression for $\mathrm{D}^\star_{\phi,\nu}(h)$.

First, we show that $\sup_{\mu\in X}\{\mu(h) - \mathrm{D}_{\phi,\nu}(\mu)\} \le C(h)$. For this, we assume that $\operatorname{ess\,im}_\Xi(h) \subseteq [\phi'(-\infty), \phi'(\infty)]$, as otherwise $C(h) = +\infty$ and there is nothing to prove. For $\mu \in X$, write $\mu = \mu_c + \mu_s^+ - \mu_s^-$ with $\mu_c \in \mathcal{M}_c(\nu)$ and $\mu_s^+, \mu_s^- \in \mathcal{M}_s^+(\nu)$, so that

$$\mu(h) - \mathrm{D}_{\phi,\nu}(\mu) = \mu_c(h) - I_{\phi,\nu}\left(\frac{d\mu_c}{d\nu}\right) + \mu_s^+(h) - \mu_s^+(\Omega)\cdot\phi'(\infty) - \mu_s^-(h) + \mu_s^-(\Omega)\cdot\phi'(-\infty). \tag{9}$$

Observe that $\mu_c(h) - I_{\phi,\nu}\Big(\frac{d\mu_c}{d\nu}\Big) = \nu\Big(\frac{d\mu_c}{d\nu}\cdot h - \phi \circ \frac{d\mu_c}{d\nu}\Big) \le \nu(\phi^\star \circ h) = I_{\phi^\star,\nu}(h)$, by the Fenchel–Young inequality applied to ϕ and monotonicity of the integral with respect to the non-negative measure ν. Since $\mu \ll \Xi$ by definition of Ξ and thus $\mu_s^+ \ll \Xi$, we have $\phi'(\infty) \geq \operatorname{ess\,sup}_\Xi h \geq \operatorname{ess\,sup}_{\mu_s^+} h$ so that $\mu_s^+(h) - \mu_s^+(\Omega) \cdot \phi'(\infty) = \mu_s^+(h - \phi'(\infty)) \leq 0$. Similarly $\mu_s^-(\Omega) \cdot \phi'(-\infty) - \mu_s^-(h) \leq 0$. Using these bounds in (9) yields $\mu(h) - \mathrm{D}_{\phi,\nu}(\mu) \leq C(h)$ as desired.

Next, we show that $\sup_{\mu \in X} \{\mu(h) - \mathrm{D}_{\phi,\nu}(\mu)\} \geq C(h)$. Observe that

$$\sup_{\mu \in X} \big\{\mu(h) - \mathrm{D}_{\phi,\nu}(\mu)\big\} \geq \sup_{\mu \in X_c(\nu)} \big\{\mu(h) - \mathrm{D}_{\phi,\nu}(\mu)\big\} = I_{\phi^\star,\nu}(h)\,, \tag{10}$$

where the equality follows from Proposition 23 applied to $X_\nu = X_c(\nu)$ and $Y_\nu = Y/{\sim_\nu}$ where $\sim_\nu$ is the equivalence relation of being equal ν-almost everywhere. If $\operatorname{ess\,im}_\Xi(h) \subseteq [\phi'(-\infty), \phi'(\infty)]$, then $I_{\phi^\star,\nu}(h) = C(h)$ and (10) gives the desired conclusion. If $\operatorname{ess\,sup}_\Xi h > \phi'(\infty)$, let $\alpha \in \mathbb{R}$ such that $\phi'(\infty) < \alpha < \operatorname{ess\,sup}_\Xi h$. Then $A = \{\omega \in \Omega \mid h(\omega) > \alpha\}$ is a measurable set in $\mathcal{F} \setminus \Xi$. If $\nu(A) > 0$, then $I_{\phi^\star,\nu}(h) = \infty = C(h)$, since $\operatorname{dom} \phi^\star \subseteq [\phi'(-\infty), \phi'(\infty)]$ and (10) again gives the desired conclusion. If $\nu(A) > 0$, then by Definition 21 there exists $\mu_A \in X^+ \setminus \{0\}$ such that $\mu_A(\Omega \setminus A) = 0$. But then

$$\begin{aligned}
\sup_{\mu \in X} \big\{\mu(h) - \mathrm{D}_{\phi,\nu}(\mu)\big\} &\geq \sup_{c>0} \big\{(\nu + c\mu_A)(h) - \mathrm{D}_{\phi,\nu}(\nu + c\mu_A)\big\} \\
&= \nu(h) + \sup_{c>0} \big\{c\mu_A(h) - c\mu_A(\Omega) \cdot \phi'(\infty)\big\} \\
&\geq \nu(h) + \sup_{c>0} \big\{c\mu_A(\Omega) \cdot \big(\alpha - \phi'(\infty)\big)\big\} = +\infty = C(h)\,,
\end{aligned}$$

where the first equality is because $I_{\phi,\nu}\big(\frac{d\nu}{d\nu}\big) = \phi(1) = 0$ and $\mu_A \in X_s^+(\nu)$, and the second is because $\mu_A(\Omega) > 0$ and $\alpha > \phi'(\infty)$. The case $\operatorname{ess\,inf}_\Xi h(\Omega) < \phi'(-\infty)$ is analogous. ■

Remark 25 *Although the expression of $\mathrm{D}^\star_{\phi,\nu}$ obtained in Proposition 24 should coincide with the one obtained in Proposition 23 when $X \subseteq \mathcal{M}_c(\nu)$ (in which case Ξ coincides with the σ-ideal of ν-null sets), it appears different at first glance because of the explicit constraint on the Ξ-essential range of g present in (8). However, this constraint is also present, though implicit, in Proposition 23 since $\overline{\operatorname{dom} \phi^\star} = [\phi'(-\infty), \phi'(\infty)]$ and thus $I_{\phi^\star,\nu}(h) = +\infty$ whenever $\operatorname{ess\,im}_\nu(h) \not\subseteq [\phi'(-\infty), \phi(\infty)]$. When X is allowed to contain measures which are not absolutely continuous with respect to ν, this implicit constraint on the ν-essential range is simply strengthened to restrict the Ξ-essential range instead. In the extreme case where $\Xi = \{\emptyset\}$ then the true range of h is constrained.*

Finally, we prove that $\mathrm{D}_{\phi,\nu}$ is lower semicontinuous over X, yielding a variational representation of $\mathrm{D}_\phi(\mu \parallel \nu)$ in the general case.

Proposition 26 *Let $\nu \in \mathcal{M}^+$ be a non-negative and finite measure and assume that (X, Y) is ν-decomposable. Then, $\mathrm{D}_{\phi,\nu}$ is lower semicontinuous over X. Equivalently, we have for all $\mu \in X$ the biconjugate representation*

$$\mathrm{D}_\phi(\mu \parallel \nu) = \sup\big\{\mu(g) - I_{\phi^\star,\nu}(g) \;\big|\; g \in Y \,\wedge\, \operatorname{ess\,im}_\Xi(g) \subseteq [\phi'(-\infty), \phi'(\infty)]\big\}\,,$$

where $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X,\ |\mu|(A) = 0\}$ is the null σ-ideal of X.

Proof Since $\mathrm{D}_{\phi,\nu}$ is proper, by the Fenchel–Moreau theorem it suffices to show that $\mathrm{D}^{\star\star}_{\phi,\nu} \geq \mathrm{D}_{\phi,\nu}$. For $\mu \in X$, write $\mu = \mu_c + \mu_s^+ - \mu_s^-$ with $\mu_c \in \mathcal{M}_c(\nu)$, and $\mu_s^+, \mu_s^- \in \mathcal{M}_s^+(\nu)$ by the Lebesgue and Hahn–Jordan decompositions. Furthermore, let $(C, P, N) \in \mathcal{F}^3$ be a partition of Ω such that $|\mu_c|(\Omega \setminus C) = \nu(\Omega \setminus C) = 0$, $\mu_s^+(\Omega \setminus P) = 0$ and $\mu_s^-(\Omega \setminus N) = 0$. By Proposition 24,

$$\begin{aligned}\mathrm{D}^{\star\star}_{\phi,\nu}(\mu) = \sup\big\{&\mu_c(g) - I_{\phi^\star,\nu}(g) + \mu_s^+(g) - \mu_s^-(g) \\ &\big|\, g \in Y \,\wedge\, \operatorname{ess\,im}_\Xi(g) \subseteq [\phi'(-\infty), \phi'(\infty)]\big\}\,.\end{aligned} \tag{11}$$

Let $\alpha \in \mathbb{R}$ such that $\alpha < I_{\phi,\nu}\Big(\frac{d\mu_c}{d\nu}\Big)$. Applying Proposition 23 with $X_\nu = \mathcal{M}_c(\nu)$ and $Y_\nu = L^\infty(\nu)$, we get the existence of $g_c \in L^\infty(\nu)$ such that $\mu_c(g_c) - I_{\phi^\star,\nu}(g_c) > \alpha$. Furthermore, since $\operatorname{dom}\phi^\star \subseteq [\phi'(-\infty), \phi'(\infty)]$, we have that $g_c \in [\phi'(-\infty), \phi'(\infty)]$ ν-almost everywhere. Consequently, there exists a representative $\tilde{g}_c \in \mathcal{L}^b(\Omega)$ of g_c such that $\tilde{g}_c(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)]$.

For $\beta, \gamma \in \mathbb{R} \cap [\phi'(-\infty), \phi'(\infty)]$ (which is nonempty since it contains $\operatorname{dom}\phi^\star$ and ϕ is convex and proper), define $\tilde{g} : \Omega \to \mathbb{R}$ by

$$\tilde{g}(\omega) = \begin{cases} \tilde{g}_c(\omega) & \text{if } \omega \in C \\ \beta & \text{if } \omega \in P \\ \gamma & \text{if } \omega \in N \end{cases}.$$

By construction $\tilde{g} \in \mathcal{L}^b(\Omega)$, hence its equivalence class g in $L^\infty(\Xi)$ belongs to Y by Definition 21. Furthermore, since $\mu \ll \Xi$ we have $\mu_c(g) - I_{\phi^\star,\nu}(g) = \mu_c(\tilde{g}_c) - I_{\phi^\star,\nu}(\tilde{g}_c) = \mu_c(g_c) - I_{\phi^\star,\nu}(g_c) > \alpha$, $\mu_s^+(g) = \mu_s^+(\Omega) \cdot \beta$, and $\mu_s^-(g) = \mu_s^-(\Omega) \cdot \gamma$. Since $\tilde{g}(\Omega) \subseteq [\phi'(-\infty), \phi'(\infty)]$ by construction, for this choice of $g \in Y$, the optimand in (11) is at least $\alpha + \mu_s^+(\Omega) \cdot \beta - \mu_s^-(\Omega) \cdot \gamma$. This concludes the proof since α, β, γ can be made arbitrarily close to $I_{\phi,\nu}\Big(\frac{d\mu_c}{d\nu}\Big)$, $\phi'(\infty)$, and $\phi'(-\infty)$ respectively. ■

4.3 Variational representations: probability measures

When applied to *probability measures*, which are the main focus of this paper, the variational representations provided by Propositions 23 and 26 are loose. This fact was first explicitly mentioned in Ruderman et al. (2012), where the authors also suggested that tighter representations could be obtained by specializing the derivation to probability measures.

Specifically, given a dual pair (X, Y) as in Section 4.2, we restrict $\mathrm{D}_{\phi,\nu}$ to probability measures by defining $\widetilde{\mathrm{D}}_{\phi,\nu} : \mu \mapsto \mathrm{D}_{\phi,\nu}(\mu) + \delta_{\mathcal{M}^1}(\mu)$ for $\mu \in X$. For $g \in Y$ we get

$$\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) = \sup_{\mu \in X}\big\{\mu(g) - \widetilde{\mathrm{D}}_{\phi,\nu}(\mu)\big\} = \sup_{\mu \in X^1}\big\{\mu(g) - \mathrm{D}_{\phi,\nu}(\mu)\big\}\,. \tag{12}$$

Observe that compared to (5), the supremum is now taken over the smaller set $X^1 = X \cap \mathcal{M}^1$, and thus $\widetilde{\mathrm{D}}^\star_{\phi,\nu} \leq \mathrm{D}^\star_{\phi,\nu}$. When $\widetilde{\mathrm{D}}_{\phi,\nu}$ is lower semicontinuous we then get for $\mu \in X^1$

$$\mathrm{D}_\phi(\mu \parallel \nu) = \widetilde{\mathrm{D}}_{\phi,\nu}(\mu) = \widetilde{\mathrm{D}}^{\star\star}_{\phi,\nu}(\mu) = \sup_{g \in Y}\big\{\mu(g) - \widetilde{\mathrm{D}}^\star_{\phi,\nu}(g)\big\}\,. \tag{13}$$

This representation should be contrasted with the one obtained in Section 4.2, $\mathrm{D}_\phi(\mu \parallel \nu) = \sup_{g\in Y}\big\{\mu(g) - \mathrm{D}^\star_{\phi,\nu}(g)\big\}$, which holds for any $\mu \in X$ and in which the optimand is smaller than in (13) for all $g \in Y$ (see also Examples 6 and 7 below for an illustration).

In the rest of this section, we carry out the above program by giving an explicit expression for $\widetilde{\mathrm{D}}^\star_{\phi,\nu}$ defined in (12) and showing that $\widetilde{\mathrm{D}}_{\phi,\nu}$ is lower semi-continuous. We will assume in the rest of this paper that $\operatorname{dom}\phi$ contains a neighborhood of 1, as otherwise the ϕ-divergence on probability measures becomes the discrete divergence $\mathrm{D}_\phi(\mu \parallel \nu) = \delta_{\{\nu\}}(\mu)$ which is only finite when $\mu = \nu$ and for which the questions studied in this work are trivial. We start with the following lemma giving a simpler expression for $\widetilde{\mathrm{D}}_{\phi,\nu}$.

Lemma 27 *Define $\phi_+ : x \mapsto \phi(x) + \delta_{\mathbb{R}_{\geq 0}}(x)$ for $x \in \mathbb{R}$. Then for all $\mu \in \mathcal{M}$*

$$\widetilde{\mathrm{D}}_{\phi,\nu}(\mu) = \mathrm{D}_{\phi_+,\nu}(\mu) + \delta_{\{1\}}\big(\mu(\Omega)\big) .$$

Proof Using the same notations as in Definition 18, and since $\phi'_+(-\infty) = -\infty$, it is easy to see that $\mathrm{D}_{\phi_+,\nu}(\mu)$ equals $+\infty$ whenever $\mu_s^- \neq 0$ or $\nu\big(\{\omega \in \Omega \mid \frac{d\mu_c}{d\nu}(\omega) < 0\}\big) \neq 0$ and equals $\mathrm{D}_{\phi,\nu}(\mu)$ otherwise. In other words, $\mathrm{D}_{\phi_+,\nu}(\mu) = \mathrm{D}_{\phi,\nu}(\mu) + \delta_{\mathcal{M}^+}(\mu)$. This concludes the proof since $\delta_{\mathcal{M}^+}(\mu) + \delta_{\{1\}}\big(\mu(\Omega)\big) = \delta_{\mathcal{M}^1}(\mu)$. ■

In the expression of $\widetilde{\mathrm{D}}_{\phi,\nu}$ given by Lemma 27, the non-negativity constraint on μ is "encoded" directly in the definition of ϕ_+ (cf. Borwein and Lewis (1991)), only leaving the constraint $\mu(\Omega) = 1$ explicit. Since $\mu(\Omega) = \int \mathbf{1}_\Omega \,\mathrm{d}\mu$, this is an affine constraint which is well-suited to a convex duality treatment. In particular, we can use Proposition 15 to compute $\widetilde{\mathrm{D}}^\star_{\phi,\nu}$.

Proposition 28 *Assume that (X, Y) is a ν-decomposable dual pair for some $\nu \in \mathcal{M}^1$. Then the convex conjugate of $\widetilde{\mathrm{D}}_{\phi,\nu}$ with respect to (X, Y) is given for all $g \in Y$, by*

$$\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) = \inf\left\{\int \phi^\star_+(g+\lambda)\,\mathrm{d}\nu - \lambda \;\middle|\; \lambda + \operatorname{ess\,sup}_\Xi g \leq \phi'(\infty)\right\}, \tag{14}$$

where $\phi_+ : x \mapsto \phi(x) + \delta_{\mathbb{R}_{\geq 0}}(x)$ and $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X,\ |\mu|(A) = 0\}$.

In (14) the infimum is reached if it is finite, which holds in particular whenever $g \in L^\infty(\Xi)$.

Proof We use Lemma 27 and apply Proposition 15 with $f = \mathrm{D}_{\phi_+,\nu}$, $y = \mathbf{1}_\Omega$ and $\varepsilon = 1$. We need to verify that $1 \in \operatorname{int}\big(\{\mu(\mathbf{1}_\Omega) \mid \mu \in \operatorname{dom}\mathrm{D}_{\phi_+,\nu}\}\big)$, but this is immediate since $(1 \pm \alpha)\nu \in \operatorname{dom}\mathrm{D}_{\phi_+,\nu}$ for sufficiently small α by the assumption that $1 \in \operatorname{int}\operatorname{dom}\phi$.

Thus, by Proposition 15, for all $g \in Y$

$$\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) = \inf_{\lambda\in\mathbb{R}}\Big\{\mathrm{D}^\star_{\phi_+,\nu}(g+\lambda) - \lambda\Big\} ,$$

where the infimum is reached whenever it is finite. Equation (14) follows by using Proposition 24 and observing that $\phi'_+(\infty) = \phi'(\infty)$ and $\phi'_+(-\infty) = -\infty$.

It remains to verify the claims about finiteness of $\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g)$. For $g \in L^\infty(\Xi)$, write $M := \operatorname{ess\,sup}_\Xi g$. Since $\operatorname{int}(\operatorname{dom}\phi^\star_+) = \big(-\infty, \phi'(\infty)\big)$, for any $A < \phi'(\infty)$, the choice of $\lambda = A - M$ makes the optimand in (14) finite. ■

Remark 29 *As in Remark 25 above, when $X \subseteq \mathcal{M}_c(\nu)$ the constraint on λ in (14) can be dropped, leading to a simpler expression for $\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g)$ in this case. Indeed, $\overline{\operatorname{dom}\phi^\star_+} = \big(-\infty, \phi'(\infty)\big]$ and thus the optimand in (14) equals $+\infty$ whenever $\operatorname{ess\,sup}_\Xi g = \operatorname{ess\,sup}_\nu g > \phi'(\infty) - \lambda$.*

Example 6 *The effect of the restriction to probability measures is particularly pronounced for the total variation distance, which is the ϕ-divergence for $\phi(x) = |x-1|$. In the unrestricted case, a simple calculation shows ϕ has convex conjugate $\phi^\star(x) = x + \delta_{[-1,1]}(x)$, so that the conjugate of the unrestricted divergence $\mathrm{D}^\star_{\phi,\nu}(g)$ is $+\infty$ unless $\operatorname{ess\,im}_\Xi(g) \subseteq [-1,1]$. In the case of probability measures, the restriction ϕ_+ of ϕ to the non-negative reals has conjugate $\phi^\star_+(x) = x$ when $|x| \le 1$, $\phi^\star_+(x) = +\infty$ when $x > 1$, but $\phi^\star_+(x) = -1$ when $x < -1$. Thus, $\mathrm{D}^\star_{\phi_+,\nu}(g) < +\infty$ whenever $\operatorname{ess\,im}_\Xi(g) \subseteq (-\infty, 1]$. Furthermore, because of the additive λ shift in Eq. (14), we have $\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) < +\infty$ whenever $\operatorname{ess\,sup}_\Xi g < +\infty$, in particular whenever $g \in L^\infty(\Xi)$.*

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 30 *Assume that (X,Y) is ν-decomposable for some $\nu \in \mathcal{M}^1$. Then, $\widetilde{\mathrm{D}}_{\phi,\nu}$ is lower semicontinuous over X. In particular for all probability measures $\mu \in X^1 = X \cap \mathcal{M}^1$*

$$\mathrm{D}_\phi(\mu \parallel \nu) = \sup_{g\in Y}\Big\{\mu(g) - \inf\big\{I_{\phi^\star_+,\nu}(g+\lambda) - \lambda \;\big|\; \lambda + \operatorname{ess\,sup}_\Xi g \le \phi'(\infty)\big\}\Big\},$$

where $\phi_+ : x \mapsto \phi(x) + \delta_{\mathbb{R}_{\ge 0}}(x)$ and $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X,\ |\mu|(A) = 0\}$.

Proof Since $\mathbf{1}_\Omega \in Y$ the linear form $\mu \mapsto \mu(\mathbf{1}_\Omega)$ is continuous for any topology compatible with the dual pair (X,Y). Consequently, the function $\mu \mapsto \delta_{\{1\}}\big(\mu(\Omega)\big)$ is lower semicontinuous as the composition of the lower semicontinuous function $\delta_{\{1\}}$ with a continuous function. Finally, $\mathrm{D}_{\phi_+,\nu}$ is lower semicontinuous by Propositions 23 and 26. Hence $\widetilde{\mathrm{D}}_{\phi,\nu}$ is lower semicontinuous as the sum of two lower semicontinuous functions, by using the expression in Lemma 27. The variational representation immediately follows by expressing $\widetilde{\mathrm{D}}_{\phi,\nu}$ as its biconjugate. ■

Example 7 *As in Example 5, we consider the case of the Kullback–Leibler divergence, given by $\phi(x) = \phi_+(x) = x\log x$. For a ν-decomposable dual pair (X,Y), since $\phi^\star(x) = e^{x-1}$ Proposition 28 implies for $\nu \in \mathcal{M}^1$ and $g \in Y$ that*

$$\widetilde{\mathrm{D}}^\star_{\phi,\nu}(g) = \inf_{\lambda\in\mathbb{R}} \int e^{g+\lambda-1}\,\mathrm{d}\nu - \lambda = \log \int e^g\,\mathrm{d}\nu\,,$$

where the last equality comes from the optimal choice of $\lambda = -\log\int e^{g-1}\,\mathrm{d}\nu$. Using Corollary 30 we obtain for all probability measure $\mu \in X^1$

$$\mathrm{D}(\mu \parallel \nu) = \sup_{g\in Y}\left\{\mu(g) - \log\int e^g\,\mathrm{d}\nu\right\} = \sup_{g\in Y}\left\{\mu(g) - \nu(g) - \log\int e^{\left(g-\nu(g)\right)}\,\mathrm{d}\nu\right\},$$

which is the Donsker–Varadhan representation of the Kullback–Leibler divergence (Donsker and Varadhan, 1976). For $\mu \in X^1$, the variational representation obtained in (7) *can be equivalently written*

$$\mathrm{D}(\mu \parallel \nu) = \sup_{g \in Y} \left\{ 1 + \mu(g) - \int e^g \, \mathrm{d}\nu \right\} .$$

Using the inequality $\log(x) \leq x - 1$ *for* $x > 0$*, we see that the optimand in the previous supremum is smaller than the optimand in the Donsker–Varadhan representation for all* $g \in Y$*. We thus obtained a "tighter" representation by restricting the divergence to probability measures.*

Example 8 *Consider the family of divergences* $\phi(x) = |x-1|^\alpha/\alpha$ *for* $\alpha \geq 1$*. A simple computation gives* $\phi^\star(y) = y + |y|^\beta/\beta$ *where* $\beta \geq 1$ *is such that* $\frac{1}{\alpha} + \frac{1}{\beta} = 1$*. Jiao et al. (2017) uses the variational representation given by Proposition 23, that is* $\mathrm{D}_\phi(\mu \parallel \nu) = \sup_g \mu(g) - \nu(\phi^\star(g))$*. However, Corollary 30 shows that the tight representation uses* $\phi_+^\star(y)$ *which has the piecewise definition* $y + |y|^\beta/\beta$ *when* $y \geq -1$ *and the constant* $-1/\alpha$ *when* $y \leq -1$*, and writes* $\mathrm{D}_\phi(\mu \parallel \nu) = \sup_g \mu(g) - \inf_\lambda \nu\big(\phi_+^\star(g+\lambda)\big)$*. Note that the additive* λ *shift, in e.g. the case* $\alpha = 2$*, reduces the second term from the raw second moment* $\nu\big(g^2\big)$ *to something no larger than the variance* $\nu\big((g-\nu(g))^2\big)$*, which is potentially much smaller.*

5. Optimal bounds for a single function and reference measure

As a first step to understand the relationship between a ϕ-divergence and an IPM, we consider the case of a single fixed probability measure $\nu \in \mathcal{M}^1$ and measurable function $g \in \mathcal{L}^0$, and study the optimal lower bound of $\mathrm{D}_\phi(\mu \parallel \nu)$ as a function of the *mean deviation* $\mu(g) - \nu(g)$. We characterize this optimal lower bound and its convex conjugate in Section 5.1 and then present implications for topological question regarding the divergence itself in subsequent sections.

