Journal of Machine Learning Research 23 (2022) 1-70 Submitted 1/21; Revised 9/21; Published 2/22

(f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics

Jeremiah Birrell BIRRELL@MATH.UMASS.EDU
TRIPODS Institute for Theoretical Foundations of Data Science
University of Massachusetts Amherst
Amherst, MA 01003, USA

Paul Dupuis DUPUIS@DAM.BROWN.EDU
Division of Applied Mathematics
Brown University
Providence, RI 02912, USA

Markos A. Katsoulakis MARKOS@MATH.UMASS.EDU
Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, MA 01003, USA

Yannis Pantazis PANTAZIS@IACM.FORTH.GR
Institute of Applied and Computational Mathematics
Foundation for Research and Technology - Hellas
Heraklion, GR-70013, Greece

Luc Rey-Bellet LUC@MATH.UMASS.EDU
Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, MA 01003, USA

Editor: Marco Cuturi

Abstract

We develop a rigorous and general framework for constructing information-theoretic divergences that subsume both f-divergences and integral probability metrics (IPMs), such as the 1-Wasserstein distance. We prove under which assumptions these divergences, hereafter referred to as (f, Γ)-divergences, provide a notion of 'distance' between probability measures and show that they can be expressed as a two-stage mass-redistribution/mass-transport process. The (f, Γ)-divergences inherit features from IPMs, such as the ability to compare distributions which are not absolutely continuous, as well as from f-divergences, namely the strict concavity of their variational representations and the ability to control heavy-tailed distributions for particular choices of f. When combined, these features establish a divergence with improved properties for estimation, statistical learning, and uncertainty