In the remainder of this work, since we are interested in probability measures, which are in particular non-negative, we assume without loss of generality that ϕ is infinite on the negative reals, that is $\phi(x) = \phi_+(x) = \phi(x) + \delta_{\mathbb{R}_{\geq 0}}(x)$. As seen in Section 4.3 (in particular Lemma 27), this does not change the value of the divergence on non-negative measures, that is $\mathrm{D}_\phi(\mu \parallel \nu) = \mathrm{D}_{\phi_+}(\mu \parallel \nu)$ for $\mu \in \mathcal{M}^+$, but yields a tighter variational representation since $\phi_+^\star \leq \phi^\star$.

Furthermore, since for probability measures $\mathrm{D}_\phi(\mu \parallel \nu)$ is invariant to affine shifts of the form $\widetilde{\phi}(x) = \phi(x) + c \cdot (x-1)$ for $c \in \mathbb{R}$, it will be convenient to assume that $0 \in \partial\phi(1)$ (e.g. $\phi'(1) = 0$), equivalently that ϕ is non-negative and has global minimum at $\phi(1) = 0$. This can always be achieved by an appropriate choice of c and is therefore without loss of generality. As an example, we now write for the Kullback–Leibler divergence $\phi(x) = x \log x - x + 1$ which is non-negative with $\phi'(1) = 0$, and equivalent to the standard definition $\phi(x) = x \log x$ for probability measures.

5.1 Derivation of the bound

We first define the optimal lower bound function, which comes in two flavors depending on whether the mean deviation or the absolute mean deviation is considered.

Definition 31 *For a probability measure $\nu \in \mathcal{M}^1$, a function $g \in \mathcal{L}^1(\nu)$, and set of probability measures M integrating g, the* optimal lower bound on $\mathrm{D}_\phi(\mu \parallel \nu)$ in terms of the mean deviation *is the function $\mathscr{L}_{g,\nu,M}$ defined for $\varepsilon \in \mathbb{R}$ by:*

$$\begin{aligned}\mathscr{L}_{g,\nu,M}(\varepsilon) &\coloneqq \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \;\Big|\; \mu \in M \,\wedge\, \mu(g) - \nu(g) = \varepsilon\Big\} \\ &= \inf_{\mu \in M}\big\{\mathrm{D}_\phi(\mu \parallel \nu) + \delta_{\{0\}}(\mu(g) - \nu(g) - \varepsilon)\big\} \qquad (15) \\ \mathscr{L}_{\{\pm g\},\nu,M}(\varepsilon) &\coloneqq \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \;\Big|\; \mu \in M \,\wedge\, |\mu(g) - \nu(g)| = \varepsilon\Big\} \\ &= \min\{\mathscr{L}_{g,\nu,M}(\varepsilon), \mathscr{L}_{g,\nu,M}(-\varepsilon)\} \qquad (16)\end{aligned}$$

where we follow the standard convention that the infimum of the empty set is $+\infty$.

Lemma 32 *For every $\nu \in \mathcal{M}^1$, $g \in \mathcal{L}^1(\nu)$, and convex set M of probability measures integrating g, the function $\mathscr{L}_{g,\nu,M}$ is convex and non-negative. Furthermore, $\mathscr{L}_{g,\nu,M}(0) = 0$ whenever $\nu \in M$, and if $\phi'(\infty) = \infty$ then $\mathscr{L}_{g,\nu,M} = \mathscr{L}_{g,\nu,M \cap \mathcal{M}_c(\nu)}$.*

Proof Convexity is immediate from Lemma 5 applied to Eq. (15), non-negativity follows from non-negativity of $\mathrm{D}_\phi(\cdot \parallel \nu)$, the choice $\mu = \nu$ implies $\mathscr{L}_{g,\nu,M}(0) = 0$ when $\nu \in M$, and if $\phi'(\infty) = \infty$ then $\mathrm{D}_\phi(\mu \parallel \nu) = +\infty$ when $\mu \in M \setminus \mathcal{M}_c(\nu)$. ■

We compute the convex conjugate of $\mathscr{L}_{g,\nu}$ by applying Fenchel duality to Eq. (15).

Proposition 33 *Let (X, Y) be a ν-decomposable pair for some probability measure $\nu \in \mathcal{M}^1$ and let $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X,\ |\mu|(A) = 0\}$. Then for all $g \in Y$ and $t \in \mathbb{R}$,*

$$\mathscr{L}^\star_{g,\nu,X^1}(t) = \inf\left\{\int \phi^\star(tg + \lambda)\,\mathrm{d}\nu - t \cdot \nu(g) - \lambda \;\middle|\; \lambda + \operatorname{ess\,sup}_\Xi(t \cdot g) \le \phi'(\infty)\right\}. \qquad (17)$$

Furthermore, $\mathscr{L}_{g,\nu,X^1}(\varepsilon) = \mathscr{L}^{\star\star}_{g,\nu,X^1}(\varepsilon)$ if and only if strong duality holds in Eq. (15).

Proof Define $\Phi : X \to \overline{\mathbb{R}}$ by $\Phi(x) = \widetilde{\mathrm{D}}_{\phi,\nu}(x + \nu)$ so that Φ is convex, lsc, non-negative, and 0 at 0. Furthermore, $\Phi^\star(h) = \widetilde{\mathrm{D}}^\star_{\phi,\nu}(h) - \nu(h)$ for $h \in Y$, and $\mathscr{L}_{g,\nu,X^1}(\varepsilon) = \inf\{\Phi(x) \mid x \in X \wedge \langle x, g\rangle = \varepsilon\}$. The result then follows by Propositions 15 and 28. ■

Remark 34 *Since $\operatorname{dom}\phi^\star \subseteq [-\infty, \phi'(\infty)]$, λ is always implicitly restricted in Eq. (17) to satisfy $\lambda + \operatorname{ess\,sup}_\nu tg \le \phi'(\infty)$. When Ξ is a proper subset of the null σ-ideal of ν, the constraint in Eq. (17) is stronger to account for measures in X which are not continuous with respect to ν.*

If $\phi'(\infty) = \infty$, then the infimum in Eq. (17) is taken over all $\lambda \in \mathbb{R}$ and in particular, does not depend on Ξ. This is consistent with the fact that, in this case, $\mathrm{D}_{\phi,\nu}$ is infinite on singular measures, hence $\mathscr{L}_{g,\nu,X^1} = \mathscr{L}_{g,\nu,X^1_c(\nu)}$ where $X_c(\nu) = X \cap \mathcal{M}_c(\nu)$.

Remark 35 *Unlike in Proposition 28, it is not always true that the interiority constraint qualification conditions hold, and indeed strong duality does not always hold for the optimization problem (15). For example, for $\Omega = (-1/2, 1/2)$, ν the Lebesgue measure, g the canonical* *injection into* $\mathbb{R}$, *and* $\phi : x \mapsto |x-1|$ *corresponding to the total variation distance, we have* $\mathscr{L}_{g,\nu,\mathcal{M}^1}(\pm 1/2) = \infty$ *but* $\mathscr{L}_{g,\nu,\mathcal{M}^1}(x) \leq 2$ *for* $|x| < 1/2$*. However, as noted in Theorem 40 below, this generally does not matter since it only affects the boundary of the domain of* $\mathscr{L}_{g,\nu}$*, which contains at most two points. Furthermore, we will show in Corollary 62 via a compactness argument that when* $\phi'(\infty) = \infty$ *and* $\operatorname{dom} \mathscr{L}^\star_{g,\nu} = \mathbb{R}$*—e.g. when* $g \in L^\infty(\nu)$*—strong duality holds in* (15)*.*

We can simplify the expressions in Proposition 33 by introducing the function $\psi : x \mapsto \phi(x+1)$. We state some useful properties of its conjugate $\psi^\star$ below.

Lemma 36 *The function* $\psi^\star : x \mapsto \phi^\star(x) - x$ *is non-negative, convex, and inf-compact. Furthermore, it satisfies* $\psi^\star(0) = 0$*,* $\psi^\star(x) \leq -x$ *when* $x \leq 0$*, and* $\operatorname{int}(\operatorname{dom}\psi^\star) = \big(-\infty, \phi'(\infty)\big)$*.*

Recall that at the beginning of Section 5 we assumed, without loss of generality, that $0 \in \partial\phi(1)$ and $\operatorname{dom}\phi \subseteq \mathbb{R}_{\geq 0}$, which is necessary for Lemma 36 to hold. The proof follows immediately from basic results in convex analysis on $\mathbb{R}$; for completeness, a proof is included in Appendix A.2.

The right-hand side of Eq. (17), expressed in terms of $\psi^\star$, will be central to our theory, so we give it a name in the following definition.

Definition 37 (Cumulant generating function) *For a* σ*-ideal* Ξ *and probability measure* $\nu \in \mathcal{M}^1_c(\Xi)$*, the* (ϕ,ν,Ξ)-cumulant generating function $K_{g,\nu,\Xi} : \mathbb{R} \to \overline{\mathbb{R}}$ *of a function* $g \in L^0(\Xi)$ *is defined for all* $t \in \mathbb{R}$ *by*

$$K_{g,\nu,\Xi}(t) \coloneqq \inf\left\{ \int \psi^\star(tg+\lambda)\,\mathrm{d}\nu \;\middle|\; \lambda + \operatorname{ess\,sup}_\Xi\big(t\cdot g\big) \leq \phi'(\infty) \right\}. \tag{18}$$

Note that since $\nu \in \mathcal{M}_c(\Xi)$*, we always have* $\Xi \subseteq N \coloneqq \{A \in \mathcal{F} \mid \nu(A) = 0\}$*, hence* $K_{g,\nu,\Xi} \geq K_{g,\nu,N}$*. In the common case where* $\Xi = N$ *we abbreviate* $K_{g,\nu} \coloneqq K_{g,\nu,N}$*.*

Note also that $\operatorname{ess\,sup}_\Xi\big(t\cdot g\big)$ *is the piecewise-linear function*

$$\operatorname{ess\,sup}_\Xi\big(t\cdot g\big) = \begin{cases} t\cdot \operatorname{ess\,sup}_\Xi g & t \geq 0 \\ t \cdot \operatorname{ess\,inf}_\Xi g & t \leq 0 \end{cases}.$$

Example 9 *For the Kullback–Leibler divergence,* $K_{g,\nu}(t) = \log\nu\big(e^{t(g-\nu(g))}\big)$ *by Example 7, which is the standard (centered) cumulant generating function, thereby justifying the name.*

Note that the (ϕ,ν)-cumulant generating function $K_{g,\nu}$ depends only on the pushforward measure $g_*\nu$ of ν through g. In particular, when ν is the probability distribution of a random variable X, as in Example 2, $K_{g,\nu}(t)$ can be equivalently written as

$$K_{g,\nu}(t) = \inf_{\lambda\in\mathbb{R}} \mathbb{E}[\psi^\star(t\cdot g(X) + \lambda)]\,, \tag{19}$$

highlighting the fact that $K_{g,\nu}$ only depends on $g(X)$. This contrasts with $K_{g,\nu,\Xi}$, for an arbitrary $\Xi \gg \nu$, for which the constraint on λ depends on the Ξ-essential range of g, which is not solely a property of the random variable $g(X)$ since it can depend on the value of g on ν-null sets.

Furthermore, since for $t \in \mathbb{R}$, the function $\lambda \mapsto I_{\psi^\star,\nu}(tg + \lambda)$ is convex in λ, the (ϕ, ν)-cumulant generating function is defined by a single-dimensional convex optimization problem whose objective function is expressed as an integral with respect to a probability measure (18, 19). Hence, the rich spectrum of stochastic approximation methods, such as stochastic gradient descent, can be readily applied, leading to efficient numerical procedures to evaluate $K_{g,\nu}(t)$, as long as the pushforward measure $g_*\nu$ is efficiently samplable.

Remark 38 *Since the mean deviation, and thus the optimal bound $\mathscr{L}_{g,\nu}$ is invariant to shifting g by a constant, we are in fact implicitly working in the quotient space $L^1(\nu)/\mathbb{R}\mathbf{1}_\Omega$. As such, $g \mapsto \inf_{\lambda\in\mathbb{R}} I_{\psi^\star,\nu}(g+\lambda)$ can be interpreted as the integral functional induced by $I_{\psi^\star,\nu}$ on this quotient space, by considering its infimum over all representatives of a given equivalence class. This is analogous to the definition of a norm on a quotient space.*

The following proposition states some basic properties of the cumulant generating function. As with Lemma 36, they follow from basic results in convex analysis, and we defer the proof to Appendix A.2.

Proposition 39 *For every σ-ideal Ξ, probability measure $\nu \in \mathcal{M}^1_c(\Xi)$, and $g \in L^0(\Xi)$, $K_{g,\nu,\Xi} : \mathbb{R} \to \overline{\mathbb{R}}$ is non-negative, convex, lower semicontinuous, and satisfies $K_{g,\nu,\Xi}(0) = 0$.*

Furthermore, if g is not ν-essentially constant then $K_{g,\nu,\Xi}$ is inf-compact. If there exists $c \in \mathbb{R}$ such that $g = c$ ν-almost surely, then there exists $t > 0$ (resp. $t < 0$) such that $K_{g,\nu,\Xi}(t) > 0$ if and only if $\phi'(\infty) < \infty$ and $\operatorname{ess\,sup}_\Xi g > c$ (resp. $\operatorname{ess\,inf}_\Xi g < c$).

With these definitions, we can state the main result of this section giving an expression for the optimal lower bound function.

Theorem 40 *Let (X, Y) be a ν-decomposable pair for some probability measure $\nu \in \mathcal{M}^1$ and let $\Xi = \{A \in \mathcal{F} \mid \forall\mu \in X,\ |\mu|(A) = 0\}$. Then for all $g \in Y$ and $\varepsilon \in \operatorname{int}(\operatorname{dom} \mathscr{L}_{g,\nu,X^1})$,*

$$\mathscr{L}_{g,\nu,X^1}(\varepsilon) = K^\star_{g,\nu,\Xi}(\varepsilon)\,. \tag{20}$$

Furthermore, if $\mathscr{L}_{g,\nu,X^1}$ is lower semi-continuous, equivalently if strong duality holds in (15)*, then* (20) *holds for all $\varepsilon \in \mathbb{R}$.*

Proof Lemma 32 implies that $\mathscr{L}_{g,\nu,X^1}$ is proper and convex, thus, by the Fenchel–Moreau theorem, we have $\mathscr{L}_{g,\nu,X^1} = \mathscr{L}^{\star\star}_{g,\nu,X^1}$ except possibly at the boundary of its domain, so this is simply a restatement of Proposition 33 using the terminology from Definition 37. ■

Proposition 33 and Theorem 40 show that the conjugate of the optimal lower bound only depends on the space of measures X, through the σ-ideal Ξ, as long as X forms a decomposable dual pair with a space Y of functions containing g. Hence, starting from a σ-ideal Ξ and a function g—or more generally a class of functions $\mathcal{G}$—a natural dual pair to consider is the space $X \subseteq \mathcal{M}_c(\Xi)$ of all measures integrating functions in $\mathcal{G}$, put in dual pairing with the subspace of $L^0(\Xi)$ of all functions integrable by measures in X. Formally, we have the following definition.

Definition 41 *Let $\mathcal{G}$ be a subset of $L^0(\Sigma)$ for some σ-ideal Σ. We define*

$$X_{\mathcal{G}} \coloneqq \{\mu \in \mathcal{M}_c(\Sigma) \mid \forall g \in \mathcal{G}, |\mu|(|g|) < \infty\}$$
$$\text{and } Y_{\mathcal{G}} \coloneqq \{h \in L^0(\Xi) \mid \forall \mu \in X_{\mathcal{G}}, |\mu|(|h|) < \infty\},$$

where $\Xi \coloneqq \{A \in \mathcal{F} \mid \forall \mu \in X_{\mathcal{G}}, |\mu|(A) = 0\}$.

For brevity, if $\mathcal{G} = \{g\}$ is a singleton, we write X_g for $X_{\{g\}}$ and Y_g for $Y_{\{g\}}$.

Remark 42 *We would like to use Σ rather than Ξ in the definition of $Y_{\mathcal{G}}$, but need to be careful since if $\Sigma \subsetneq \Xi$ then using Σ would prevent $(X_{\mathcal{G}}, Y_{\mathcal{G}})$ from being in separating duality. Unfortunately, there exist pathological σ-ideals for which $\Sigma \subsetneq \Xi$ (Szpilrajn (1934, 2. Corollaire)), but since for non-pathological choices of Σ (e.g. when it is the null ideal of a σ-finite, semifinite, or s-finite measure) we indeed have $\Sigma = \Xi$, we do not dwell on this distinction.*

Lemma 43 *Consider a subset $\mathcal{G} \subseteq L^0(\Sigma)$ for some σ-ideal Σ. Then for every $\nu \in X^+_{\mathcal{G}}$, the pair $\big(X_{\mathcal{G}}, Y_{\mathcal{G}}\big)$ is ν-decomposable.*

Proof That $\mu(h) < \infty$ for all $\mu \in X_{\mathcal{G}}$ and $h \in Y_{\mathcal{G}}$ is by definition. As discussed in Remark 22, it suffices to verify that item 2 in Definition 21 is true for all $\nu \in X_{\mathcal{G}}$, and indeed just for all $\nu \in X^+_{\mathcal{G}}$ since $\nu \in X_{\mathcal{G}}$ implies $|\nu| \in X^+_{\mathcal{G}}$. Item 3 and the separability of the duality between $(X_{\mathcal{G}}, Y_{\mathcal{G}})$ then follow immediately.

For item 2, consider $\nu \in X^+_{\mathcal{G}}$ and $\mu \in \mathcal{M}_c(\nu)$ such that $\frac{d\mu}{d\nu} \in L^\infty(\nu)$. Then for all $g \in \mathcal{G}$ we have $|\mu|(|g|) = \nu\Big(\Big|\frac{d\mu}{d\nu}\Big| \cdot |g|\Big) < \infty$ by Hölder's inequality, hence $\mu \in X_{\mathcal{G}}$. That $L^\infty(\Xi) \subseteq Y_{\mathcal{G}}$ holds is immediate since every $\mu \in \mathcal{M}_c(\Sigma)$ integrates every $h \in L^\infty(\Xi)$. ■

The following easy corollary is an "operational" restatement of Theorem 40, specialized to the dual pair of Definition 41, and highlighting the duality between upper bounding the cumulant generating function and lower bounding the ϕ-divergence by a convex lower semicontinuous function of the mean deviation.

Corollary 44 *Consider a measurable function $g \in L^0(\Sigma)$ for some σ-ideal Σ and let $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X_g, |\mu|(A) = 0\}$. Then for every probability measure $\nu \in X^1_g \coloneqq X_g \cap \mathcal{M}^1$ and convex lower semicontinuous function $L : \mathbb{R} \to \overline{\mathbb{R}}_{\geq 0}$, the following are equivalent:*

(i) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(\mu(g) - \nu(g))$ *for every $\mu \in X^1_g$.*

(ii) $K_{g,\nu,\Xi} \leq L^\star$.

Example 10 *The Hammersley–Chapman–Robbins bound in statistics is an immediate corollary of Corollary 44 applied to the χ^2-divergence given by $\phi(x) = (x-1)^2 + \delta_{\mathbb{R}_{\geq 0}}(x)$: The convex conjugate of $\psi(x) = x^2 + \delta_{[-1,\infty)}(x)$ is*

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

and satisfies in particular $\psi^\star(x) \leq x^2/4$, so that $K_{g,\nu}(t) \leq \inf_\lambda \int (tg+\lambda)^2/4 \,\mathrm{d}\nu = t^2 \operatorname{Var}_\nu(g)/4$. Since the convex conjugate of $t \mapsto t^2 \operatorname{Var}_\nu(g)/4$ is $t \mapsto t^2/\operatorname{Var}_\nu(g)$, we obtain for all $\mu, \nu \in \mathcal{M}^1$ and $g \in L^1(\nu)$ that $\chi^2(\mu \parallel \nu) \geq (\mu(g) - \nu(g))^2/\operatorname{Var}_\nu(g)$.

Theorem 40 also gives a useful characterization of the existence of a non-trivial lower bound by the *absolute* mean deviation.

Corollary 45 *Consider a measurable function $g \in L^0(\Sigma)$ for some σ-ideal Σ and let $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X_g, |\mu|(A) = 0\}$. Then for every $\nu \in X_g^1$, the optimal lower bound $\mathscr{L}_{\{\pm g\},\nu,X_g^1}$ is non-zero if and only if $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu,\Xi})$. In other words, the following are equivalent*

(i) there exists a non-zero function $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}_{\geq 0}$ such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L\big(|\mu(g) - \nu(g)|\big)$ for every $\mu \in X_g^1$.

(ii) the function $K_{g,\nu,\Xi}$ is finite on an open interval around 0.

Proof Writing $M = X_g^1$, we have by Eq. (16) that the function $\mathscr{L}_{\{\pm g\},\nu,M}$ is non-zero if and only if there exists $\varepsilon > 0$ such that $\mathscr{L}_{g,\nu,M}(\varepsilon) \neq 0 \neq \mathscr{L}_{g,\nu,M}(-\varepsilon)$. Since $\mathscr{L}_{g,\nu,M}$ is convex, non-negative, and 0 at 0 by Lemma 32, such an ε exists if and only if 0 is contained in the interval $\big(\mathscr{L}'_{g,\nu,M}(-\infty), \mathscr{L}'_{g,\nu,M}(\infty)\big)$, the interior of the domain of $\mathscr{L}^\star_{g,\nu,M}$. ■

Remark 46 *Throughout this section, we have seen the σ-ideal Ξ appear in our results, in particular via $K_{g,\nu,\Xi}$ in Corollary 45. We will see in Theorem 83 however that when we consider a true IPM where we require the bound L to hold jointly for all measures ν and μ, we can ignore the σ-ideal Ξ and consider only $K_{g,\nu}$.*

5.2 Subexponential functions and connections to Orlicz spaces

In Sections 5.2 to 5.4 , we explore properties of the set of functions satisfying the conditions of Corollary 45, i.e. for which there is a non-trivial lower bound of the ϕ-divergence in terms of the absolute mean deviation, and show its relation to topological properties of the divergence. A reader primarily interested in quantitative bounds for IPMs can skip to Section 6.

In light of Corollary 45, we need to consider the set of functions g such that $\operatorname{dom} K_{g,\nu,\Xi}$ contains a neighborhood of zero. The following lemma shows that this is the case for bounded functions, and that furthermore, when $\phi'(\infty) < \infty$, boundedness is necessary. In other words, when $\phi'(\infty) < \infty$, the ϕ-divergence cannot upper bound the absolute mean deviation of an unbounded function. This is in sharp contrast with the KL divergence (satisfying $\phi'(\infty) = \infty$), for which such upper bounds exist as long as the function satisfies Gaussian-type tail bounds (Boucheron et al., 2013, §4.10).

Lemma 47 *Let Ξ be a σ-ideal and $\nu \in \mathcal{M}_c^1(\Xi)$. If $g \in L^\infty(\Xi)$ then $\operatorname{dom} K_{g,\nu,\Xi}$ is all of $\mathbb{R}$, and in particular contains a neighborhood of zero. Furthermore, when $\phi'(\infty) < \infty$, we have conversely that if 0 is in the interior of the domain of $K_{g,\nu,\Xi}$, then $g \in L^\infty(\Xi)$, in which case $K_{g,\nu}(t) = K_{g,\nu,\Xi}(t)$ whenever $|t| \cdot \big(\operatorname{ess\,sup}_\Xi g - \operatorname{ess\,inf}_\Xi g\big) \leq \phi'(\infty)$.*

Remark 48 *As already discussed, Lemma 47 implies that when $\phi'(\infty) < \infty$, boundedness of g is necessary for the existence of a non-trivial lower bound on $\mathrm{D}_\phi(\mu \parallel \nu)$ in terms of the $|\mu(g) - \nu(g)|$. Moreover, we can deduce from Lemma 47 that in this case, any non-trivial lower bound must depend on $\operatorname{ess\,sup}_\Xi |g|$ and cannot depend only on properties of g such as its ν-variance. In particular, any non-trivial lower bound must converge to 0 as $\operatorname{ess\,sup}_\Xi |g|$ converges to $+\infty$, for if it were not the case, one could obtain a non-trivial lower bound for an unbounded function g by approximating it with bounded functions $g \cdot \mathbf{1}\{|g| \leq n\}$.*

Proof Recall that $(-\infty, 0] \subseteq \operatorname{dom} \psi^\star$ and that $\psi^\star(x) \leq -x$ for $x \leq 0$ by Lemma 36. For $g \in L^\infty(\Xi)$, write B for $\operatorname{ess\,sup}_\Xi |g|$, and for $t \in \mathbb{R}$, write $\lambda \coloneqq -|t| \cdot B$. Then we have that $-2|t|B \leq t \cdot g + \lambda \leq 0 \leq \phi'(\infty)$ holds Ξ-a.s., and thus also $\psi^\star(tg + \lambda) \leq 2|t|B$ holds Ξ-a.s. Thus $K_{g,\nu,\Xi}(t)$ is at most $2|t| \cdot B < \infty$ by definition, and since t is arbitrary, we get $\operatorname{dom} K_{g,\nu,\Xi} = \mathbb{R}$.

We now assume $\phi'(\infty) < \infty$ and prove the converse claim. If $K_{g,\nu,\Xi}(t)$ is finite for some $t \in \mathbb{R}$, then $tg + \lambda \leq \phi'(\infty)$ holds Ξ-a.s. for some $\lambda \in \mathbb{R}$. In particular, if it holds for some $t > 0$, then $\operatorname{ess\,sup}_\Xi g$ is finite, and if it holds for some $t < 0$, then $\operatorname{ess\,inf}_\Xi g$ is finite.

For the remaining claim, since $\psi^\star$ is non-decreasing on the non-negative reals we have that $K_{g,\nu}(t) = \inf\bigl\{I_{\psi^\star,\nu}(tg + \lambda) \bigm| \lambda \in \mathbb{R}\bigr\} = \inf\bigl\{I_{\psi^\star,\nu}(tg + \lambda) \bigm| \operatorname{ess\,inf}_\Xi tg + \lambda \leq 0\bigr\}$. But if $\operatorname{ess\,sup}_\Xi\bigl(t\cdot g\bigr) - \operatorname{ess\,inf}_\Xi\bigl(t\cdot g\bigr) \leq \phi'(\infty)$, then $\operatorname{ess\,inf}_\Xi t\cdot g + \lambda \leq 0$ implies $\operatorname{ess\,sup}_\Xi tg + \lambda \leq \phi'(\infty)$ and $K_{g,\nu}(t) \geq K_{g,\nu,\Xi}(t) \geq K_{g,\nu}(t)$. ■

Since Lemma 47 completely characterizes the existence of a non-trivial lower bound when $\phi'(\infty) < \infty$, we focus on the case $\phi'(\infty) = \infty$ in the remainder of this section. Recall that $K_{g,\nu} = K_{g,\nu,\Xi}$ in this case, so we only need to consider $K_{g,\nu}$ in the following definition.

Definition 49 ((ϕ, ν)-subexponential functions) *Let $\nu \in \mathcal{M}^1$ be a probability measure. We say that the function $g \in L^0(\nu)$ is* (ϕ, ν)-subexponential *if $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu})$ and we denote by $S^\phi(\nu)$ the space of all such functions. We further say that $g \in L^0(\nu)$ is* strongly (ϕ, ν)-subexponential *if $\operatorname{dom} K_{g,\nu} = \mathbb{R}$ and denote by $S^\phi_\heartsuit(\nu)$ the space of all such functions.*

Example 11 *For the case of the KL-divergence, if the pushforward $g_*\nu$ of ν induced by g on $\mathbb{R}$ is the Gaussian distribution (respectively the gamma distribution), then g is strongly subexponential (respectively subexponential). Furthermore, it follows from Example 9 that $g \in S^\phi(\nu)$ iff the moment-generating function of g is finite on a neighborhood of 0, which is the standard definition of subexponential functions (see e.g. Vershynin (2018, §2.7)) and thus justifies our terminology.*

Example 12 *Lemma 47 shows that $L^\infty(\nu) \subseteq S^\phi_\heartsuit(\nu)$ and that furthermore, if $\phi'(\infty) < \infty$, then $L^\infty(\nu) = S^\phi_\heartsuit(\nu) = S^\phi(\nu)$.*

We start with the following key lemma allowing us to relate the finiteness of $K_{g,\nu}$ to the finiteness of the function $t \mapsto I_{\psi^\star,\nu}(tg)$.

Lemma 50 *For $\nu \in \mathcal{M}^1$, $g \in L^0(\nu)$, and $t \in \operatorname{dom} K_{g,\nu}$, we have that if $\phi'(\infty) = \infty$ (resp. $\phi'(\infty) > 0$) then $\alpha t g \in \operatorname{dom} I_{\psi^\star,\nu}$ for all $\alpha \in (0, 1)$ (resp. for sufficiently small $\alpha > 0$).*

Proof Let $\lambda \in \mathbb{R}$ be such that $\int \psi^\star(tg + \lambda)\,\mathrm{d}\nu < \infty$ (such a λ exists since $t \in \operatorname{dom} K_{g,\nu}$). Using the convexity of $\psi^\star$, we get for any $\alpha \in (0, 1)$

$$\begin{aligned}\int \psi^\star(\alpha t g)\,\mathrm{d}\nu &= \int \psi^\star\biggl(\alpha(tg + \lambda) + (1-\alpha)\frac{-\alpha\lambda}{1-\alpha}\biggr)\,\mathrm{d}\nu\\ &\leq \alpha \int \psi^\star(tg + \lambda)\,\mathrm{d}\nu + (1-\alpha)\psi^\star\biggl(\frac{-\alpha\lambda}{1-\alpha}\biggr).\end{aligned}$$

The first summand is finite by definition, and if $-\alpha\lambda/(1-\alpha) \in \operatorname{dom}\psi^\star \supseteq (-\infty, \phi'(\infty))$ then so is the second summand. If $\phi'(\infty) = \infty$ this holds for all $\alpha \in (0,1)$, and if $\phi'(\infty) > 0$ it holds for sufficiently small $\alpha > 0$. ■

Remark 51 *When $\phi'(\infty) < \infty$, it is not necessarily true that any $\alpha \in (0,1)$ can be used in Lemma 50. For example, Lemma 47 implies that* $\operatorname{dom} K_{g,\nu} = \mathbb{R}$ *for all $g \in L^\infty(\nu)$, but since* $\operatorname{dom}\psi^\star \subseteq \big(-\infty, \phi'(\infty)\big]$ *we have $I_{\psi^\star,\nu}(tg) = \infty$ for sufficiently large (possibly only positive or negative) t, unless g is zero ν-a.s.*

The following proposition gives useful characterizations of subexponential functions in terms of the finiteness of different integral functionals of g.

Proposition 52 *Suppose that $\phi'(\infty) = \infty$ and fix $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$. Then the following are equivalent:*

(i) g is (ϕ,ν)-subexponential

(ii) $K_{|g|,\nu}(t) < \infty$ for some $t > 0$

(iii) $g \in L^\theta(\nu)$ for $\theta : x \mapsto \max\{\psi^\star(x), \psi^\star(-x)\}$ (here $L^\theta(\nu)$ is the Orlicz space defined in Section 3.3)

Proof *(i)* $\implies$ *(ii)* If $\operatorname{dom} K_{g,\nu}$ contains an open interval around 0, Lemma 50 and the convexity of $\operatorname{dom} I_{\psi^\star,\nu}$ imply that there exists $s > 0$ such that $\int \psi^\star(tg)\,\mathrm{d}\nu < \infty$ for all $|t| < s$. By non-negativity of $\psi^\star$, $\int \psi^\star(t|g|)\,\mathrm{d}\nu \le \int \psi^\star(tg) + \psi^\star(-tg)\,\mathrm{d}\nu < \infty$ for all $t \in (-s,s)$, which in turns implies $(-s,s) \subseteq \operatorname{dom} K_{|g|,\nu}$.

(ii) $\implies$ *(iii)* Define $\eta(x) \coloneqq \psi^\star(|x|)$. Since $\psi^\star(x) \le -x$ for $x \le 0$ by Lemma 36, we have that $\eta(x) \le \theta(x) \le \eta(x) + |x|$ for all $x \in \mathbb{R}$. Since we also have $L^\eta(\nu) \subseteq L^1(\nu)$, this implies that $g \in L^\eta(\nu)$ if and only if $L^\theta(\nu)$. We conclude after observing that $K_{|g|,\nu}(t) < \infty$ for some $t > 0$ implies that $g \in L^\eta(\nu)$ by Lemma 50.

(iii) $\implies$ *(i)* Observe that for all $t \in \mathbb{R}$,

$$\max\{K_{g,\nu}(t), K_{g,\nu}(-t)\} \le \max\left\{\int \psi^\star(tg)\,\mathrm{d}\nu, \int \psi^\star(-tg)\,\mathrm{d}\nu\right\} \le \int \theta(tg)\,\mathrm{d}\nu\,, \tag{21}$$

where the first inequality is by definition of $K_{g,\nu}$ and the second inequality is by monotonicity of the integral and the definition of θ. Since $\operatorname{dom} K_{g,\nu}$ is convex, if there exists $t > 0$ such that $I_{\theta,\nu}(tg) < \infty$, then (21) implies that $[-t,t] \subseteq \operatorname{dom} K_{g,\nu}$ and g is (ϕ,ν)-subexponential. ■

Remark 53 *Though Proposition 52 implies that the set of (ϕ,ν)-subexponential functions is the same as the set $L^\theta(\nu)$ for $\theta(x) = \max\{\psi^\star(x), \psi^\star(-x)\}$, we emphasize that the Luxemburg norm $\|\cdot\|_\theta$ does* not *capture the relationship between $\mathrm{D}_\phi(\mu \,\|\, \nu)$ and the absolute mean deviation $|\mu(g) - \nu(g)|$. First, the function θ, being a symmetrization of $\psi^\star$, induces integral functionals which are potentially much larger than those defined by $\psi^\star$, in particular it is possible to have* $\max\{K_{g,\nu}(t), K_{g,\nu}(-t)\} < \inf_{\lambda\in\mathbb{R}} I_{\theta,\nu}(tg+\lambda) < I_{\theta,\nu}(tg)$*. Furthermore, the Luxemburg norm summarizes the growth of $t \mapsto I_{\theta,\nu}(tg)$ with a single number (specifically its inverse at 1), whereas Theorem 40 shows that the relationship with the mean deviation is controlled by $K^\star_{g,\nu}$, which depends on the growth of $K_{g,\nu}(t)$ with t.*

We are now ready to prove the main result of this section, which is that the space $S^\phi(\nu)$ of (ϕ,ν)-subexponential functions is the largest space of functions which can be put in dual pairing with (the span of) all measures $\mu \in \mathcal{M}_c(\nu)$ such that $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$, i.e. $\operatorname{dom} I_{\phi,\nu}$.

Theorem 54 *For $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$, the following are equivalent:*

(i) g is (ϕ,ν)-subexponential, i.e. $g \in S^\phi(\nu)$.

(ii) g is μ-integrable for every $\mu \in \mathcal{M}_c(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$.

(iii) g is μ-integrable for every $\mu \in \mathcal{M}_c^1(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$.

Proof $(i) \implies (ii)$ If $\phi'(\infty) < \infty$ this follows since $L^\infty(\nu) = S^\phi(\nu)$, so assume that $\phi'(\infty) = \infty$. If $g \in S^\phi(\nu)$ then $g \in L^\theta(\nu)$ for $\theta(x) = \max\{\psi^\star(x), \psi^\star(-x)\}$ by Proposition 52. Since $\theta \geq \psi^\star$ we have $\theta^\star \leq \psi$, and thus for $\mu \in \mathcal{M}_c(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$,

$$I_{\theta^\star,\nu}\left(\frac{d\mu}{d\nu} - 1\right) \leq I_{\psi,\nu}\left(\frac{d\mu}{d\nu} - 1\right) = \mathrm{D}_\phi(\mu \parallel \nu) < \infty\,,$$

implying that $\frac{d\mu}{d\nu} - 1 \in L^{\theta^\star}(\nu)$. Furthermore, since $1 \in L^\infty(\nu) \subseteq L^{\theta^\star}(\nu)$ we get that $\frac{d\mu}{d\nu} \in L^{\theta^\star}(\nu)$. Property 2. then follows from the fact that $(L^{\theta^\star}, L^\theta)$ form a dual pair.

$(ii) \implies (iii)$ Immediate.

$(iii) \implies (i)$ Define $C \coloneqq \{\mu \in \mathcal{M}_c^1(\nu) \mid \mathrm{D}_\phi(\mu \parallel \nu) \leq 1\}$, which is closed and convex as a sublevel set of the convex lower semicontinuous functional $\widetilde{\mathrm{D}}_{\phi,\nu}$ on the Banach space $\mathcal{M}_c(\nu)$ with the total variation norm (recall that this space is isomorphic to $L^1(\nu)$ by the Radon–Nikodym theorem). Since furthermore $C \subseteq \mathcal{M}^1$, it is bounded in $\mathcal{M}_c(\nu)$ and so is cs-compact (Jameson, 1972, Proposition 2). Then by assumption, the linear function $\mu \mapsto \mu(|g|)$ is well-defined and bounded below by 0 on C, so Lemma 16 implies that there exists $B \in \mathbb{R}$ such that $\mu(|g|) \leq B$ for all $\mu \in C$. Thus, we get that for all $\mu \in C$, $|\mu(g) - \nu(g)| \leq \mu(|g|) + \nu(|g|) \leq B + \nu(|g|)$. In particular, if $|\mu(g) - \nu(g)| > B + \nu(|g|)$ then $\mathrm{D}_\phi(\mu \parallel \nu) > 1$, proving the existence of a non-zero function L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L\big(|\mu(g) - \nu(g)|\big)$. This implies that $g \in S^\phi(\nu)$ by Corollary 45. ■

We have the following characterization of the space $S_\heartsuit^\phi(\nu)$ of strongly subexponential functions. In particular $S_\heartsuit^\phi(\nu)$ can be identified as a set with $L^\infty(\nu)$ or the Orlicz heart $L_\heartsuit^\theta(\nu)$ depending on whether $\phi'(\infty)$ is finite or infinite (with the finite case from Lemma 47).

Proposition 55 *Suppose that $\phi'(\infty) = \infty$ and fix $\nu \in \mathcal{M}^1$ and $g \in L^0(\nu)$. Then the following are equivalent:*

(i) g is strongly (ϕ,ν)-subexponential, i.e. $g \in S_\heartsuit^\phi(\nu)$.

(ii) $K_{|g|,\nu}(t) < \infty$ for all $t > 0$.

(iii) $g \in L_\heartsuit^\theta(\nu)$ for $\theta : x \mapsto \max\{\psi^\star(x), \psi^\star(-x)\}$.

Proof $(i) \implies (ii)$ Since $\phi'(\infty) = \infty$, Lemma 50 implies that $tg \in \operatorname{dom} I_{\psi^\star,\nu}$ for all $t \in \mathbb{R}$, and since $\psi^\star$ is non-negative we have for each $t > 0$ that $K_{|g|,\nu}(t) \leq \int \psi^\star(t|g|)\,\mathrm{d}\nu \leq \int \psi^\star(tg) + \psi^\star(-tg)\,\mathrm{d}\nu < \infty$.

(ii) $\implies$ *(iii)* Define $\eta : x \mapsto \psi^\star(|x|)$, so that by Lemma 50 we have $\int \eta(tg)\,\mathrm{d}\nu = \int \psi^\star(t|g|)\,\mathrm{d}\nu < \infty$ for all $t > 0$, and hence Property 2. implies $g \in L^\eta_\heartsuit(\nu)$. As in the proof of Proposition 52, $\eta(x) \le \theta(x) \le \eta(x) + |x|$ for all $x \in \mathbb{R}$ and since $L^\eta_\heartsuit(\nu) \subseteq L^1(\nu)$, we have that $g \in L^\eta_\heartsuit(\nu)$ iff $g \in L^\theta_\heartsuit(\nu)$.

(iii) $\implies$ *(i)* Immediate since for $t \in \mathbb{R}$, $K_{g,\nu}(t) \le \int \psi^\star(tg)\,\mathrm{d}\nu \le \int \theta(tg)\,\mathrm{d}\nu < \infty$. ■

Finally, we collect several statements from this section and express them in a form which will be convenient for subsequent sections.

Corollary 56 *Define $\theta(x) \coloneqq \max\{\psi^\star(x), \psi^\star(-x)\}$. Then we have $S^\phi_\heartsuit(\nu) \subseteq S^\phi(\nu) \subseteq L^1(\nu)$ and $\operatorname{dom} I_{\phi,\nu} \subseteq L^{\theta^\star}(\nu) \subseteq L^1(\nu)$. Furthermore, $L^{\theta^\star}(\nu)$ is in dual pairing with both $S^\phi(\nu)$ and $S^\phi_\heartsuit(\nu)$, and when $\phi'(\infty) = \infty$ the topology induced by $\|\cdot\|_\theta$ on $S^\phi_\heartsuit(\nu)$ is complete and compatible with the pairing.*

Proof The containment $S^\phi(\nu) \subseteq L^1(\nu)$ is because $S^\phi(\nu)$ is equal as a set to the Orlicz space $L^\theta(\nu)$ by Proposition 52, and the containment $\operatorname{dom} I_{\phi,\nu} \subseteq L^{\theta^\star}(\nu)$ can be found in the proof of *(i)* $\implies$ *(ii)* of Theorem 54. The fact that $\big(L^{\theta^\star}(\nu), S^\phi(\nu)\big)$ form a dual pair is also immediate from the identification of $S^\phi(\nu)$ with $L^\theta(\nu)$ as a set. Finally, the last claim follows from the identification of $S^\phi_\heartsuit(\nu)$ with $L^\theta_\heartsuit(\nu)$ as a set and the fact that when $\phi'(\infty) = \infty$, then $\operatorname{dom}\theta = \mathbb{R}$ implying that the topological dual of the Banach space $(L^\theta_\heartsuit(\nu), \|\cdot\|_\theta)$ is isomorphic to $(L^{\theta^\star}(\nu), \|\cdot\|_{\theta^\star})$. ■

5.3 Inf-compactness of divergences and connections to strong duality

In this section, we study the question of inf-compactness of the functional $\mathrm{D}_{\phi,\nu}$ and that of its restriction $\widetilde{\mathrm{D}}_{\phi,\nu}$ to probability measures. Specifically, we wish to understand under which topology the information "ball" $\mathcal{B}_{\phi,\nu}(\tau) \coloneqq \{\mu \in \mathcal{M} \mid \mathrm{D}_\phi(\mu \,\|\, \nu) \le \tau\}$ is compact. Beyond being a natural topological question, it also has implications for strong duality in Theorem 40, since the following lemma shows that compactness of the ball under suitable topologies implies strong duality.

Lemma 57 *For every g, ν, and M as in Definition 31, if $\mu \mapsto \mathrm{D}_\phi(\mu \,\|\, \nu)$ is inf-compact (or even countably inf-compact) with respect to a topology on M such that $\mu \mapsto \mu(g)$ is continuous, then $\mathscr{L}_{g,\nu,M}$ is inf-compact (and in particular lower semicontinuous), so that strong duality holds in Theorem 40.*

Proof Recall from Eq. (15) that

$$\mathscr{L}_{g,\nu,M}(\varepsilon) = \inf_{\mu \in M} \mathrm{D}_\phi(\mu \,\|\, \nu) + \delta_{\{0\}}(\mu(g) - \nu(g) - \varepsilon)$$

where $f(\varepsilon, \mu) = \mathrm{D}_\phi(\mu \,\|\, \nu) + \delta_{\{0\}}(\mu(g) - \nu(g) - \varepsilon)$ is convex. Furthermore, under the stated assumption, we have that f is also inf-compact so that Lemma 5 gives the claim. ■

Throughout this section, we assume that $\phi'(\infty) = \infty$,[5] which implies that $\operatorname{dom}\psi^\star = \mathbb{R}$ by Lemma 36, and furthermore that $\mu \in \mathcal{M}_c(\nu)$ whenever $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$ and hence $\mathrm{D}_{\phi,\nu} = I_{\phi,\nu}$ and $\mathcal{B}_{\phi,\nu}(\tau) \subset \mathcal{M}_c(\nu)$ for all $\tau \geq 0$. It is well known that in this case, $\mathcal{B}_{\phi,\nu}(\tau)$ is compact in the weak topology $\sigma\big(L^1(\nu), L^\infty(\nu)\big)$ (e.g. Rockafellar (1971, Corollary 2B) or Teboulle and Vajda (1993)). This fact can be derived as a simple consequence of the Dunford–Pettis theorem since $\mathcal{B}_{\phi,\nu}(\tau)$ is uniformly integrable by the de la Vallée-Poussin theorem (see e.g. Valadier (1970, pages 67–68)). In light of Lemma 57, it is however useful to understand whether $\mathcal{B}_{\phi,\nu}(\tau)$ is compact under topologies for which $\mu \mapsto \mu(g)$ is continuous, where g could be unbounded. Léonard (Léonard, 2001b, Theorem 3.4) showed, in the context of convex integral functionals on Orlicz spaces, that strong duality holds when $g \in S^\phi_\heartsuit(\nu)$, and in this section we reprove this result in the language of ϕ-divergences by noting (as is implicit in Léonard (2001b, Lemma 3.1)) that $\mathcal{B}_{\phi,\nu}(\tau)$ is compact for the initial topology induced by the maps of the form $\mu \mapsto \mu(g)$ for all strongly subexponential function $g \in S^\phi_\heartsuit(\nu)$.

Proposition 58 *Fix $\nu \in \mathcal{M}^1$ and define $\theta : x \mapsto \max\{\psi^\star(x), \psi^\star(-x)\}$ as in Proposition 55. If $\phi'(\infty) = \infty$, then the functional $I_{\phi,\nu}$ is $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\heartsuit(\nu)\big)$ inf-compact.*

Proof By Corollary 56, we know that $\big(S^\phi_\heartsuit(\nu), \|\cdot\|_\theta\big)$ is a Banach space in dual pairing with $L^{\theta^\star}(\nu)$. Thus, from Proposition 23, the integral functional $I_{\phi^\star,\nu}$ defined on $S^\phi_\heartsuit(\nu)$ is convex, lower semicontinuous, and has conjugate $I_{\phi^{\star\star},\nu} = I_{\phi,\nu}$ on $L^{\theta^\star}(\nu)$. Furthermore, from Lemma 50 we know for every $g \in S^\phi_\heartsuit(\nu)$ that $I_{\phi^\star,\nu}(g) < \infty$, so $I_{\phi^\star,\nu}$ is convex, lsc, and finite everywhere on a Banach space, and thus continuous everywhere by Brøndsted (1964, 2.10). Finally, Moreau (1964, Proposition 1) implies that its conjugate $I_{\phi,\nu}$ is inf-compact on $L^{\theta^\star}(\nu)$ with respect to the weak topology $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\heartsuit(\nu)\big)$. ■

Remark 59 *This result generalizes Rockafellar (1971, Corollary 2B) since $L^\infty(\nu) \subseteq S^\phi_\heartsuit(\nu)$ whenever $\phi'(\infty) = \infty$ (see Example 12).*

Corollary 60 *Under the same assumptions and notations as Proposition 58, the functional $\widetilde{\mathrm{D}}_{\phi,\nu}$ is $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\heartsuit(\nu)\big)$ inf-compact.*

Proof Observe that since $\phi(x) = \infty$ for $x < 0$, we have for every $\tau \in \mathbb{R}$ that $\{\mu \in L^{\theta^\star}(\nu) \mid \widetilde{\mathrm{D}}_{\phi,\nu}(\mu) \leq \tau\} = \{\mu \in \mathcal{M}^1 \cap L^{\theta^\star}(\nu) \mid I_{\phi,\nu}(\mu) \leq \tau\} = \{\mu \in L^{\theta^\star}(\nu) \mid I_{\phi,\nu}(\mu) \leq \tau\} \cap f^{-1}(1)$ where $f : \mu \to \mu(\mathbf{1}_\Omega)$ is continuous in the weak topology $\sigma\big(L^{\theta^\star}(\nu), S^\phi_\heartsuit(\nu)\big)$ since $L^\infty(\nu) \subseteq S^\phi_\heartsuit(\nu)$ by Lemma 47. Hence, $\mathcal{M}^1 \cap \mathcal{B}_{\phi,\nu}(\tau)$ is compact as a closed subset of a compact set. ■

Corollary 61 *If $\phi'(\infty) = \infty$, then for every $\tau \in \mathbb{R}$ the sets $\mathcal{B}_{\phi,\nu}(\tau)$ and $\mathcal{M}^1 \cap \mathcal{B}_{\phi,\nu}(\tau)$ are compact in the initial topology induced by $\{\mu \mapsto \mu(g) \mid g \in S^\phi_\heartsuit(\nu)\}$.*

Proof Immediate from Proposition 58 and Corollary 60. ■

[5] When $\phi'(\infty) < \infty$, compactness of information balls is very dependent on the specific measure space $(\Omega, \mathcal{F}, \nu)$, and in this work we avoid such conditions.

Corollary 62 *Let $\nu \in \mathcal{M}^1$ be a probability measure and assume that $\phi'(\infty) = \infty$. If $g \in L^0(\nu)$ is strongly (ϕ, ν)-subexponential and $M \subseteq \mathcal{M}^1_c(\nu)$ is a convex set of probability measures containing every $\mu \in \mathcal{M}^1_c(\nu)$ with $\mathrm{D}_\phi(\mu \parallel \nu) < \infty$, then the function $\mathscr{L}_{g,\nu,M}$ is lower semicontinuous.*

Proof Follows from Lemma 57 and Corollary 61. ■

Remark 63 *Corollary 62 does not apply when $\phi'(\infty) < \infty$ or $g \in S^\phi(\nu) \setminus S^\phi_\heartsuit(\nu)$ (e.g. when the pushforward measure $g_*\nu$ is gamma-distributed in the case of the KL divergence), and it would be interesting to identify conditions other than inf-compactness of $\mathrm{D}_{\phi,\nu}$ under which $\mathscr{L}_{g,\nu}$ is lower semicontinuous.*

5.4 Convergence in ϕ-divergence and weak convergence

Our goal in this section is to relate two notions of convergence for a sequence of probability measures $(\nu_n)_{n\in\mathbb{N}}$ and $\nu \in \mathcal{M}^1$: (i) $\mathrm{D}_\phi(\nu_n \parallel \nu) \to 0$,[6] and (ii) $|\nu_n(g)-\nu(g)| \to 0$ for $g \in \mathcal{L}^0(\Omega)$. Specifically, we would like to identify the largest class of functions $g \in \mathcal{L}^0(\Omega)$ such that *convergence in ϕ-divergence* (i) implies (ii). In other words, we would like to identify the finest initial topology induced by linear forms $\mu \mapsto \mu(g)$ for which (sequential) convergence is implied by (sequential) convergence in ϕ-divergence.[7] This question is less quantitative than computing the best lower bound of the ϕ-divergence in terms of the absolute mean deviation, since it only characterizes when $|\nu_n(g) - \nu(g)|$ converges to 0, whereas the optimal lower bound quantifies the *rate of convergence* to 0 when it occurs.

This has been studied in the specific case of the Kullback–Leibler divergence by Harremoës, who showed (Harremoës, 2007, Theorem 25) that $\mathrm{D}(\nu_n \parallel \nu) \to 0$ implies $|\nu_n(g) - \nu(g)| \to 0$ for every non-negative function g whose moment generating function is finite at some positive real (in fact, the converse was also shown in the same paper under a so-called *power-dominance* condition on ν). In this section, we generalize this to an arbitrary ϕ-divergence and show that convergence in ϕ-divergence implies $\nu_n(g) \to \nu(g)$ if and only if g is (ϕ, ν)-subexponential.

This question is also closely related to the one of understanding the relationship between weak convergence and *modular convergence* in Orlicz spaces (e.g. Nakano (1950) or Musielak (1983)). Although convergence in ϕ-divergence as defined above only formally coincides with the notion of modular convergence when ϕ is symmetric about 1 (though this can sometimes be relaxed (Herda, 1967)) and satisfies the so-called Δ_2 growth condition, it is possible that this line of work could be adapted to the question studied in this section.

We start with the following proposition, showing that this question is equivalent to the differentiability of $\mathscr{L}^\star_{g,\nu}$ at 0.

[6] Throughout this section, we restrict our attention to ϕ which are not the constant 0 on a neighborhood of 1, i.e. such that $1 \notin \operatorname{int}\{x \in \mathbb{R} \mid \phi(x) = 0\}$, as otherwise it is easy to construct probability measures $\mu \neq \nu$ such that $\mathrm{D}_\phi(\mu \parallel \nu) = 0$, hence $\mathrm{D}_\phi(\nu_n \parallel \nu) \to 0$ does not define a meaningful convergence notion.

[7] The natural notion of convergence in ϕ-divergence defines a topology on the space of probability measures for which continuity and sequential continuity coincide (see e.g. Kisyński (1960); Dudley (1964); Harremoës (2007)), so it is without loss of generality that we consider only sequences rather than nets in the rest of this section. Note that the information balls $\{\mu \in \mathcal{M}^1 \mid \mathrm{D}_\phi(\mu \parallel \nu) < \tau\}$ for $\tau > 0$ need not be neighborhoods of ν in this topology, and the information balls do not in general define a basis of neighborhoods for a topology on the space of probability measures (Csiszár, 1962, 1964, 1967; Dudley, 1998).

Proposition 64 *Let $\nu \in \mathcal{M}^1$, $g \in \mathcal{L}^1(\nu)$, and $M \subseteq \mathcal{M}^1$ be a convex set of measures integrating g and containing ν. Then the following are equivalent:*

(i) $\lim_{n\to\infty} \nu_n(g) = \nu(g)$ *for all* $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$ *such that* $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \parallel \nu) = 0$.

(ii) $\mathscr{L}_{g,\nu,M}$ *is strictly convex at* 0*, that is* $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$ *if and only if* $\varepsilon = 0$.

(iii) $\partial\mathscr{L}^\star_{g,\nu,M}(0) = \{0\}$*, that is* $\mathscr{L}^\star_{g,\nu,M}$ *is differentiable at* 0 *and* $\mathscr{L}^{\star\prime}_{g,\nu,M}(0) = 0$.

Proof *(i)* $\implies$ *(ii)* Assume for the sake of contradiction that $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$ for some $\varepsilon \neq 0$. Then by definition of $\mathscr{L}_{g,\nu,M}$, there exists a sequence $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$ such that for all $n \in \mathbb{N}$, $\mathrm{D}_\phi(\nu_n \parallel \nu) \leq 1/n$ and $\nu_n(g) - \nu(g) = \varepsilon$, thus contradicting *(i)*. Hence, $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$ if and only if $\varepsilon = 0$, which is equivalent to strict convexity at 0 since $\mathscr{L}_{g,\nu,M}$ is convex with global minimum $\mathscr{L}_{g,\nu,M}(0) = 0$ by Lemma 32.

(ii) $\implies$ *(i)* Let $(\nu_n)_{n\in\mathbb{N}} \in M^{\mathbb{N}}$ be a sequence such that $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \parallel \nu) = 0$. By definition of $\mathscr{L}_{g,\nu,M}$, we have that $\mathrm{D}_\phi(\nu_n \parallel \nu) \geq \mathscr{L}_{g,\nu,M}\big(\nu_n(g) - \nu(g)\big) \geq 0$ for all $n \in \mathbb{N}$, and in particular $\lim_{n\to\infty} \mathscr{L}_{g,\nu,M}\big(\nu_n(g) - \nu(g)\big) = 0$. Assume for the sake of contradiction that $\nu_n(g)$ does not converge to $\nu(g)$. This implies the existence of $\varepsilon > 0$ such that $|\nu_n(g) - \nu(g)| \geq \varepsilon$ for infinitely many $n \in \mathbb{N}$. But then $\mathscr{L}_{g,\nu,M}\big(\nu_n(g) - \nu(g)\big) \geq \min\{\mathscr{L}_{g,\nu,M}(\varepsilon), \mathscr{L}_{g,\nu,M}(-\varepsilon)\} > 0$ for infinitely many $n \in \mathbb{N}$, a contradiction.

(ii) $\iff$ *(iii)* By a standard characterization of the subdifferential (see e.g. Zălinescu (2002, Theorem 2.4.2(iii))), we have that $\partial\mathscr{L}^\star_{g,\nu,M}(0) = \{x \in \mathbb{R} \mid \mathscr{L}^\star_{g,\nu,M}(0) + \mathscr{L}^{\star\star}_{g,\nu,M}(x) = 0 \cdot x\} = \{x \in \mathbb{R} \mid \mathscr{L}^{\star\star}_{g,\nu,M}(x) = 0\}$. Since $\mathscr{L}_{g,\nu,M}$ is convex, non-negative, and 0 at 0, this subdifferential contains $\varepsilon \neq 0$ if and only if there exists $\varepsilon \neq 0$ with $\mathscr{L}_{g,\nu,M}(\varepsilon) = 0$. ■

The above proposition characterizes continuity in terms of the differentiability at 0 of the conjugate of the optimal lower bound function, or equivalently by Proposition 33, differentiability of the function $K_{g,\nu,\Xi}$. In the previous section we investigated in detail the finiteness (or equivalently by convexity, the continuity) of these functions around 0; in this section we show that continuity at 0 is equivalent to differentiability at 0 assuming that ϕ is not the constant 0 on a neighborhood of 1.

Proposition 65 *Assume that* $1 \notin \operatorname{int}\{x \in \mathbb{R} \mid \phi(x) = 0\}$*. Then for every σ-ideal Ξ, probability measure* $\nu \in \mathcal{M}^1_c(\Xi)$*, and* $g \in L^0(\Xi)$*, we have that* $0 \in \operatorname{int}\operatorname{dom} K_{g,\nu,\Xi}$ *if and only if* $K'_{g,\nu,\Xi}(0) = 0$.

Proof The if direction is immediate, since differentiability at 0 implies continuity at 0. Thus, for the remainder of the proof we assume that $K_{g,\nu,\Xi}$ is finite on a neighborhood of 0.

We first consider the case $\phi'(\infty) < \infty$, where Lemma 47 implies $g \in L^\infty(\Xi)$. Define $B \coloneqq \operatorname{ess\,sup}_\Xi |g|$, and let $\sigma \in \{-1, 1\}$ be such that $\phi(1 + \sigma x) > 0$ for all $x > 0$ as exists by assumption on ϕ. Since ψ is non-negative and 0 at 0, a standard characterization of the subdifferential (e.g. Zălinescu (2002, Theorem 2.4.2(iii))) implies that the function $t \mapsto \psi^\star(\sigma|t|)$ has derivative 0 at 0. Then for all $t \in \mathbb{R}$, by considering $\lambda = \sigma t B$ in (18), we obtain $K_{g,\nu,\Xi}(t)$ is at most $\nu(\psi^\star(tg + \sigma t B)) + \delta_{[-\infty,\phi'(\infty)]}(2\sigma|t|B) \leq \psi^\star(2\sigma|t|B) + \delta_{[-\infty,\phi'(\infty)]}(2\sigma|t|B)$. Now, if $\sigma = -1$ then $2\sigma|t|B \leq 0 \leq \phi'(\infty)$ for all t, and if $\sigma = 1$ then necessarily $\phi'(\infty) > 0$ and so $2\sigma|t|B \leq \phi'(\infty)$ for sufficiently small $|t|$. Thus, we have for sufficiently small $|t|$ that $K_{g,\nu,\Xi}(t)$ is between 0 and $\psi^\star(2\sigma|t|B)$, both of which are 0 with derivative 0 at 0, completing the proof in this case.

Now, assume that $\phi'(\infty) = \infty$, so that we have $K_{g,\nu,\Xi} = K_{g,\nu} = \inf_{\lambda\in\mathbb{R}} f(\cdot,\lambda)$ for $f(t,\lambda) \coloneqq \nu(\psi^\star(tg+\lambda))$. Note that $\psi \geq 0$ implies $f \geq 0$, so since $K_{g,\nu}(0) = f(0,0) = 0$ we have by standard results in convex analysis (e.g. Zălinescu (2002, Theorem 2.6.1(ii))) that $\partial K_{g,\nu}(0) = \{t^\star \mid (t^\star,0) \in \partial f(0,0)\}$. Furthermore, by assumption $K_{g,\nu}$ is finite on a neighborhood of 0, so since $K_{g,\nu} = K_{g+c,\nu}$ for all $c \in \mathbb{R}$, Lemma 50 implies $\operatorname{int}(\operatorname{dom} K_{g,\nu}) \times \mathbb{R} \subseteq \operatorname{dom} f$ and in particular $(0,0) \in \operatorname{int}\operatorname{dom} f$. Thus, defining for each $\omega \in \Omega$ the function $f_\omega(t,\lambda) \coloneqq \psi^\star(t\cdot g(\omega)+\lambda)$ (where here and in the rest of the proof we fix some representative $g \in \mathcal{L}^0(\Omega)$), standard results on convex integral functionals (e.g. Levin (1968, Theorem 1) or Ioffe and Tikhomirov (1969, Formula (7))) imply that $(t^\star,\lambda^\star) \in \partial f(0,0)$ if and only $(t^\star,\lambda^\star) = \big(\nu(t^\star_\omega),\nu(\lambda^\star_\omega)\big)$ for measurable functions $t^\star_\omega, \lambda^\star_\omega : \Omega \to \mathbb{R}$ such that $(t^\star_\omega,\lambda^\star_\omega) \in \partial f_\omega(0,0)$ holds ν-a.s.

Now, for each $\omega \in \Omega$, we have that $(t^\star_\omega,\lambda^\star_\omega) \in \partial f_\omega(0,0)$ if and only if $\psi^\star(t\cdot g(\omega)+\lambda) \geq t^\star_\omega \cdot t + \lambda^\star_\omega \cdot \lambda$ for all $(t,\lambda) \in \mathbb{R}^2$. By considering $t = 0$, this implies that $\lambda^\star_\omega \in \partial\psi^\star(0) = \{x \in \mathbb{R} \mid \psi(x) = 0\}$, which is contained in either $\mathbb{R}_{\geq 0}$ or $\mathbb{R}_{\leq 0}$ since ψ is not 0 on a neighborhood of 0. Then since the integral of a function of constant sign is zero if and only if it is zero almost surely, we have that $(t^\star,0) = \big(\nu(t^\star_\omega),\nu(\lambda^\star_\omega)\big)$ if and only if $\lambda^\star_\omega = 0$ holds ν-a.s. But $(t^\star_\omega,0) \in \partial f_\omega(0,0)$ if and only if for all $t \in \mathbb{R}$ we have $t^\star_\omega \cdot t \leq \inf_\lambda \psi^\star(t\cdot g(\omega)+\lambda) = \psi^\star(0) = 0$, i.e. if and only if $t^\star_\omega = 0$.

Putting this together, we get that $\partial K_{g,\nu}(0) = \{t^\star \mid (t^\star,0) \in \partial f(0,0)\} = \{\nu(t^\star_\omega) \mid (t^\star_\omega,0) \in \partial f_\omega(0,0)\ \nu\text{-a.s.}\} = \{\nu(t^\star_\omega) \mid t^\star_\omega = 0\ \nu\text{-a.s.}\} = \{0\}$ and $K'_{g,\nu}(0) = 0$ as desired. ■

Remark 66 *If ϕ is 0 on a neighborhood of 1, then it is easy to show that $K_{g,\nu}$ is not differentiable at 0 unless g is ν-essentially constant. Thus, the above proposition shows that the following are equivalent: (i) $1 \notin \operatorname{int}\{x \in \mathbb{R} \mid \phi(x) = 0\}$, (ii) for every g, continuity of $K_{g,\nu}$ at 0 implies differentiability at 0, (iii) $\mathrm{D}_\phi(\mu \parallel \nu) = 0$ for probability measures μ and ν if and only if $\mu = \nu$.*

A similar (but simpler) proof shows that the following are equivalent: (i) ϕ strictly convex at 1, (ii) for every g, continuity of $t \mapsto I_{\psi^\star,\nu}(tg)$ at 0 implies differentiability at 0, and (iii) $\mathrm{D}_\phi(\mu \parallel \nu) = 0$ for finite measures μ and ν if and only if $\mu = \nu$. The similarity of the statements in both cases suggest there may be a common proof of the equivalences using more general techniques in convex analysis.

Thus, combining the previous two propositions and Proposition 33 computing the convex conjugate of the optimal lower bound function, we obtain the following theorem.

Theorem 67 *Assume that $1 \notin \operatorname{int}(\{x \in \mathbb{R} \mid \phi(x) = 0\})$. Then for a σ-ideal Σ, $g \in L^0(\Sigma)$ and $\nu \in X^1_g$, writing $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X_g, |\mu|(A) = 0\}$, the following are equivalent:*

(i) for all $(\nu_n)_{n\in\mathbb{N}} \in \mathcal{M}^1_c(\Xi)^{\mathbb{N}}$, $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \parallel \nu) = 0$ implies that g is ν_n-integrable for sufficiently large n and $\lim_{n\to\infty} \nu_n(g) = \nu(g)$.

(ii) for all $(\nu_n)_{n\in\mathbb{N}} \in (X^1_g)^{\mathbb{N}}$, $\lim_{n\to\infty} \mathrm{D}_\phi(\nu_n \parallel \nu) = 0$ implies $\lim_{n\to\infty} \nu_n(g) = \nu(g)$.

(iii) $\partial K_{g,\nu,\Xi}(0) = \{0\}$, i.e. $K_{g,\nu,\Xi}$ is differentiable at 0 with $K'_{g,\nu,\Xi}(0) = 0$.

(iv) $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu,\Xi})$, that is, $g \in L^\infty(\Xi)$ when $\phi'(\infty) < \infty$ and $g \in S^\phi(\nu)$ when $\phi'(\infty) = \infty$.

Proof The equivalence of (ii)-(iv) is immediate from Propositions 64 and 65 since Proposition 33 implies $\mathscr{L}^\star_{g,\nu,X^1_g} = K_{g,\nu,\Xi}$. That (i) implies (ii) is immediate by definition of X^1_g. The reformulation of $0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu,\Xi})$ depending on the finiteness of $\phi'(\infty)$ uses Lemma 47 and Definition 49. Finally that (ii) and (iv) implies (i) is immediate when $\phi'(\infty) < \infty$—since every $\mu \in \mathcal{M}^1_c(\Xi)$ integrates every $g \in L^\infty(\Xi)$—and follows from Theorem 54 otherwise. ■

6. Optimal bounds relating ϕ-divergences and IPMs

In this section we generalize Theorem 40 on the optimal lower bound function for a single measure and function to the case of sets of measures and measurable functions.

6.1 On the choice of definitions

When considering a class of functions $\mathcal{G}$, there are several ways to define a lower bound of the divergence in terms of the mean deviation of functions in $\mathcal{G}$. The first one is to consider the IPM $d_{\mathcal{G}}$ induced by $\mathcal{G}$ and to ask for a function L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L\big(d_{\mathcal{G}}(\mu,\nu)\big)$ for all probability measures μ and ν, leading to the following definition of the optimal bound.

Definition 68 *Let $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ be a non-empty set of measurable functions and let $N, M \subseteq \mathcal{M}^1$ be two sets of probability measures such that $\mathcal{G} \subseteq L^1(\nu)$ for every $\nu \in N \cup M$. The optimal lower bound function $\mathscr{L}_{\mathcal{G},N,M} : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}_{\geq 0}$ is defined by*

$$\mathscr{L}_{\mathcal{G},N,M}(\varepsilon) \coloneqq \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \;\Big|\; (\nu,\mu) \in N \times M \;\wedge\; \sup_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big) = \varepsilon\Big\}\,.$$

We also for convenience extend the definition to the negative reals by

$$\mathscr{L}_{\mathcal{G},N,M}(\varepsilon) \coloneqq \mathscr{L}_{-\mathcal{G},N,M}(-\varepsilon) = \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \;\Big|\; (\nu,\mu) \in N \times M \;\wedge\; \inf_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big) = \varepsilon\Big\}$$

for $\varepsilon < 0$ where $-\mathcal{G} \coloneqq \{-g \mid g \in \mathcal{G}\}$.

Remark 69 *To motivate the definition of $\mathscr{L}_{\mathcal{G},N,M}$ on the negative reals, note that the equality $\sup_{g\in\mathcal{G}} \mu(g) - \nu(g) = \varepsilon$ for $\varepsilon \geq 0$ constrains by how "much above 0" an element of $\mathcal{G}$ can distinguish μ and ν, whereas the constraint $\inf_{g\in\mathcal{G}} \mu(g) - \nu(g) = -\varepsilon$ analogously constrains how much* below 0 *an element of $\mathcal{G}$ can distinguish them. When $\mathcal{G}$ is closed under negation, then $\sup_{g\in\mathcal{G}} \mu(g) - \nu(g) = d_{\mathcal{G}}(\mu,\nu) = -\inf_{g\in\mathcal{G}} \mu(g) - \nu(g)$ and $\mathscr{L}_{\mathcal{G},N,M}$ is even and exactly quantifies the smallest value taken by the ϕ-divergence given a constraint on the IPM defined by $\mathcal{G}$.*

An alternative definition, using the notations of Definition 68, is to consider the largest function L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(\mu(g) - \nu(g))$ for all $(\nu,\mu) \in N \times M$ and $g \in \mathcal{G}$. It is easy to see that this function can simply be expressed as

$$\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon) = \inf_{\substack{g\in\mathcal{G}\\ \nu\in N}} \mathscr{L}_{g,\nu,M}(\varepsilon) = \inf\Big\{\mathrm{D}_\phi(\mu \parallel \nu) \Big| (\nu,\mu,g) \in N \times M \times \mathcal{G} \wedge \mu(g) - \nu(g) = \varepsilon\Big\}\,.$$

Observe that $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M} = \mathscr{L}_{\mathcal{G},N,M}$ when $\mathcal{G}=\{g\}$ or $\mathcal{G}=\{-g,g\}$. More generally, the goal of this section is to explore the relationship between $\mathscr{L}_{\mathcal{G},N,M}$ and $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$. In particular, we will show that assuming a basic convexity condition on the set of measures M, these functions can differ only on their (at most countably many) discontinuity points.

Lemma 70 *Let $N, M\subseteq\mathcal{M}^1$ be two sets of probability measures with $N\subseteq M$ and M convex. Then the functions $\mathscr{L}_{\mathcal{G},N,M}$ and $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$ are non-negative, 0 at 0, and are non-decreasing on the non-negative reals.*

Proof It is sufficient to prove the result for $\mathscr{L}_{\mathcal{G},N,M}$, since the result for $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$ follows from the fact that taking infima preserves sign and monotonicity. Non-negativity and being 0 at 0 follow from non-negativity of $\mathrm{D}_\phi(\mu\parallel\nu)$ with $\mathrm{D}_\phi(\nu\parallel\nu)=0$.

Fix $0\le x\le y$ and consider $\alpha>\mathscr{L}_{\mathcal{G},N,M}(y)$, so that by definition there exist $\mu\in M$ and $\nu\in N$ with $\mathrm{D}_\phi(\mu\parallel\nu)<\alpha$ and $\sup_{g\in\mathcal{G}}\bigl(\mu(g)-\nu(g)\bigr)=y$. Define $\mu' = x/y\cdot\mu+(1-x/y)\cdot\nu$, which is a probability measure in M since $\nu\in N\subseteq M$ and M is convex. Then we have for every $g\in\mathcal{G}$ that $\mu'(g)-\nu(g)=x/y\cdot\bigl(\mu(g)-\nu(g)\bigr)$, and thus $\sup_{g\in\mathcal{G}}\bigl(\mu'(g)-\nu(g)\bigr)=x$. Furthermore, by convexity of $\mathrm{D}_{\phi,\nu}$ we have $\mathrm{D}_\phi(\mu'\parallel\nu)\le x/y\cdot\mathrm{D}_\phi(\mu\parallel\nu)+(1-x/y)\cdot\mathrm{D}_\phi(\nu\parallel\nu)< x/y\cdot\alpha\le\alpha$ since $x/y\le 1$. This implies that $\mathscr{L}_{\mathcal{G},N,M}(x)<\alpha$ and since α can be made arbitrarily close to $\mathscr{L}_{\mathcal{G},N,M}(y)$ that $\mathscr{L}_{\mathcal{G},N,M}(x)\le\mathscr{L}_{\mathcal{G},N,M}(y)$. ■

Remark 71 *For convex sets of measures M and N and a single function $g\in L^1(\nu)$, a simple adaptation of Lemma 32 shows that $\mathscr{L}_{g,N,M}$ is convex, non-decreasing, and non-negative on the non-negative reals. Lemma 70 extends the latter two properties to the case of $\mathscr{L}_{\mathcal{G},N,M}$ for a set of functions $\mathcal{G}$, and in fact its proof shows that $\mathscr{L}_{\mathcal{G},N,M}(y)/y$ is non-decreasing, which is necessary for convexity. It would be interesting to characterize the set of $\mathcal{G}$, N, and M for which $\mathscr{L}_{\mathcal{G},N,M}$ and $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$ are in fact convex.*

Proposition 72 *Under the assumptions of Lemma 70, we have for every $\varepsilon>0$ that*

$$\lim_{\varepsilon'\to\varepsilon^-}\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}(\varepsilon')\le\mathscr{L}_{\mathcal{G},N,M}(\varepsilon)\le\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}(\varepsilon)\,,$$

with equality if $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$ is lower semicontinuous (equivalently left-continuous) at ε or if $\mathcal{G}$ is compact in the initial topology on $\mathcal{L}^0$ induced by the maps $\langle\mu-\nu,\,\cdot\,\rangle$ for $\mu\in M$ and $\nu\in N$.

Proof Under the assumptions of Lemma 70 we have $\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}$ and $\mathscr{L}_{\mathcal{G},N,M}$ are non-decreasing on the positive reals. Thus, we have

$$\begin{aligned}
\inf_{g\in\mathcal{G}}\mathscr{L}_{g,N,M}(\varepsilon) &= \inf\Bigl\{\mathrm{D}_\phi(\mu\parallel\nu)\Bigm|(\nu,\mu)\in N\times M\ \wedge\ \exists g\in\mathcal{G},\mu(g)-\nu(g)=\varepsilon\Bigr\}\\
&\ge\inf\Bigl\{\mathrm{D}_\phi(\mu\parallel\nu)\Bigm|(\nu,\mu)\in N\times M\ \wedge\ \sup_{g\in\mathcal{G}}\bigl(\mu(g)-\nu(g)\bigr)\ge\varepsilon\Bigr\} &(22)\\
&=\inf_{\varepsilon'\ge\varepsilon}\mathscr{L}_{\mathcal{G},N,M}(\varepsilon')=\mathscr{L}_{\mathcal{G},N,M}(\varepsilon) &(23)\\
&=\inf\Bigl\{\mathrm{D}_\phi(\mu\parallel\nu)\Bigm|(\nu,\mu)\in N\times M\ \wedge\ \forall\varepsilon'<\varepsilon\ \exists g\in\mathcal{G},\ \mu(g)-\nu(g)\ge\varepsilon'\Bigr\}
\end{aligned}$$

$$\begin{aligned}
&\geq \sup_{\varepsilon'<\varepsilon} \inf\Big\{ \mathrm{D}_\phi(\mu \parallel \nu) \,\Big|\, (\nu,\mu) \in N \times M \,\wedge\, \exists g \in \mathcal{G}, \mu(g) - \nu(g) \geq \varepsilon' \Big\} \\
&= \sup_{\varepsilon'<\varepsilon} \inf\Big\{ \mathrm{D}_\phi(\mu \parallel \nu) \,\Big|\, (\nu,\mu,g) \in N \times M \times \mathcal{G} \,\wedge\, \mu(g) - \nu(g) \geq \varepsilon' \Big\} \\
&= \sup_{\varepsilon'<\varepsilon} \inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon') = \lim_{\varepsilon'\to\varepsilon^-} \inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon') \qquad (24)
\end{aligned}$$

where Eq. (22) is since if there is $g \in \mathcal{G}$ with $\mu(g) - \nu(g) = \varepsilon$ then $\sup_{g\in\mathcal{G}} \mu(g) - \nu(g) \geq \varepsilon$, Eq. (23) is because $\mathscr{L}_{\mathcal{G},N,M}$ is non-decreasing, and Eq. (24) is because $\inf_{g\in G} \mathscr{L}_{g,N,M}$ is non-decreasing.

For the equality claims, since $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$ is non-decreasing, lower semicontinuity at ε is equivalent to left-continuity, and $\lim_{\varepsilon'\to\varepsilon^-} \inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}(\varepsilon') = \inf_{g\in G} \mathscr{L}_{g,N,M}(\varepsilon)$ in this case. If $\mathcal{G}$ is compact in the claimed topology, then $\sup_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big)$ is the supremum of the continuous function $\langle \mu - \nu, \,\cdot\, \rangle$ on the compact set $\mathcal{G}$, so that $\sup_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big) = \max_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big)$ and thus Eq. (22) is an equality. ■

Corollary 73 *Under the assumptions of Lemma 70 we have that* $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$ *and* $\mathscr{L}_{\mathcal{G},N,M}$ *are non-increasing on the non-positive reals, non-decreasing on the non-negative reals,* 0 *at* 0, *and differ only on their (at most countably many) discontinuity points, at which* $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M} \geq \mathscr{L}_{\mathcal{G},N,M}$. *In particular, they have the same convex conjugate and biconjugate.*

Proof Applying Proposition 72 to $\mathscr{L}_{-\mathcal{G},N,M}(-\varepsilon)$ for $\varepsilon < 0$ gives the claim for the negative reals. Since the functions share the same lsc regularization (the largest lsc function lower bounding them pointwise), they also share their convex conjugate and biconjugate. ■

Remark 74 *Corollary 73 is key because, as we will see in Section 6.2, it lets us reduce the problem of computing the optimal lower bound on an IPM to the case of a single function g considered in Section 5.*

Remark 75 *Corollary 73 also implies that* $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$ *and* $\mathscr{L}_{\mathcal{G},N,M}$ *have the same* (generalized) inverse. *This inverse consists simply of the best bounds on the mean deviation, that is the largest non-positive function V and smallest non-negative function U such that* $V\big(\mathrm{D}_\phi(\mu \parallel \nu)\big) \leq \mu(g) - \nu(g) \leq U\big(\mathrm{D}_\phi(\mu \parallel \nu)\big)$ *for all* $(\mu,\nu,g) \in M \times N \times \mathcal{G}$, *or equivalently such that* $d_{\mathcal{G}}(\mu,\nu) \leq U\big(\mathrm{D}_\phi(\mu \parallel \nu)\big)$ *for all* $(\mu,\nu) \in M \times N$ *when* $\mathcal{G}$ *is closed under negation. In this language, any discontinuity of* $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,M}$ *corresponds to an interval on which U is constant, i.e. in which changing the value of the divergence does not change the largest possible value of* $d_{\mathcal{G}}(\mu,\nu)$.

We conclude this section with two lemmas showing how the lower bound is preserved under natural transformations of the sets of functions $\mathcal{G}$ or measures M, N.

Lemma 76 *For every set* $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ *and pair of measures* $\mu, \nu \in X_{\mathcal{G}}$, *we have that*

$$\sup_{g\in\mathcal{G}}\big(\mu(g) - \nu(g)\big) = \sup_{g\in\overline{\operatorname{co}\mathcal{G}}}\big(\mu(g) - \nu(g)\big)$$

where $\overline{\operatorname{co}\mathcal{G}}$ *is the* $\sigma(Y_{\mathcal{G}}, X_{\mathcal{G}})$*-closed convex hull of* $\mathcal{G}$.

Proof We have $\mathcal{G} \subseteq \overline{\operatorname{co} \mathcal{G}}$, and furthermore since $\langle \mu - \nu, \,\cdot\, \rangle$ is a $\sigma(Y_{\mathcal{G}}, X_{\mathcal{G}})$-continuous linear function we have that the set $\big\{h \in Y_{\mathcal{G}} \;\big|\; \langle \mu - \nu, h\rangle \le \sup_{g\in\mathcal{G}}(\mu(g) - \nu(g))\big\}$ is convex, $\sigma(Y_{\mathcal{G}}, X_{\mathcal{G}})$-closed, and contains $\mathcal{G}$, and so also contains $\overline{\operatorname{co} \mathcal{G}}$. ■

Lemma 77 *For every $g \in \mathcal{L}^0(\Omega)$, we have $\mathscr{L}_{g,X_g^1,X_g^1} = \mathscr{L}_{\mathrm{Id}_{\mathbb{R}},g_*X_g^1,g_*X_g^1}$ where $g_*X_g^1 = \big\{g_*\nu \;\big|\; \nu \in X_g^1\big\}$ is the set of probability measures on $\mathbb{R}$ obtained by pushing forward through g the probability measures $\nu \in X_g^1$. Furthermore, for every $\nu \in \mathcal{M}^1$ and $g \in L^1(\nu)$ we have that $\mathscr{L}_{g,\nu,X_g^1} = \mathscr{L}_{\mathrm{Id}_{\mathbb{R}},g_*\nu,g_*X_g^1}$.*

Proof We first prove the main claim. As in Example 2, we have for every $\mu, \nu \in X_g^1$ that $\mu(g)-\nu(g) = \int \mathrm{Id}_{\mathbb{R}}\, \mathrm{d}g_*\mu - \int \mathrm{Id}_{\mathbb{R}}\, \mathrm{d}g_*\nu$, so it suffices to show for every $\mu_0, \nu_0 \in X_g^1$ the existence of $\mu, \nu \in X_g^1$ with $g_*\mu = g_*\mu_0$, $g_*\nu = g_*\nu_0$, and $\mathrm{D}_\phi(g_*\mu_0 \parallel g_*\nu_0) = \mathrm{D}_\phi(\mu \parallel \nu) \le \mathrm{D}_\phi(\mu_0 \parallel \nu_0)$.

For this, write $\xi = \frac{1}{2}(\mu_0 + \nu_0)$ so that $\mu_0, \nu_0 \ll \xi$ and $\xi \in X_g^1$, and define the measures $\mu, \nu \in \mathcal{M}_c^1(\xi)$ by $\frac{d\mu}{d\xi} = \frac{dg_*\mu_0}{dg_*\xi} \circ g$ and $\frac{d\nu}{d\xi} = \frac{dg_*\nu_0}{dg_*\xi} \circ g$ (note that these are just the conditional expectations of $\frac{d\mu_0}{d\xi}$ and $\frac{d\nu_0}{d\xi}$ with respect to g). It remains to show that μ and ν have the desired properties, for which we first note that for every (Borel) measurable function $h : \mathbb{R}^3 \to \mathbb{R} \cup \{+\infty\}$ we have

$$\begin{aligned}\int h\bigg(\frac{d\mu}{d\xi}, \frac{d\nu}{d\xi}, g\bigg)\, \mathrm{d}\xi &= \int h\bigg(\frac{dg_*\mu_0}{dg_*\xi} \circ g, \frac{dg_*\nu_0}{dg_*\xi} \circ g, g\bigg)\, \mathrm{d}\xi \\ &= \int h\bigg(\frac{dg_*\mu_0}{dg_*\xi}, \frac{dg_*\nu_0}{dg_*\xi}, \mathrm{Id}_{\mathbb{R}}\bigg)\, \mathrm{d}g_*\xi\,.\end{aligned}$$

Then taking $h(x, y, z) = x$ we get $\mu(\Omega) = \mu(\mathbf{1}_\Omega) = g_*\mu_0(\mathbf{1}_{\mathbb{R}}) = \mu_0(\mathbf{1}_\Omega) = 1$, and similarly by taking $h(x, y, z) = y$ we get $\nu(\Omega) = 1$. Taking $h(x, y, z) = x \cdot |z|$ we get $\mu(|g|) = \mu_0(|g|) < \infty$ so that $\mu \in X_g^1$, and similarly by taking $h(x, y, z) = y \cdot |z|$ we get $\nu(|g|) = \nu_0(|g|) < \infty$ and $\nu \in X_g^1$. Finally, as in Remark 19, taking $h(x, y, z) = y \cdot \phi(x/y)$ if $y \ne 0$ and $h(x, y, z) = x \cdot \phi'(\infty)$ if $y = 0$ gives $\mathrm{D}_\phi(\mu \parallel \nu) = \mathrm{D}_\phi(g_*\mu_0 \parallel g_*\nu_0)$, and furthermore Jensen's inequality implies that $\mathrm{D}_\phi(\mu \parallel \nu) \le \mathrm{D}_\phi(\mu_0 \parallel \nu_0)$ since h is convex.

The furthermore claim is analogous after noting that since when $\mu \ll \nu$ and $g \in L^1(\nu)$ we can take $\xi = \nu_0 = \nu$. ■

6.2 Derivation of the bound

In this section we give our main results computing optimal lower bounds on a ϕ-divergence given an integral probability metric. Note that from Section 6.1, the optimal lower bound is simply the infimum of the optimal lower bound $\mathscr{L}_{g,\nu}$ for each $g \in \mathcal{G}$ and $\nu \in N$. Since $\mathscr{L}_{g,\nu}^\star = K_{g,\nu}$ by Proposition 33, and given the order-reversing property of convex conjugacy, it is natural to consider the best *upper bound* on $K_{g,\nu}$ which holds *uniformly* over all $g \in \mathcal{G}$ and $\nu \in N$. Formally, we have the following definition.

Definition 78 *Let Ξ be a σ-ideal, $\mathcal{G} \subseteq L^0(\Xi)$ be a set of measurable functions, and $N \subseteq \mathcal{M}_c^1(\Xi)$ be a set of measures. We write $K_{\mathcal{G},N,\Xi}(t) \coloneqq \sup\{K_{g,\nu,\Xi}(t) \mid (g, \nu) \in \mathcal{G} \times N\}$ and $K_{\mathcal{G},N}(t) \coloneqq \sup\{K_{g,\nu}(t) \mid (g, \nu) \in \mathcal{G} \times N\}$.*

Note that $K_{\mathcal{G},N,\Xi}$ is convex and lower semicontinuous as a supremum of convex and lower semicontinuous functions. Furthermore, as alluded to before Definition 78, we expect $K_{\mathcal{G},N,\Xi}$ to be equal to the conjugate of the optimal lower bound functions. This is stated formally in the following theorem which also gives a sufficient condition under which the optimal lower bound functions are convex and lower semicontinuous (see also Remark 81 below).

Theorem 79 *Let (X, Y) be a dual pair with $X \subseteq \mathcal{M}$ and $Y \subseteq L^0(\Xi)$ where $\Xi = \{A \in \mathcal{F} \mid \forall \mu \in X, |\mu|(A) = 0\}$. Consider $\mathcal{G} \subseteq Y$ and $N \subseteq X^1 := X \cap \mathcal{M}^1$ and assume that (X, Y) is decomposable with respect to all measures in N. Then we have*

$$\mathscr{L}^\star_{\mathcal{G},N,X^1} = \left(\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}\right)^\star = K_{\mathcal{G},N,\Xi}\,. \tag{25}$$

Proof The first equality in (25) follows from Corollary 73. For the second equality, we have

$$\left(\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}\right)^\star = \left(\inf_{\substack{g\in\mathcal{G}\\ \nu\in N}} \mathscr{L}_{g,\nu,X^1}\right)^\star = \sup_{\substack{g\in\mathcal{G}\\ \nu\in N}} \mathscr{L}^\star_{g,\nu,X^1} = \sup_{\substack{g\in\mathcal{G}\\ \nu\in N}} K_{g,\nu,\Xi} = K_{\mathcal{G},N,\Xi}\,,$$

where we used successively the definition of $\mathscr{L}_{g,N,X^1}$, the fact that $(\inf_{\alpha\in A} f_\alpha)^\star = \sup_{\alpha\in A} f^\star_\alpha$ for any collection $(f_\alpha)_{\alpha\in A}$ of functions, Proposition 33, and Definition 78. ■

Remark 80 *When starting from a set of function $\mathcal{G} \subseteq L^0(\Xi)$ for some σ-ideal Ξ, a natural pair to which Theorem 79 can be applied is the pair $(X_\mathcal{G}, Y_\mathcal{G})$ from Definition 41.*

Remark 81 *Theorem 79 computes the conjugate of the optimal lower bound functions, but if this function is not convex or lsc and so does not coincide with its biconjugate, it is also useful to discuss what we can be said about $\mathscr{L}_{\mathcal{G},N,X^1}$ or $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}$ themselves.*

First note that for all $g \in \mathcal{G}$ and $\nu \in N$, $\mathscr{L}_{g,\nu,X^1}$ is convex and non-decreasing over the non-negative reals by Lemma 32 and under the assumptions of Theorem 79, $\mathscr{L}^{\star\star}_{g,\nu,X^1} = K^\star_{g,\nu,\Xi}$ by Proposition 33. Thus, we can apply Lemma 14 and obtain for all ε that

$$\liminf_{\varepsilon'\to\varepsilon} \inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}(\varepsilon') \le \inf_{\substack{g\in\mathcal{G}\\ \nu\in N}} K^\star_{g,\nu,\Xi}(\varepsilon) \le \inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}(\varepsilon)\,.$$

Thus the function $\inf_{(g,\nu)\in\mathcal{G}\times N} K^\star_{g,\nu,\Xi}$ allows us to recover the function $\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}$ up to its points of discontinuity which are countable by monotonicity. Similarly, by Corollary 73 we also recover $\mathscr{L}_{\mathcal{G},N,X^1}$ up to its countably many points of discontinuity.

More can be said under additional assumptions. If $\mathscr{L}_{g,\nu,X^1}$ is lower semicontinuous for each $g \in \mathcal{G}$ and $\nu \in N$ (e.g. when $\phi'(\infty) = \infty$, $X \subseteq \mathcal{M}_c(\nu)$, and $\mathcal{G} \subseteq S^\phi_\heartsuit(\nu)$ for all $\nu \in N$ by Corollary 62), then

$$\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}(\varepsilon) = \inf_{\substack{g\in\mathcal{G}\\ \nu\in N}} \mathscr{L}_{g,\nu,X^1}(\varepsilon) = \inf_{\substack{g\in\mathcal{G}\\ \nu\in N}} K^\star_{g,\nu,\Xi}(\varepsilon)\,,$$

Furthermore, if we also know that the function $\inf_{(g,\nu)\in\mathcal{G}\times N} K^\star_{g,\nu,\Xi}$ is itself convex and lsc, then

$$\inf_{g\in\mathcal{G}} \mathscr{L}_{g,N,X^1}(\varepsilon) = \mathscr{L}_{\mathcal{G},N,X^1}(\varepsilon) = K^\star_{\mathcal{G},N,\Xi}(\varepsilon)\,.$$

Similarly to Corollary 44, we give in the following corollary an "operational" restatement of Theorem 79 emphasizing the duality between upper bounds on $K_{\mathcal{G},N}$ and lower bounds on $\mathrm{D}_\phi(\mu \parallel \nu)$ in terms of $d_\mathcal{G}(\mu, \nu)$.

Corollary 82 *Under the assumptions of Theorem 79, for every convex and lower semicontinuous function $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$, the following are equivalent:*

(i) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(d_\mathcal{G}(\mu, \nu))$ *for all* $\nu \in N$ *and* $\mu \in X^1$.

(ii) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(|\mu(g) - \nu(g)|)$ *for all* $g \in \mathcal{G}$, $\nu \in N$, *and* $\mu \in X^1$.

(iii) $K_{g,\nu,\Xi}(t) \leq L^\star(|t|)$ *for all* $t \in \mathbb{R}$, $g \in \mathcal{G}$, *and* $\nu \in N$.

Proof The equivalence of (i) and (iii) follows from applying Theorem 79 to $\mathcal{G}' = \mathcal{G} \cup -\mathcal{G}$, since $d_\mathcal{G}(\mu, \nu) = \sup_{g \in \mathcal{G}'} \mu(g) - \nu(g) \geq 0$, $\mathscr{L}_{\mathcal{G}',N,X^1}$ is even, and $K_{\mathcal{G}',N,\Xi}(t) = \max\{K_{\mathcal{G},N,\Xi}(t), K_{\mathcal{G},N,\Xi}(-t)\}$. The equivalence of (i) and (iii) for $\{g, -g\}$ for each $g \in \mathcal{G}$ gives the equivalence of (ii) and (iii). ■

Example 13 (Subgaussian functions) *For the Kullback–Leibler divergence, Boucheron et al. (2013, Lemma 4.18) shows that* $\mathrm{D}(\mu \parallel \nu) \geq \frac{1}{2} d_\mathcal{G}(\mu, \nu)^2$ *for all* $\mu \in \mathcal{M}^1$ *if and only if* $\log \int e^{t(g - \nu(g))} \, \mathrm{d}\nu \leq t^2/2$ *for all* $g \in \mathcal{G}$ *and* $t \in \mathbb{R}$. *Such a quadratic upper bound on the log moment-generating function is one of the characterizations of the so-called* subgaussian *functions, which contain as a special case the class of bounded functions by Hoeffding's lemma (Hoeffding, 1963) (see also Example 21). Corollary 82 recovers this result by considering the (self-conjugate) function* $L : t \mapsto t^2/2$, *thus showing that Pinsker's inequality generalize to all subgaussian functions.*

Theorem 79 generalizes this further to an arbitrary ϕ-divergence, showing that a subset $\mathcal{G} \subseteq \mathcal{L}^0(\Omega)$ *of measurable functions satisfies* $\mathrm{D}_\phi(\mu \parallel \nu) \geq \frac{1}{2} d_\mathcal{G}(\mu, \nu)^2$ *for all* $\mu \in \mathcal{M}^1$ *if and only if* $K_{g,\nu}(t) \leq t^2/2$ *for all* $g \in \mathcal{G}$ *and* $t \in \mathbb{R}$. *By analogy, we refer to functions whose cumulant generating function admits such a quadratic upper bound as* ϕ-subgaussian *functions.*

Example 14 *Recall from Example 10 that the χ^2-divergence given by* $\phi(x) = (x-1)^2 + \delta_{\mathbb{R}_{\geq 0}}(x)$ *satisfies*

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

and $K_{g,\nu}(t) \leq \inf_\lambda \int (tg + \lambda)^2/4 \, \mathrm{d}\nu = t^2 \operatorname{Var}_\nu(g)/4$, *showing that the class of χ^2-subgaussian functions (see Example 13) includes all those with bounded variance.*

Example 15 *As a step towards understanding the Wasserstein distance, Bolley and Villani (2005) define a "weighted total variation distance" between probability measures μ and ν as* $\int g \, \mathrm{d}|\mu - \nu|$ *for some non-negative measurable function* $g \in \mathcal{L}^0(\Omega)$, *and their main result (Bolley and Villani, 2005, Theorem 2.1) bounds this weighted total variation in terms of the KL divergence.*

We rederive their result by noting that the g-weighted total variation is $d_{g\mathcal{B}}(\mu, \nu)$ *for* $g\mathcal{B} = \{g \cdot b \,|\, b \in \mathcal{B}\}$ *where $\mathcal{B}$ is the set of measurable functions taking values in* $[-1, 1]$, *so that* *it suffices by Theorem 79 to upper bound $K_{g\cdot b,\nu}(t)$ for each $b \in \mathcal{B}$ in terms of $\log \int e^g \, \mathrm{d}\nu$ or $\log \int e^{g^2} \, \mathrm{d}\nu$. But since $g \geq 0$, we have $g \cdot b \leq |g| = g$ and we conclude by using the fact that finiteness of $\log \int e^h \, \mathrm{d}\nu$ (resp. $\log \int e^{h^2} \, \mathrm{d}\nu$) implies a quadratic upper bound on the centered log-moment generating function $K_{h,\nu}(t)$ for $|t| \leq 1/4$ (resp. all $t \in \mathbb{R}$) for any non-negative function h (see e.g. Vershynin (2018, Propositions 2.5.2 and 2.7.1)).*

Finally, we show that when we take $N = X^1$ in Theorem 79, that is, we want a lower bound L such that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(d_{\mathcal{G}}(\mu, \nu))$ for all probability measures μ and ν in X, we no longer need to consider pairs of measures such that $\mu \not\ll \nu$, and in particular we can ignore the σ-ideal Ξ in the derivation of the bound. Intuitively, this is because when $\mu \not\ll \nu$, we now have sufficiently many measures in N to approximate ν with a measure ν' such that $\mu \ll \nu'$.

Theorem 83 *Let (X, Y) be a dual pair with $X \subseteq \mathcal{M}$ and assume that (X, Y) is decomposable with respect to all probability measures in X. Then for all subsets of functions $\mathcal{G} \subseteq Y$,*

$$\mathscr{L}^\star_{\mathcal{G},X^1,X^1} = \left(\inf_{g \in \mathcal{G}} \mathscr{L}_{g,X^1,X^1}\right)^\star = K_{\mathcal{G},X^1}\,.$$

In particular, for any σ-ideal Σ and $\mathcal{G} \subseteq L^0(\Sigma)$, the following are equivalent for every convex lsc $L : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$:

(i) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(d_{\mathcal{G}}(\mu, \nu))$ *for all $\mu, \nu \in \mathcal{M}_c(\Sigma)$ integrating all of $\mathcal{G}$.*

(ii) $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(|\mu(g) - \nu(g)|)$ *for all $g \in \mathcal{G}$ and $\mu, \nu \in \mathcal{M}_c(\Sigma)$ integrating all of $\mathcal{G}$.*

(iii) $K_{g,\nu}(t) \leq L^\star(|t|)$ *for all $t \in \mathbb{R}$, $g \in \mathcal{G}$, and $\nu \in \mathcal{M}_c(\Sigma)$ integrating all of $\mathcal{G}$.*

Proof The in particular claim follows from the main claim applied to $(X_{\mathcal{G}}, Y_{\mathcal{G}})$ by an argument analogous to that of Corollary 82. For the main claim, by Theorem 79, it suffices to show that $\inf_{g \in \mathcal{G}} \mathscr{L}_{g,X^1,X^1} = \inf_{g \in \mathcal{G}, \nu \in X^1} \mathscr{L}_{g,\nu,X^1}$ and $\inf_{g \in \mathcal{G}, \nu \in X^1} \mathscr{L}_{g,\nu,X^1 \cap \mathcal{M}_c(\nu)}$ have the same conjugate, or simply the same lsc regularization. Since the former is definitionally no larger than the latter, it suffices to show that any lsc lower bound L for the latter also lower bounds the former, equivalently, that if $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(\mu(g) - \nu(g))$ for all $\mu \ll \nu \in X^1$ and $g \in \mathcal{G}$, then this also holds for all $\mu, \nu \in X^1$.

Given any $\mu, \nu \in X^1$ and $\delta \in [0, 1]$, let $\nu_\delta = (1 - \delta) \cdot \nu + \delta \cdot \mu$ so that $\nu_\delta \in X^1$. Then for each $\delta \in [0, 1]$ we have that $\mu(g) - \nu_\delta(g) = (1 - \delta)(\mu(g) - \nu(g))$ for all $g \in \mathcal{G}$, and furthermore, by convexity of $\mathrm{D}_\phi(\mu \parallel \cdot\,)$ we have for $\delta \in (0, 1]$ that

$$(1 - \delta)\,\mathrm{D}_\phi(\mu \parallel \nu) = (1 - \delta)\,\mathrm{D}_\phi(\mu \parallel \nu) + \delta\,\mathrm{D}_\phi(\mu \parallel \mu) \geq \mathrm{D}_\phi(\mu \parallel \nu_\delta) \geq L\big((1 - \delta)(\mu(g) - \nu(g))\big)$$

where the last inequality is because $\mu \ll \nu_\delta$. But since L is lower semicontinuous, we have that $L(\mu(g) - \nu(g)) \leq \liminf_{\delta \to 0} L\big((1 - \delta)(\mu(g) - \nu(g))\big) \leq \lim_{\delta \to 0^-} L\big((1 - \delta)(\mu(g) - \nu(g))\big)$, and so we get that $\mathrm{D}_\phi(\mu \parallel \nu) \geq L(\mu(g) - \nu(g))$ as desired. ■

6.3 Application to bounded functions and the total variation

In this section, we consider the problem of lower bounding the ϕ-divergence by a function of the total variation distance. Though it is a well-studied problem and most of the results we derive are already known, we consider this case to demonstrate the applicability of the results obtained in Section 6.2. In Section 6.3.1, we study Vajda's problem (Vajda, 1972): obtaining the best lower bound of the ϕ-divergence by a function of the total variation distance, and in Section 6.3.2 we show how to obtain quadratic relaxations of the best lower bound as in Pinsker's inequality and Hoeffding's lemma.

6.3.1 Vajda's problem

The Vajda problem (Vajda, 1972) is to quantify the optimal relationship between the ϕ-divergence and the total variation, that is to compute the function

$$\begin{aligned}\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(\varepsilon) &= \inf\bigl\{\mathrm{D}_\phi(\mu \parallel \nu) \bigm| (\mu,\nu) \in \mathcal{M}^1 \times \mathcal{M}^1 \,\wedge\, \mathrm{TV}(\mu,\nu) = \varepsilon\bigr\}\\ &= \inf\bigl\{\mathrm{D}_\phi(\mu \parallel \nu) \bigm| (\mu,\nu) \in \mathcal{M}^1 \times \mathcal{M}^1 \,\wedge\, d_{\mathcal{B}}(\mu,\nu) = \varepsilon\bigr\}\end{aligned}$$

where $\mathcal{B}$ is the set of measurable functions $\Omega \to [-1,1]$. In this section, we use Theorem 83 to give for an arbitrary ϕ an expression for the Vajda function as the convex conjugate of a natural geometric quantity associated with the function $\psi^\star$, the inverse of its *sublevel set volume function*, which we call the *height-for-width* function.

Definition 84 *The* sublevel set volume *function* $\mathrm{sls}_{\psi^\star} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *maps* $h \in \mathbb{R}$ *to the Lebesgue measure of the sublevel set* $\{x \in \mathbb{R} \mid \psi^\star(x) \leq h\}$*. Since* $\psi^\star$ *is convex and inf-compact, the sublevel sets are compact intervals and their Lebesgue measure is simply their length.*

The height-for-width *function* $\mathrm{hgt}_{\psi^\star} : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$ *is the (right) inverse of the sublevel set volume function given by* $\mathrm{hgt}_{\psi^\star}(w) = \inf\bigl\{h \in \overline{\mathbb{R}} \bigm| \mathrm{sls}_{\psi^\star}(h) \geq w\bigr\}$*.*

To understand this definition, note that since $\psi^\star$ is defined on $\mathbb{R}$, the sublevel set volume function can be interpreted as giving for each height h the length of longest horizontal line segment that can be placed in the epigraph of $\psi^\star$ but no higher than h. The inverse, the height-for-width function, asks for the minimal height at which one can place a horizontal line segment of length w in the epigraph of $\psi^\star$. See Fig. 1 for an illustration of this in the case of $\psi^\star(x) = e^x - x - 1$, corresponding to the Kullback–Leibler divergence.

The following lemma shows that the height-for-width function can be equivalently formulated as the optimal value of a simple convex optimization problem.

Lemma 85 *For all* $w \in \mathbb{R}_{\geq 0}$*,* $\mathrm{hgt}_{\psi^\star}(w) = \inf_{\lambda \in \mathbb{R}} \max\{\psi^\star(\lambda + w/2), \psi^\star(\lambda - w/2)\}$*. Furthermore, if for* $w > 0$ *there exists* λ_w *such that* $\psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$*, then* $\mathrm{hgt}_{\psi^\star}(w) = \psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$*.*

Proof For every $w \geq 0$, define the function $h_w : \lambda \mapsto \max\{\psi^\star(\lambda - w/2), \psi^\star(\lambda + w/2)\}$ which is the supremum of two convex inf-compact functions with overlapping domain, and so is itself proper, convex, and inf-compact. In particular, h_w achieves its global minimum $y_w \in \mathbb{R}$, where by definition and convexity of $\psi^\star$ we have y_w is the smallest number such that there exists an interval $[\lambda - w/2, \lambda + w/2]$ of length w such that $\psi^\star([\lambda - w/2, \lambda + w/2]) \subseteq (-\infty, y_w]$, and thus $y_w = \inf\bigl\{x \in \overline{\mathbb{R}} \bigm| \mathrm{sls}_{\psi^\star}(x) \geq w\bigr\} = \mathrm{hgt}_{\psi^\star}(w)$ as desired.

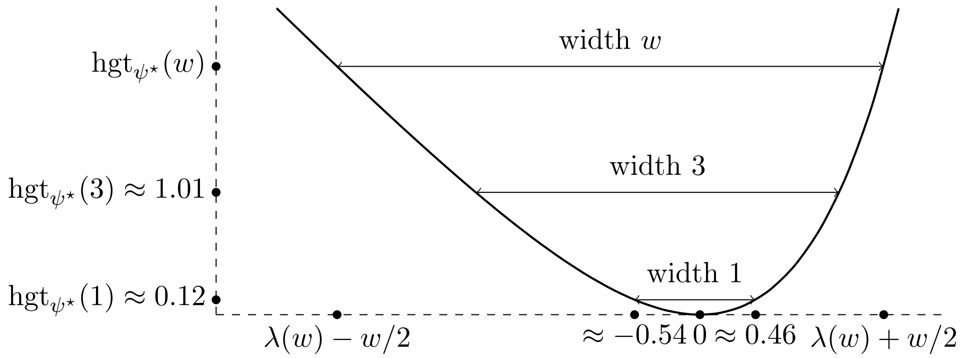

Figure 1: Illustration of height-for-width function for $\psi^\star(x) = e^x - x - 1$

For the remaining claim, consider $w > 0$ for which there is $\lambda_w \in \mathbb{R}$ such that $\psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$. By convexity of $\psi^\star$ we have for every $\lambda < \lambda_w$ that $\psi^\star(\lambda - w/2) \geq \psi^\star(\lambda_w - w/2)$, and analogously for every $\lambda > \lambda_w$ that $\psi^\star(\lambda + w/2) \geq \psi^\star(\lambda_w + w/2)$. Thus for every λ we have $\max\{\psi^\star(\lambda - w/2), \psi^\star(\lambda + w/2)\} \geq \min\{\psi^\star(\lambda_w - w/2), \psi^\star(\lambda_w + w/2)\} = \psi^\star(\lambda_w - w/2) = \psi^\star(\lambda_w + w/2)$, so the result follows from the main claim. ■

Example 16 *For the case of the KL divergence for which $\psi^\star(w) = e^w - w - 1$, one can compute that $\psi^\star(\lambda(w) + w/2) = \psi^\star(\lambda(w) - w/2)$ for $\lambda(w) = -\log \frac{e^{w/2} - e^{-w/2}}{w} = -\log \frac{2\sinh(w/2)}{w}$, so that $\mathrm{hgt}_{\psi^\star}(w) = -1 + \frac{w}{2} \coth \frac{w}{2} + \log \frac{2\sinh(w/2)}{w}$.*

The duality result of Theorem 83 computes the biconjugate of the optimal bound $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}$, so we first prove that this function is convex and lsc.

Lemma 86 *Let M the set of probability measures supported on $\{-1, 1\}$. Then $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M}$ is convex and lower semicontinuous. In particular $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(\varepsilon) = K^\star_{\mathcal{B},\mathcal{M}^1}(\varepsilon)$ for $\varepsilon \geq 0$.*

Proof By Theorem 83 we have that $\mathscr{L}^\star_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = K_{\mathcal{B},\mathcal{M}^1}$, so the in particular statement follows immediately from the main claim. The main claim, that $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M}$ is convex and lower semicontinuous, is well-known and can easily be derived using the methods of e.g. Vajda (1972), but we include a proof here in our language for completeness and to illustrate how it could be generalized beyond the total variation.

Note that the set $\mathcal{B} = [-1, 1]^\Omega \cap \mathcal{L}^0(\Omega)$ is convex, and furthermore is $\sigma(\mathcal{L}^b(\Omega), \mathcal{M})$-compact by the Banach–Alaoglu theorem, and so by the Krein–Milman theorem $\mathcal{B}$ is the $\sigma(\mathcal{L}^b(\Omega), \mathcal{M})$-closed convex hull of its extreme points $\mathrm{ext}(\mathcal{B}) = \{-1, 1\}^\Omega \cap \mathcal{L}^0(\Omega)$ the set of measurable $\{-1, 1\}$-valued functions. Thus, Lemma 76 implies $d_{\mathcal{B}} = d_{\mathrm{ext}(\mathcal{B})}$, and so $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}_{\mathrm{ext}(\mathcal{B}),\mathcal{M}^1,\mathcal{M}^1}$.

We now prove that $\inf_{g \in \mathrm{ext}(\mathcal{B})} \mathscr{L}_{g,\mathcal{M}^1}$ is convex and lsc, which by Proposition 73 also implies $\mathscr{L}_{\mathrm{ext}(\mathcal{B}),\mathcal{M}^1,\mathcal{M}^1} = \inf_{g \in \mathrm{ext}(\mathcal{B})} \mathscr{L}_{g,\mathcal{M}^1}$ is convex and lsc. By Lemma 77, for each $g \in \mathrm{ext}(\mathcal{B})$ we have $\mathscr{L}_{g,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M_g,M_g}$ for $M_g = \{g_* \mu \mid \mu \in \mathcal{M}^1\}$. In particular, if g is constant this set is the singleton $M_g = \{\delta_{g(\Omega)}\}$, and if g is non-constant then it is exactly the set M of probability measures supported on $\{-1, 1\}$. Thus, $\inf_{g \in \mathrm{ext}(\mathcal{B})} \mathscr{L}_{g,\mathcal{M}^1} = \mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M}$.

Note that the set M with the total variation norm is homeomorphic to the unit interval $[0,1]$ via the linear map $p \mapsto p \cdot \delta_{\{1\}} + (1-p) \cdot \delta_{\{-1\}}$. Then the function $f : \mathbb{R} \times M^2 \to \overline{\mathbb{R}}$ given by $f\big(\varepsilon, (\mu,\nu)\big) = \mathrm{D}_\phi(\mu \parallel \nu) + \delta_{\{0\}}\big(\mu(\mathrm{Id}_{\{-1,1\}}) - \nu(\mathrm{Id}_{\{-1,1\}}) - \varepsilon\big)$ is jointly convex and lower semicontinuous, and hence since M is compact also inf-compact. Thus, by Lemma 5, the function $\mathscr{L}_{\mathrm{Id}_{\{-1,1\}},M,M} = \inf_{(\mu,\nu)\in M^2} f\big(\,\cdot\,,(\mu,\nu)\big)$ is convex and inf-compact as desired. ■

Lemma 86 implies that it suffices to compute $K_{\mathcal{B},\mathcal{M}^1}$.

Lemma 87 *$K_{\mathcal{B},\mathcal{M}^1}(t) = \mathrm{hgt}_{\psi^\star}(2t)$ for every $t \geq 0$.*

Proof For $M = \{p \cdot \delta_{\{1\}} + (1-p) \cdot \delta_{\{-1\}} \mid p \in [0,1]\}$, we have by Lemma 86 and Theorem 83 that $K_{\mathcal{B},\mathcal{M}^1} = \mathscr{L}^\star_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1} = \mathscr{L}^\star_{\mathrm{Id}_{\{-1,1\}},M,M} = \sup_{\nu\in M} K_{\mathrm{Id}_{\{-1,1\}},\nu}$. For $p \in [0,1]$ we have $K_{\mathrm{Id}_{\{-1,1\}},p\cdot\delta_{\{1\}}+(1-p)\cdot\delta_{\{-1\}}} = \inf_{\lambda\in\mathbb{R}}\big(p \cdot \psi^\star(t+\lambda) + (1-p)\cdot\psi^\star(-t+\lambda)\big)$, so that

$$K_{\mathcal{B},\mathcal{M}^1}(t) = \sup_{p\in[0,1]} \inf_{\lambda\in\mathbb{R}} \big(p \cdot \psi^\star(\lambda+t) + (1-p)\cdot\psi^\star(\lambda-t)\big)\,. \tag{26}$$

This mixed optimization problem is convex in λ for each p and linear in p for each $\lambda \in \mathbb{R}$, and the interval $[0,1]$ is compact, so by the Sion minimax theorem (Sion, 1958) we can swap the supremum and infimum to get

$$\begin{aligned} K_{\mathcal{B},\mathcal{M}^1}(t) &= \inf_{\lambda\in\mathbb{R}} \sup_{p\in[0,1]} \big(p \cdot \psi^\star(\lambda+t) + (1-p)\cdot\psi^\star(\lambda-t)\big) \\ &= \inf_{\lambda\in\mathbb{R}} \max\{\psi^\star(\lambda+t), \psi^\star(\lambda-t)\} \end{aligned}$$

so the claim follows from Lemma 85. ■

Example 17 *For the Kullback–Leibler divergence, since $K_{g,\nu}(t) = \log\nu\big(e^{t(g-\nu(g))}\big)$ as in Example 7, Lemma 87 and Example 16 imply that the optimal bound on the cumulant generating function of a random variable g with $\nu(g) = 0$ and $m \leq g \leq M$ ν-a.s. is $\log\nu\big(e^{tg}\big) \leq \mathrm{hgt}_{\psi^\star}\left[(M-m)t\right] = -1 + \frac{M-m}{2}\coth\frac{M-m}{2} + \log\frac{2\sinh((M-m)t/2)}{t}$. This is a refinement of Hoeffding's lemma, which gives the upper bound of $(M-m)^2t^2/8$, which we will also derive as consequence of a general quadratic relaxation on the height function in Example 21.*

Corollary 88 *$\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(\varepsilon) = \mathrm{hgt}^\star_{\psi^\star}(\varepsilon/2)$ for all $\varepsilon \geq 0$. In particular, if $\mathrm{hgt}_{\psi^\star}$ is differentiable then $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(2\,\mathrm{hgt}'_{\psi^\star}(x)) = x\,\mathrm{hgt}'_{\psi^\star}(x) - \mathrm{hgt}_{\psi^\star}(x)$.*

Proof The main claim is immediate from Lemmas 86 and 87, and the supplemental claim follows from the explicit expression for the convex conjugate for differentiable functions. ■

Example 18 *For the Kullback–Leibler divergence, using Example 16, the supplemental claim of Corollary 88 applied to $x = 2t$ gives $\mathscr{L}_{\mathcal{B},\mathcal{M}^1,\mathcal{M}^1}(V(t)) = \log\frac{t}{\sinh t} + t\coth t - \frac{t^2}{\sinh^2 t}$ for $V(t) = 2\coth t - \frac{t}{\sinh^2 t} - 1/t$, which is exactly the formula derived by Fedotov et al. (2003).*

Remark 89 *Corollary 88 shows that lower bounds on the ϕ-divergence in terms of the total variation are equivalent to upper bounds on the height-for-width function* $\operatorname{hgt}_{\psi^\star}$*, equivalently to lower bounds on the sublevel set volume function of $\psi^\star$. The complementary problem of obtaining upper bounds on the sublevel set volume function is of interest in harmonic analysis due to its connection to studying oscillatory integrals (e.g. Stein (1993, Chapter 8, Proposition 2) and Carbery et al. (1999, §1-2)), and it would be interesting to see if techniques from that literature could be applied in this context.*

Remark 90 *Since the total variation* $\mathrm{TV}(\mu, \nu)$ *is symmetric in terms of μ and ν, the optimal lower bound on* $\mathrm{D}_\phi(\mu \parallel \nu)$ *in terms of* $\mathrm{TV}(\mu,\nu)$ *is the same as the optimal lower bound on* $\mathrm{D}_\phi(\nu \parallel \mu) = \mathrm{D}_{\phi^\dagger}(\mu \parallel \nu)$ *for $\phi^\dagger = x\phi(1/x)$. By Corollary 88, this implies that* $\operatorname{hgt}_{\psi^\star} = \operatorname{hgt}_{(\psi^\dagger)^\star}$ *(note that this can also be derived directly from the definition).*

6.3.2 Application to Pinsker-type inequalities

Corollary 88 implies that to obtain Pinsker-type inequalities, it suffices to upper bound the height function $\operatorname{hgt}_{\psi^\star}(t)$ by a quadratic function of t. In this section, we show such bounds under mild assumptions on $\psi^\star$, both rederiving optimal Pinsker-type inequalities for the Kullback–Leibler divergence and α-divergences for $-1 \leq \alpha \leq 2$ due to Gilardoni (2010), and deriving new but not necessarily optimal Pinsker-type inequalities for all $\alpha \in \mathbb{R}$. We proceed by giving two arguments approximating the minimizer $\lambda(t)$ in the optimization problem defining the height (Lemma 85), and an argument that works directly with the optimal $\lambda(t)$.

We begin with the crudest but most widely applicable bound.

Corollary 91 *If ϕ is twice differentiable on its domain and ϕ'' is monotone, then* $\operatorname{hgt}_{\psi^\star}(t) \leq t^2/(2\phi''(1))$ *for all $t \geq 0$. Equivalently, for such ϕ we have that* $\mathrm{D}_\phi(\mu \parallel \nu) \geq \frac{\phi''(1)}{8} \cdot \mathrm{TV}(\mu,\nu)^2$ *for all $\mu, \nu \in \mathcal{M}^1$.*

Proof If $\phi''(1) = 0$, then the claim is trivial, so we assume that $\phi''(1) > 0$. If ϕ'' is non-decreasing, we have by Taylor's theorem that $\phi(x) \geq \frac{\phi''(1)}{2}(x-1)^2$ for $x \geq 1$, equivalently $\psi(x) \geq \frac{\phi''(1)}{2}x^2$ for $x \geq 0$, so that $\psi^\star(x) \leq \frac{1}{2\phi''(1)}x^2$ for $x \geq 0$. Then $\operatorname{hgt}_{\psi^\star}(t) = \inf_{\lambda\in\mathbb{R}} \max\{\psi^\star(\lambda - t/2), \psi^\star(\lambda + t/2)\} \leq \max\{\psi^\star(0), \psi^\star(t)\} \leq t^2/(2\phi''(1))$. On the other hand, if ϕ'' is non-increasing, then analogously we have $\psi^\star(x) \leq \frac{1}{2\phi''(1)}x^2$ for $x \leq 0$, so that Then $\operatorname{hgt}_{\psi^\star}(t) = \inf_{\lambda\in\mathbb{R}} \max\{\psi^\star(\lambda - t/2), \psi^\star(\lambda + t/2)\} \leq \max\{\psi^\star(0), \psi^\star(-t)\} \leq t^2/(2\phi''(1))$. ■

Example 19 *Most of the standard ϕ-divergences satisfy the condition of Corollary 91, in particular the α-divergences given by $\phi_\alpha = \frac{x^\alpha - \alpha(x-1)-1}{\alpha(\alpha-1)}$ have $\phi''_\alpha(x) = x^{\alpha-2}$ which is monotone for all α. As a result, we get for all α the (possibly suboptimal) Pinsker inequality* $\mathrm{D}_{\phi_\alpha}(\mu \parallel \nu) \geq \frac{1}{8} \cdot \mathrm{TV}(\mu,\nu)^2$ *for all $\mu, \nu \in \mathcal{M}^1$. Such a bound appears to be new for $\alpha > 2$, but for* $\alpha \in [-1, 2]$ *Gilardoni (2010) established the better bound* $\mathrm{D}_{\phi_\alpha}(\mu \parallel \nu) \geq \frac{1}{2} \cdot \mathrm{TV}(\mu,\nu)^2$*, extending the standard case of the Kullback–Leibler divergence $\alpha = 1$. We rederive this optimal constant for these divergences below, and also give general conditions under which such bounds hold.*

Corollary 91 used the crude linear relaxation $-t/2 \leq \lambda(t) \leq t/2$. In the following Corollary, we derive a tighter Pinsker-type inequality by using a Taylor expansion of $\lambda(t)$.

Corollary 92 *Suppose that ϕ strictly convex and twice differentiable on its domain, thrice differentiable at 1 and that*

$$\frac{27\phi''(1)}{\left(3 - z\phi'''(1)/\phi''(1)\right)^3} \leq \phi''(1+z)$$

for all $z \geq -1$. Then $\mathrm{hgt}_{\psi^\star}(t) \leq t^2/(8\phi''(1))$ for all $t \geq 0$, equivalently, for such ϕ we have $\mathrm{D}_\phi(\mu \parallel \nu) \geq \frac{\phi''(1)}{2} \cdot \mathrm{TV}(\mu,\nu)^2$ for all $\mu, \nu \in \mathcal{M}^1$.

Remark 93 *The Pinsker constant in Corollary 92 is best-possible, since if ϕ is twice-differentiable at 1, then Taylor's theorem gives the local expansion $\phi(x) = \phi''(1)/2 \cdot (x-1)^2 + o\big((x-1)^2\big)$, and thus the distributions $\mu_\varepsilon = (1/2+\varepsilon/2, 1/2-\varepsilon/2)$ and $\nu = (1/2, 1/2)$ on the set $\{0,1\}$ have $\mathrm{TV}(\mu_\varepsilon, \nu) = \varepsilon$ and $\mathrm{D}_\phi(\mu_\varepsilon \parallel \nu) = \phi''(1)/2 \cdot \varepsilon^2 + o(\varepsilon^2)$.*

Proof Under suitable regularity assumptions on ϕ and $\psi^\star$, one can easily show that the second order expansion of the function $\lambda(t)$ implicitly defined by $\psi^\star(\lambda(t)+t/2) = \psi^\star(\lambda(t)-t/2)$ is $L(t) = -\frac{ct^2}{24}$ for $c = (\psi^\star)'''(0)/(\psi^\star)''(0) = -\phi'''(1)/\phi''(1)^2$. Taking this as given, we show under the stated assumptions of the proposition that for $L(t) = -\frac{ct^2}{24}$ and $c = -\phi'''(1)/\phi''(1)^2$, we have that $\psi^\star(L(t)+st/2) \leq t^2/(8\phi''(1))$ for $s \in \{\pm 1\}$. Since both sides are 0 at 0, it thus suffices to show $\big(L'(t)+s/2\big)(\psi^\star)'(L(t)+st/2) \leq t/(4\phi''(1))$. Now, let $\lesseqgtr$ indicate $\leq$ if $L'(t)+s/2 \geq 0$ and $\geq$ if $L'(t)+s/2 \leq 0$. Since ϕ strictly convex implies $\psi' = ((\psi^\star)')^{-1}$ is strictly increasing, we thus have that this is equivalent to

$$L(t) + st/2 \lesseqgtr \psi'\left(\frac{t/(4\phi''(1))}{L'(t)+s/2}\right) \tag{27}$$

Write $z = \frac{t/(4\phi''(1))}{L'(t)+s/2} = \frac{t/(4\phi''(1))}{-ct/12+s/2}$ so that z has the same sign as $L'(t)+s/2$ and $t = \frac{6sz\phi''(1)}{3+cz\phi''(1)}$. Plugging this in and using the fact that $s^2 = 1$, we wish to show that

$$\frac{3z\phi''(1)(6+cz\phi''(1))}{2(3+cz\phi''(1))^2} - \psi'(z) \lesseqgtr 0 \tag{28}$$

for all z such that $t \geq 0$. The left hand side of Eq. (28) is 0 at 0, so since $z > 0$ implies $\lesseqgtr$ is $\leq$ and $z < 0$ implies $\lesseqgtr$ is $\geq$, it suffices to show that the derivative of the left-hand side of Eq. (28) with respect to z is non-positive for all z. This derivative is

$$\frac{27\phi''(1)}{\left(3+cz\phi''(1)\right)^3} - \psi''(z) = \frac{27\phi''(1)}{\left(3-z\phi'''(1)/\phi''(1)\right)^3} - \phi''(1+z) \tag{29}$$

which since $\operatorname{dom}\psi \subseteq [-1,\infty)$ is non-positive for all z if and only if it is non-positive for all $z \geq -1$. ■

Example 20 (Gilardoni (2010)) *For the α-divergences, we have $\phi''_\alpha(x) = x^{\alpha-2}$, and $\phi'''_\alpha(x) = (\alpha-2)x^{\alpha-3}$ so that Corollary 92 is equivalent to the condition $\frac{27}{(3+(2-\alpha)z)^3} \leq (1+z)^{\alpha-2}$ for $z \geq -1$. Note that this is true for $z = 0$ for all α, and the derivative of $\frac{27(1+z)^{2-\alpha}}{(3+(2-\alpha)z)^3}$ with respect to z is $\frac{27(\alpha-2)(\alpha+1)z(1+z)^{1-\alpha}}{(3+(2-\alpha)z)^4}$. Thus, for $\alpha \in [-1,2]$ the sign of the derivative is the opposite of the sign of z, and the condition holds for all $z \geq -1$, recovering the result of Gilardoni (2010) as desired.*

Example 21 *For the case of the Kullback–Leibler divergence, Example 20 rederives Pinsker's inequality and Hoeffding's lemma.*

Finally, we show that one can also obtain optimal Pinsker-type inequalities while arguing directly about the optimal $\lambda(t)$, for which we need the following lemma.

Lemma 94 *Suppose that $f : \mathbb{R} \to \overline{\mathbb{R}}$ is a convex function continuously differentiable on (a,b) the interior of its domain with a unique global minimum and such that $\lim_{x\to a^+} f(x) = \infty = \lim_{x\to b^-} f(x)$. Then there is a continuously differentiable function $\lambda : (a-b, b-a) \to \mathbb{R}$ such that $\mathrm{hgt}_f(t) = f(\lambda(t)+t/2) = f(\lambda(t)-t/2)$ and*

$$\lambda'(t) = \frac{f'\left(\lambda(t)+t/2\right) + f'\left(\lambda(t)-t/2\right)}{2\left(f'\left(\lambda(t)-t/2\right) - f'\left(\lambda(t)+t/2\right)\right)} \tag{30}$$

$$\mathrm{hgt}'_f(t) = \frac{f'(\lambda(t)+t/2)f'(\lambda(t)-t/2)}{f'(\lambda(t)-t/2) - f'(\lambda(t)+t/2)}\,. \tag{31}$$

Proof For each $t \in (a-b, b-a)$, the function $\lambda \mapsto f(\lambda+t/2) - f(\lambda-t/2)$ is continuously differentiable on its domain $(a+\frac{|t|}{2}, b-\frac{|t|}{2})$, with limits $-\infty$ and ∞. Thus, for all such t there exists λ satisfying the implicit equation $f(\lambda(t)+t/2) = f(\lambda(t)-t/2)$, which by Lemma 85 also defines $\mathrm{hgt}_f(t)$. Furthermore, the fact that f has a unique global minimum implies this function is strictly increasing in λ for each t, and thus the implicit function theorem guarantees the existence of the claimed continuously differentiable $\lambda(t)$.

Given the existence of $\lambda(t)$, we have by its definition that $\frac{d}{dt}f(\lambda(t)+t/2) = \frac{d}{dt}f(\lambda(t)-t/2)$, which implies by the chain rule the claimed value for $\lambda'(t)$, which since $\mathrm{hgt}'_f(t) = \frac{d}{dt}f(\lambda(t)+t/2)$ implies the claimed expressions for the derivative of hgt_f. ■

Using the previous lemma, we obtain the same optimal Pinsker-type inequality as in Corollary 92 under related but incomparable assumptions.

Corollary 95 *If ϕ is strictly convex, has a positive second derivative on its domain, $1/\phi''$ is concave, and $\lim_{x\to\phi'(\infty)^-} \psi^\star(x) = \infty$ (e.g. if $\phi'(\infty) = \infty$), then $\mathrm{hgt}_{\psi^\star}(t) \leq t^2/(8\phi''(1))$ for all $t \geq 0$. Equivalently, for such ϕ we have $\mathrm{D}_\phi(\mu \parallel \nu) \geq \frac{\phi''(1)}{2} \cdot \mathrm{TV}(\mu,\nu)^2$ for all $\mu,\nu \in \mathcal{M}^1$.*

Proof By standard results in convex analysis, the existence and positivity of ψ'' imply that $\psi^\star$ is itself twice differentiable (e.g. Hiriart-Urruty and Lemaréchal (1993, Proposition 6.2.5) or Gorni (1991, Proposition 1.1)). Thus, by Lemma 94, it suffices to show that $\mathrm{hgt}'_{\psi^\star}(t) \leq t/(4\phi''(1))$, or equivalently

$$\frac{(\psi^\star)'(\lambda(t)+t/2)(\psi^\star)'(\lambda(t)-t/2)}{(\psi^\star)'(\lambda(t)-t/2) - (\psi^\star)'(\lambda(t)+t/2)} \leq \frac{t}{4\phi''(1)}. \tag{32}$$

Since $\psi^\star(\lambda(t)+t/2) = \psi^\star(\lambda(t)-t/2)$ and $\psi^\star$ has global minimum at 0, we have $\lambda(t)-t/2 \le 0$ and $\lambda(t)+t/2 \ge 0$, and $(\psi^\star)'(\lambda(t)-t/2) \le 0$ and $(\psi^\star)'(\lambda(t)+t/2) \ge 0$. Thus, we have that the left-hand side of Eq. (32) is half the harmonic mean of $(\psi^\star)'(\lambda(t)+t/2)$ and $-(\psi^\star)'(\lambda(t)-t/2)$, so it suffices by the arithmetic mean–harmonic mean inequality to prove

$$(\psi^\star)'(\lambda(t)+t/2) - (\psi^\star)'(\lambda(t)-t/2) \le \frac{t}{\phi''(1)}. \tag{33}$$

Since Eq. (33) holds when $t=0$, it suffices to prove that

$$(1/2+\lambda'(t)) \cdot (\psi^\star)''(\lambda(t)+t/2) + (1/2-\lambda'(t)) \cdot (\psi^\star)''(\lambda(t)-t/2) \le \frac{1}{\phi''(1)}. \tag{34}$$

By the relationship between the second derivative of a function and the one of its conjugate (e.g. Hiriart-Urruty and Lemaréchal (1993, Proposition 6.2.5)), this is equivalent to

$$\frac{1/2+\lambda'(t)}{\psi''\big((\psi^\star)'(\lambda(t)+t/2)\big)} + \frac{1/2-\lambda'(t)}{\psi''\big((\psi^\star)'(\lambda(t)-t/2)\big)} \le \frac{1}{\phi''(1)}. \tag{35}$$

Now, by Eq. (30), we have that $\lambda'(t) \in [-1/2, 1/2]$, so that by Jensen's inequality and the concavity of $1/\psi''$, the left-hand side of Eq. (35) is at most

$$1/\psi''\Big((1/2+\lambda'(t))(\psi^\star)'(\lambda(t)+t/2) - (\lambda'(t)-1/2)(\psi^\star)'(\lambda(t)-t/2)\Big). \tag{36}$$

Finally, since by definition $\psi^\star(\lambda(t)+t/2) = \psi^\star(\lambda(t)-t/2)$, the term inside $1/\psi''$ in Eq. (36) is 0, so since $\psi(x) = \phi(1+x)$ we are done. ■

Example 22 *For the α-divergences, we have $1/\phi''_\alpha(x) = x^{2-\alpha}$ which is concave for $\alpha \in [1,2]$, so Corollary 95 applies for these divergences. Furthermore, by Remark 90, we can consider the reverse α-divergences with $\phi^\dagger_\alpha(x) = x\phi_\alpha(1/x)$ which has $1/(\phi^\dagger_\alpha)''(x) = x^{1+\alpha}$, which is concave for $\alpha \in [-1,0]$.*

7. Discussion

Throughout this paper, the ϕ-cumulant generating function has proved central in explicitating the relationship between ϕ-divergences and integral probability metrics. As a starting point, the identity $K_{g,\nu} = \mathscr{L}^\star_{g,\nu}$ (Theorem 40) expresses the cumulant generating function as the convex conjugate of the best lower bound of $\mathrm{D}_\phi(\mu \parallel \nu)$ in terms of $\mu(g) - \nu(g)$. This establishes a "correspondence principle" by which properties of the relationship between ϕ-divergences and integral probability metrics translate by duality into properties of the cumulant generating function, and vice versa. An advantage of this correspondence is that the function $K_{g,\nu}$, being expressed as the solution of a single-dimensional convex optimization problem (Definition 37), is arguably easier to evaluate and analyze than its counterpart $\mathscr{L}_{g,\nu}$, expressed as the solution to an infinite-dimensional optimization problem. Following Theorem 40, several results from the present paper can be seen as instantiations of this "correspondence principle" and we summarize some of them in Table 2.

Ref.	Property of the ϕ-cumulant generating function	Property of the ϕ-divergence
§5.1	$K_{g,\nu}(t) \leq B(t)$ for all $t \in \mathbb{R}$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq B^\star\big(\mu(g) - \nu(g)\big)$ for all $\mu \in X_g^1$
§5.2	$0 \in \operatorname{int}(\operatorname{dom} K_{g,\nu})$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq L\big(\lvert\mu(g) - \nu(g)\rvert\big)$ for some $L \not\equiv 0$, all $\mu \in X_g^1$
§5.4	$K_{g,\nu}$ differentiable at 0	$\mathrm{D}_\phi(\nu_n \parallel \nu) \to 0$ implies $\nu_n(g) \to \nu(g)$ for all $(\nu_n) \in \big(X_g^1\big)^{\mathbb{N}}$
§6.2	$K_{g,\nu}(t) \leq E(t)$ for all $t \in \mathbb{R}$, $g \in \mathcal{G}$, $\nu \in X_{\mathcal{G}}^1$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq E^\star\big(d_{\mathcal{G}}(\mu, \nu)\big)$ for all $\mu, \nu \in X_{\mathcal{G}}^1$
§6.3	$\mathrm{hgt}_{\psi^\star}(2t) \leq B(t)$ for all $t \in \mathbb{R}$	$\mathrm{D}_\phi(\mu \parallel \nu) \geq B^\star\big(\mathrm{TV}(\mu, \nu)\big)$ for all $\mu, \nu \in \mathcal{M}^1$

Table 2: Several examples, proved in this paper, of the dual correspondence between properties of the ϕ-cumulant generating function and properties of the relationship between the ϕ-divergence and mean deviations. Throughout, $\mu \in \mathcal{M}^1$, $g \in L^1(\nu)$, $B : \mathbb{R} \to \mathbb{R}$ is arbitrary, $E : \mathbb{R} \to \mathbb{R}$ is even, $\mathcal{G} \subseteq \mathcal{L}^0$, and X_g^1 and $X_{\mathcal{G}}^1$ are as in Definition 41.

A limitation of this correspondence is that it only describes the optimal lower bound function $\mathscr{L}_{g,\nu}$ via its convex conjugate. When $\mathscr{L}_{g,\nu}$ is lower semicontinuous, this is without any loss of information by the Fenchel–Moreau theorem, but in general this only provides information about the *biconjugate* $\mathscr{L}_{g,\nu}^{\star\star}$. While $\mathscr{L}_{g,\nu}$ and $\mathscr{L}_{g,\nu}^{\star\star}$ differ in at most two points, as discussed in Section 5.1, the difference between the optimal lower bound and its biconjugate is potentially much more important when considering a class of functions $\mathcal{G}$ or a class of measures N as in Section 6.1. Some conditions under which this lower bound $\mathscr{L}_{\mathcal{G},N}$ is necessarily convex and lower semicontinuous were derived in Sections 5.3 and 6.3, and we gave a characterization of $\mathscr{L}_{\mathcal{G},N}$ up to countably many points in Remark 81 regardless, but this does not completely answer the question (cf. Remarks 63 and 71). We believe that an interesting direction for future work would be to identify natural necessary or sufficient conditions under which $\mathscr{L}_{\mathcal{G},N}$ is convex or lower semicontinuous.

Acknowledgments

The authors are grateful to Flavio du Pin Calmon, Benjamin L. Edelman, Julien Fageot, Salil Vadhan, and the anonymous reviewers for their helpful comments and suggestions on an earlier draft of this paper, and to Julien Fageot and Salil Vadhan for suggesting the problem considered in Section 5.4.

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

Appendix A. Deferred proofs

In this section, for the sake of completeness, we include proofs of results that follow from standard tools in convex analysis.

A.1 Proof of Lemma 16

Lemma 16 follows immediately from König (1986, Remark 1.9) stated in the general context of superconvex structures, which applies to cs-compact sets by König (1986, Example 1.6(0)). For completeness, we include a proof here in the language of topological vector spaces.

First, a convex function which is upper bounded on a cs-closed set in the sense of Jameson (1972) satisfies an infinite-sum version of convexity called cs-convexity (convex-series convexity) by Simons (1990).

Lemma 96 *Let C be a cs-closed subset of a real Hausdorff topological vector space and let $f : C \to \mathbb{R}$ be a convex function such that $\sup_{x\in C} f(x) < \infty$. Then f is cs-convex.*

Proof Let $(\lambda_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ be a sequence of real numbers such that $\sum_{i=0}^{\infty} \lambda_i = 1$ and $\lambda_n \geq 0$ for all $n \in \mathbb{N}$. Let $(x_n)_{n\in\mathbb{N}} \in C^{\mathbb{N}}$ be a sequence of elements in C such that $r_0 \coloneqq \sum_{i=0}^{\infty} \lambda_i x_i$ exists, and thus is in C since C is cs-closed. We wish to show that

$$f(r_0) \leq \liminf_{n\to\infty} \sum_{i=0}^{n} \lambda_i f(x_i)\,. \tag{37}$$

Define for each $n \in \mathbb{N}$ the partial sums

$$\Lambda_n \coloneqq \sum_{i=0}^{n} \lambda_i \quad \text{and} \quad s_n \coloneqq \Lambda_n^{-1} \sum_{i=0}^{n} \lambda_i x_i\,.$$

If $\Lambda_n = 1$ for some $n \in \mathbb{N}$ then Eq. (37) is immediate from convexity. Otherwise, we have that for each $n \geq 0$,

$$r_n \coloneqq \frac{r_0 - \Lambda_n \cdot s_n}{1 - \Lambda_n} = \sum_{i=n+1}^{\infty} \frac{\lambda_i}{1 - \Lambda_n} \cdot x_i$$

is an element of C since C is cs-closed, so that by convexity of f

$$\begin{aligned} f(r_0) &\leq \Lambda_n \cdot f(s_n) + (1 - \Lambda_n) \cdot f(r_n) \\ &\leq \sum_{i=0}^{n} \lambda_i f(x_i) + (1 - \Lambda_n) \cdot \sup_{x\in C} f(x)\,. \end{aligned}$$

Since $\sup_{x\in C} f(x) < \infty$ and $\lim_{n\to\infty} \Lambda_n = 1$ by assumption, the previous inequality implies Eq. (37) as desired. ■

Second, cs-convex functions are necessarily bounded below on cs-compact sets. Together with Lemma 96, this implies Lemma 16.

Lemma 97 *Let $f : C \to \mathbb{R}$ be a cs-convex function on a cs-compact subset C of a real Hausdorff topological vector space. Then* $\inf_{x\in C} f(x) > -\infty$.

Proof We prove the contrapositive, that if $\inf_{x\in C} f(x) = -\infty$ then f is not cs-convex. Indeed, if $\inf_{x\in C} f(x) = -\infty$, then for each $n \in \mathbb{N}$ there exists $x_n \in C$ with $f(x_n) \leq -4^n$. Since C is cs-compact, the element $\overline{x} \coloneqq \sum_{i=1}^{\infty} 2^{-i} \cdot x_i$ exists and is in C. But then

$$\liminf_{n\to\infty} \sum_{i=1}^{n} 2^{-i} \cdot f(x_i) \leq \liminf_{n\to\infty} \sum_{i=1}^{n} 2^{-i} \cdot -4^i = -\infty < f(\overline{x}),$$

proving that f is not cs-convex. ■

A.2 ϕ-cumulant generating function

Lemma 98 (Lemma 36 restated) *The function $\psi^\star : x \mapsto \phi^\star(x) - x$ is non-negative, convex, and inf-compact. Furthermore, it satisfies $\psi^\star(0) = 0$, $\psi^\star(x) \leq -x$ when $x \leq 0$, and* $\operatorname{int}(\operatorname{dom}\psi^\star) = \big(-\infty, \phi'(\infty)\big)$.

Proof We have that $\psi^\star(x) = \sup_{y\in\mathbb{R}}(y \cdot x - \phi(y+1)) = \sup_{y\in\mathbb{R}}((y-1)\cdot x - \phi(y)) = -x + \sup_{y\in\mathbb{R}}(y\cdot x - \phi(y)) = \phi^\star(x) - x$. Non-negativity of $\psi^\star$ holds since $\psi^\star(x) \geq 0 \cdot x - \psi(0) = 0$, and convexity and lower semicontinuity hold for any convex conjugate. For inf-compactness, we have since $0 \in \operatorname{int}\operatorname{dom}\phi$ by assumption that there exists $\alpha > 0$ with $[-\alpha, \alpha] \subseteq \operatorname{dom}\psi$, so that $\psi^\star(y) \geq \max\{\alpha \cdot y - \psi(\alpha), -\alpha\cdot y - \psi(-\alpha)\} \geq \alpha \cdot |y| - \max\{\psi(\alpha), \psi(-\alpha)\}$, so that the sublevel sets of $\psi^\star$ are closed and bounded and thus compact.

The claim about $\operatorname{dom}\psi$ is immediate from Lemma 13 since $\operatorname{dom}\phi \subseteq \mathbb{R}_{\geq 0}$ implies $\phi'(-\infty) = -\infty$. Finally, $\operatorname{dom}\phi \subseteq \mathbb{R}_{\geq 0}$ also implies for $x \leq 0$ that $\psi^\star(x) = \sup_{y\geq -1}\big(y\cdot x - \psi(y)\big) \leq \sup_{y\geq -1} y \cdot x - \inf_{y\geq -1}\psi(y) = -x$ where the last equality is because $\psi \geq 0$ and $x \leq 0$. ■

Proposition 99 (Proposition 39 restated) *For every σ-ideal Ξ, probability measure $\nu \in \mathcal{M}_c^1(\Xi)$, and $g \in L^0(\Xi)$, $K_{g,\nu,\Xi} : \mathbb{R} \to \overline{\mathbb{R}}$ is non-negative, convex, lsc, and satisfies $K_{g,\nu,\Xi}(0) = 0$.*

Furthermore, if g is not ν-essentially constant then $K_{g,\nu,\Xi}$ is inf-compact. If there exists $c \in \mathbb{R}$ such that $g = c$ ν-almost surely, then there exists $t > 0$ (resp. $t < 0$) such that $K_{g,\nu,\Xi}(t) > 0$ if and only if $\phi'(\infty) < \infty$ and $\operatorname{ess\,sup}_\Xi g > c$ *(resp.* $\operatorname{ess\,inf}_\Xi g < c$*).*

We prove this in steps, using the following important function:

Definition 100 *For every σ-ideal Ξ, probability measure $\nu \in \mathcal{M}_c^1(\Xi)$, and $g \in L^0(\Xi)$, define*

$$\begin{aligned} F_{g,\nu,\Xi}(t,\lambda) &\coloneqq \begin{cases} \int \psi^\star(tg+\lambda)\,\mathrm{d}\nu & \text{if } \operatorname{ess\,sup}_\Xi(tg+\lambda) \leq \phi'(\infty) \\ +\infty & \text{otherwise} \end{cases} \\ &= \int \psi^\star(tg+\lambda)\,\mathrm{d}\nu + \begin{cases} 0 & \text{if } tg+\lambda \in [-\infty, \phi'(\infty)]\ \Xi\text{-a.e.} \\ +\infty & \text{otherwise} \end{cases}, \end{aligned}$$

so that $K_{g,\nu,\Xi} = \inf_{\lambda\in\mathbb{R}} F_{g,\nu,\Xi}(\,\cdot\,,\lambda)$.

Lemma 101 *For every σ-ideal Ξ, probability measure $\nu \in \mathcal{M}_c^1(\Xi)$, and $g \in L^0(\Xi)$, the function $F_{g,\nu,\Xi}$ is non-negative, convex, lsc, and the set $\{\lambda \in \mathbb{R} \mid F_{g,\nu,\Xi}(0,\lambda) = 0\}$ is compact and contains 0.*

Proof Non-negativity of $F_{g,\nu,\Xi}$ is immediate from non-negativity of $\psi^\star$. The function

$$(t,\lambda) \mapsto \begin{cases} 0 & \text{if } tg + \lambda \in [-\infty, \phi'(\infty)] \text{ } \Xi\text{-a.e.} \\ +\infty & \text{otherwise} \end{cases}$$

is convex and lsc since $[-\infty, \phi'(\infty)]$ is a closed interval.

Similarly, the convexity of $\psi^\star$ implies the convexity of $(t,\lambda) \mapsto \int \psi^\star(tg+\lambda)\,\mathrm{d}\nu$. Furthermore, by Fatou's lemma and since $\psi^\star$ is lsc we have for every sequence $(t_n, \lambda_n) \to (t,\lambda)$ that

$$\liminf_{n\to\infty} \int \psi^\star(t_n g + \lambda_n)\,\mathrm{d}\nu \geq \int \liminf_{n\to\infty} \psi^\star(t_n g + \lambda_n)\,\mathrm{d}\nu \geq \int \psi^\star(tg+\lambda)\,\mathrm{d}\nu\,,$$

so that this function is also lower semicontinuous.

Finally, $\{\lambda \in \mathbb{R} \mid F_{g,\nu,\Xi}(0,\lambda) = 0\}$ is a sublevel set of a non-negative lsc function and so is closed, it contains 0 since $\psi^\star(0) = 0$ and $\phi'(\infty) \geq 0$, and is bounded since it is contained in the compact set $\{\lambda \in \mathbb{R} \mid \psi^\star(\lambda) = 0\}$. ■

Lemma 102 *For every σ-ideal Ξ, probability measure $\nu \in \mathcal{M}_c^1(\Xi)$, and $g \in L^0(\Xi)$, we have $\mathbb{R}_{\geq 0} \subseteq \{t \in \mathbb{R} \mid \exists \lambda \in \mathbb{R} \wedge F_{g,\nu,\Xi}(t,\lambda) = 0\}$ if and only if g is ν-essentially constant and either $\phi'(\infty) = \infty$ or $\operatorname{ess\,sup}_\Xi g = \operatorname{ess\,sup}_\nu g$.*

Proof If $g = c$ holds ν-a.s. for some $c \in \mathbb{R}$ and either $\phi'(\infty) = \infty$ or $\operatorname{ess\,sup}_\Xi g = \operatorname{ess\,sup}_\nu g = c$, then for all $t \geq 0$ we have $F_{g,\nu,\Xi}(t, -t\cdot c) = 0$ since $tg + \lambda$ is 0 ν-a.s. and at most $\phi'(\infty) \geq 0$ Ξ-a.e.

Conversely, suppose $\mathbb{R}_{\geq 0} \subseteq \{t \in \mathbb{R} \mid \exists \lambda \in \mathbb{R} \wedge F_{g,\nu,\Xi}(t,\lambda) = 0\}$. Then for every $t \geq 0$ there is $\lambda \in \mathbb{R}$ such that $tg + \lambda \in \{x \in \mathbb{R} \mid \psi^\star(x) = 0\} \subseteq [-\infty, \psi^\star(\infty)]$ holds ν-a.s. and $tg + \lambda \in [-\infty, \phi'(\infty)]$ holds Ξ-a.e. Since $\psi^\star$ is non-negative, convex, and inf-compact, the set $\{x \in \mathbb{R} \mid \psi^\star(x) = 0\}$ is a compact interval $[a,b]$, and thus there is $\lambda \in \mathbb{R}$ such that $tg + \lambda \in [a,b]$ holds ν-a.s. if only if $|t| \cdot \big(\operatorname{ess\,sup}_\nu g - \operatorname{ess\,inf}_\nu g\big) \leq b - a < \infty$. Thus, since this holds for all $t \in \mathbb{R}$, we have $\operatorname{ess\,sup}_\nu g = \operatorname{ess\,inf}_\nu g$, equivalently that $g = c$ holds ν-a.s. for some $c \in \mathbb{R}$. Thus, the condition on t reduces to the existence of $\lambda \in \mathbb{R}$ such that $tc + \lambda \in [a,b]$ and $\operatorname{ess\,sup}_\Xi tg + \lambda = tc + \lambda + t \cdot (\operatorname{ess\,sup}_\Xi g - c) \leq \phi'(\infty)$. In particular, this implies that $a + t \cdot (\operatorname{ess\,sup}_\Xi g - c) \leq \phi'(\infty)$ for all $t \geq 0$, which implies either $\phi'(\infty) = \infty$ or $\operatorname{ess\,sup}_\Xi g \leq c = \operatorname{ess\,sup}_\nu g$ as desired. ■

We can finally prove Proposition 39.

Proof [Proof of Proposition 39] The main claim is immediate by applying standard results in convex analysis (e.g. (Rockafellar and Wets, 1998a, Propositions 1.17 and 3.32)) to Lemma 101. Furthermore, these results imply that $\{t \in \mathbb{R} \mid K_{g,\nu,\Xi}(t) = 0\} = \{t \in \mathbb{R} \mid \exists \lambda \in \mathbb{R} \wedge F_{g,\nu,\Xi}(t,\lambda) = 0\}$.

For the supplemental claim, we have since $K_{g,\nu,\Xi}$ is non-negative, convex, lsc, and 0 at 0 that it is inf-compact if and only if there exist $t_+ > 0$ and $t_- < 0$ such that $K_{g,\nu,\Xi}(t_+), K_{g,\nu,\Xi}(t_-) > 0$. The claimed characterization thus follows from applying Lemma 102 to g and $-g$. ■

Bibliography

R. Agrawal and T. Horel. Optimal Bounds between f-Divergences and Integral Probability Metrics. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, July 2020.

S. M. Ali and S. D. Silvey. 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.

Y. Altun and A. Smola. Unifying divergence minimization and statistical inference via convex duality. In G. Lugosi and H. U. Simon, editors, *Learning Theory*, pages 139–153, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-35296-9. doi: 10.1007/11776420_13.

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

A. Ben-Tal and A. Charnes. A dual optimization framework for some problems of information theory and statistics. Technical report, Center for Cybernetic Studies, University of Texas, Austin, 1977.

A. Ben-Tal and A. Charnes. A dual optimization framework for some problems of information theory and statistics. *Problems of Control and Information Theory. Problemy Upravlenija i Teorii Informacii*, 8(5-6):387–401, 1979. ISSN 0370-2529.

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

S. G. Bobkov and F. Götze. Exponential integrability and transportation cost related to logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, Apr. 1999. ISSN 0022-1236. doi: 10.1006/jfan.1998.3326.

F. Bolley and C. Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, 14(3):331–352, 2005.

J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338, 1991. doi: 10.1137/0329017.

J. M. Borwein and A. S. Lewis. Partially-finite programming in L_1 and the existence of maximum entropy estimates. *SIAM Journal on Optimization*, 3(2):248–267, 1993. doi: 10.1137/0803012.

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

N. Bourbaki. *Topological Vector Spaces*. Elements of Mathematics. Springer-Verlag, Berlin, 1987. ISBN 978-3-540-13627-9. doi: 10.1007/978-3-642-61715-7. Translated by H.G. Eggleston & S. Madan from *Espaces vectoriels topologiques*, Masson, Paris, 1981.

J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, Jan 1979. ISSN 1432-2064. doi: 10.1007/BF00535278.

A. Brøndsted. Conjugate convex functions in topological vector spaces. *Matematisk-fysiske Meddelelser udgivet af det Kongelige Danske Videnskabernes Selskab*, 34(2):27, 1964. ISSN 0023-3323.

M. Broniatowski and A. Keziou. Minimization of ϕ-divergences on sets of signed measures. *Studia Scientiarum Mathematicarum Hungarica*, 43(4):403–442, 2006. doi: 10.1556/SScMath.43.2006.4.2.

A. Carbery, M. Christ, and J. Wright. Multidimensional van der Corput and sublevel set estimates. *Journal of the American Mathematical Society*, 12(4):981–1015, 1999. ISSN 1088-6834. doi: 10.1090/S0894-0347-99-00309-4.

I. Csiszár. Informationstheoretische Konvergenzbegriffe im Raum der Wahrscheinlichkeitsverteilungen. *A Magyar Tudományos Akadémia. Matematikai Kutató Intézetének Közleményei*, 7:137–158, 1962. ISSN 0541-9514.

I. Csiszár. Über topologische und metrische Eigenschaften der relativen Information der Ordnung α. In *Transactions of the Third Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, 1962*, pages 63–73. Publishing House of the Czechoslovak Academy of Science, Prague, 1964.

I. Csiszár. On topological properties of f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, 2:329–339, 1967.

I. Csiszár and F. Matúš. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, June 2003. ISSN 0018-9448. doi: 10.1109/TIT.2003.810633.

I. Csiszár and F. Matúš. Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. *Kybernetika*, 48(4):637–689, 2012. ISSN 0023-5954.

I. Csiszár. 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.

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

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3(1):146–158, 02 1975. doi: 10.1214/aop/1176996454.

I. Csiszár, F. Gamboa, and E. Gassiat. MEM pixel correlated solutions for generalized moment and interpolation problems. *IEEE Transactions on Information Theory*, 45(7): 2253–2270, Nov. 1999. doi: 10.1109/18.796367.

M. D. Donsker and S. R. S. Varadhan. 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.

R. M. Dudley. On sequential convergence. *Transactions of the American Mathematical Society*, 112(3):483–507, 1964. ISSN 0002-9947, 1088-6850. doi: 10.1090/ S0002-9947-1964-0175081-6.

R. M. Dudley. Consistency of M-Estimators and One-Sided Bracketing. In E. Eberlein, M. Hahn, and M. Talagrand, editors, *High Dimensional Probability*, pages 33–58. Birkhäuser Basel, Basel, 1998. ISBN 978-3-0348-9790-7 978-3-0348-8829-5. doi: 10.1007/978-3-0348-8829-5_3.

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

G. A. Edgar and L. Sucheston. On maximal inequalities in Orlicz spaces. In R. D. Mauldin, R. M. Shortt, and C. E. Silva, editors, *Contemporary Mathematics*, volume 94, pages 113–129. American Mathematical Society, Providence, Rhode Island, 1989. ISBN 978-0-8218-5099-2 978-0-8218-7682-4. doi: 10.1090/conm/094/1012982.

I. Ekeland and R. Témam. *Convex Analysis and Variational Problems.* Society for Industrial and Applied Mathematics, 1999. doi: 10.1137/1.9781611971088.

A. A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, June 2003. doi: 10.1109/TIT.2003. 811927.

G. L. Gilardoni. 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.

G. L. Gilardoni. An improvement on Vajda's inequality. In V. Sidoravicius and M. E. Vares, editors, *In and Out of Equilibrium 2*, volume 60 of *Progress in Probability*, pages 299–304. Birkhäuser Basel, Basel, 2008. ISBN 978-3-7643-8786-0. doi: 10.1007/978-3-7643-8786-0_ 14.

G. L. Gilardoni. 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.

G. Gorni. Conjugation and second-order properties of convex functions. *Journal of Mathematical Analysis and Applications*, 158(2):293–315, July 1991. ISSN 0022-247X. doi: 10.1016/0022-247X(91)90237-T.

N. Gozlan and C. Léonard. Transport inequalities. A survey. *Markov Processes and Related Fields*, 16(4):635–736, 2010. ISSN 1024-2953. URL `http://math-mprf.org/journal/articles/id1224/`.

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

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

P. Harremoës. Information Topologies with Applications. In I. Csiszár, G. O. H. Katona, G. Tardos, and G. Wiener, editors, *Entropy, Search, Complexity*, volume 16, pages 113–150. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-32573-4. doi: 10.1007/978-3-540-32777-6_5. Series Title: Bolyai Society Mathematical Studies.

P. Harremoës and I. Vajda. 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.

H. H. Herda. On non-symmetric modular spaces. *Colloquium Mathematicum*, 17(2):333–346, 1967. ISSN 0010-1354. doi: 10.4064/cm-17-2-333-346.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I*, volume 305 of *Grundlehren Der Mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993. ISBN 978-3-642-08161-3 978-3-662-02796-7. doi: 10.1007/978-3-662-02796-7.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459. doi: 10.2307/2282952.

A. D. Ioffe and V. M. Tikhomirov. On minimization of integral functionals. *Functional Analysis and Its Applications*, 3(3):218–227, Jul 1969. ISSN 1573-8485. doi: 10.1007/BF01676623.

G. J. O. Jameson. Convex series. *Mathematical Proceedings of the Cambridge Philosophical Society*, 72(1):37–47, July 1972. ISSN 1469-8064, 0305-0041. doi: 10.1017/S0305004100050933.

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

J. H. B. Kemperman. On the optimum rate of transmitting information. *The Annals of Mathematical Statistics*, 40(6):2156–2177, 12 1969. doi: 10.1214/aoms/1177697293.

M. Khosravifard, D. Fooladivanda, and T. A. Gulliver. Exceptionality of the Variational Distance. In *Proceedings of the 2006 IEEE Information Theory Workshop*, pages 274–276, Chengdu, China, Oct. 2006. IEEE. ISBN 978-1-4244-0067-6 978-1-4244-0068-3. doi: 10.1109/ITW2.2006.323802.

M. Khosravifard, D. Fooladivanda, and T. A. Gulliver. Confliction of the Convexity and Metric Properties in f-Divergences. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E90-A(9):1848–1853, Sept. 2007. ISSN 0916-8508. doi: 10.1093/ietfec/e90-a.9.1848.

J. Kisyński. Convergence du type $\mathcal{L}$. *Colloquium Mathematicum*, 7(2):205–211, 1960. ISSN 0010-1354. doi: 10.4064/cm-7-2-205-211.

H. König. Theory and applications of superconvex spaces. In *Aspects of Positivity in Functional Analysis (Tübingen, 1985)*, volume 122 of *North-Holland Math. Stud.*, pages 79–118. North-Holland, Amsterdam, 1986.

S. Kullback. *Information theory and statistics*. Wiley, New York, 1959.

S. Kullback. A lower bound for discrimination information in terms of variation. *IEEE Transactions on Information Theory*, 13(1):126–127, January 1967. doi: 10.1109/TIT.1967.1053968.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, Mar. 1951. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177729694.

C. Léonard. Minimizers of energy functionals. *Acta Mathematica Hungarica*, 93(4):281–325, 2001a. ISSN 02365294. doi: 10.1023/A:1017919422086.

C. Léonard. Minimization of Energy Functionals Applied to Some Inverse Problems. *Applied Mathematics and Optimization*, 44(3):273–297, Jan. 2001b. ISSN 0095-4616, 1432-0606. doi: 10.1007/s00245-001-0019-5.

C. Léonard. Orlicz Spaces. https://leonard.perso.math.cnrs.fr/papers/Leonard-Orlicz%20spaces.pdf, Apr. 2007. URL `https://leonard.perso.math.cnrs.fr/papers/Leonard-Orlicz%20spaces.pdf`.

V. L. Levin. Some properties of support functionals. *Mathematical Notes of the Academy of Sciences of the USSR*, 4(6):900–906, Dec. 1968. ISSN 0001-4346, 1573-8876. doi: 10.1007/BF01110826.

W. Luxemburg and A. Zaanen. Conjugate spaces of Orlicz spaces. *Indagationes Mathematicae (Proceedings)*, 59:217–228, 1956. ISSN 1385-7258. doi: 10.1016/S1385-7258(56)50029-7.

K. Marton. A simple proof of the blowing-up lemma (corresp.). *IEEE Transactions on Information Theory*, 32(3):445–446, May 1986. ISSN 1557-9654. doi: 10.1109/TIT.1986.1057176.

J. J. Moreau. Sur la fonction polaire d'une fonction semi-continue supérieurement. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 258:1128–1130, 1964.

T. Morimoto. Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, Mar. 1963. ISSN 0031-9015, 1347-4073. doi: 10.1143/JPSJ.18.328.

M. Morse and W. Transue. Functionals f Bilinear Over the Product $A \times B$ of Two Pseudo-Normed Vector Spaces: II. Admissible Spaces A. *Annals of Mathematics*, 51(3):576–614, 1950. ISSN 0003-486X. doi: 10.2307/1969370.

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

J. Musielak. *Orlicz Spaces and Modular Spaces*, volume 1034 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983. ISBN 978-3-540-38692-6. doi: 10.1007/BFb0072210.

H. Nakano. *Modulared Semi-Ordered Linear Spaces*. Tokyo Mathematical Book Series,v. 1. Maruzen Co., Tokyo, 1950.

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

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

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

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

M. S. Pinsker. Informatsiya i informatsionnaya ustoichivost' sluchainykh velichin i protsessov. *Probl. Peredachi Inf.*, 7, 1960.

M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, 1964. Translation of Pinsker (1960) by Amiel Feinstein.

M. M. Rao and Z. D. Ren. *Theory of Orlicz Spaces*. Number 146 in Monographs and Textbooks in Pure and Applied Mathematics. M. Dekker, New York, 1991. ISBN 978-0-8247-8478-2.

M. D. Reid and R. C. Williamson. Generalised Pinsker inequalities. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12:731–817, 2011. ISSN 1532-4435.

A. Rényi. On measures of entropy and information. In J. Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 547–561. University of California Press, 1961.

R. T. Rockafellar. Level sets and continuity of conjugate convex functions. *Transactions of the American Mathematical Society*, 123(1):46–63, 1966. ISSN 0002-9947, 1088-6850. doi: 10.1090/S0002-9947-1966-0192318-X.

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

R. T. Rockafellar. Integrals which are convex functionals. II. *Pacific J. Math.*, 39(2):439–469, 1971.

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

R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren Der Mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998a. ISBN 978-3-642-02431-3. doi: 10.1007/978-3-642-02431-3.

R. T. Rockafellar and R. J.-B. Wets. Measurability. In Rockafellar and Wets (1998a), chapter 14, pages 642–683. ISBN 978-3-642-02431-3. doi: 10.1007/978-3-642-02431-3_14.

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

D. Russo and J. Zou. 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.

I. N. Sanov. On the probability of large deviations of random variables. *Mat. Sb. (N.S.)*, 42 (84):11–44, 1957.

I. Sason. On f-Divergences: Integral Representations, Local Behavior, and Inequalities. *Entropy*, 20(5):383, May 2018. doi: 10.3390/e20050383.

I. Sason and S. Verdú. f-Divergence Inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, Nov. 2016. doi: 10.1109/TIT.2016.2603151.

S. Simons. The occasional distributivity of $\circ$ over $\overset{+}{e}$ and the change of variable formula for conjugate functions. *Nonlinear Analysis: Theory, Methods & Applications*, 14(12): 1111–1120, Jan. 1990. ISSN 0362546X. doi: 10.1016/0362-546X(90)90071-N.

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

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

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

E. M. Stein. *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals.* Number 43 in Princeton Mathematical Series. Princeton University Press, Princeton, NJ, 1993. ISBN 978-0-691-03216-0.

E. Szpilrajn. Remarques sur les fonctions complètement additives d'ensemble et sur les ensembles jouissant de la propriété de Baire. *Fundamenta Mathematicae*, 22(1):303–311, 1934. ISSN 0016-2736. doi: 10.4064/fm-22-1-303-311.

M. Teboulle and I. Vajda. Convergence of best ϕ-entropy estimates. *IEEE Transactions on Information Theory*, 39(1):297–301, 1993.

I. Vajda. Note on discrimination information and variation. *IEEE Transactions on Information Theory*, 16(6):771–773, November 1970. doi: 10.1109/TIT.1970.1054557.

I. Vajda. 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.

I. Vajda. χ^α-divergence and generalized Fisher's information. In *Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 873–886. Academia, Prague, 1973.

M. Valadier. Intégration de convexes fermés notamment d'épigraphes. Inf-convolution continue. *Rev. Française Informat. Recherche Opérationnelle*, 4(Sér. R-2):57–73, 1970.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4.

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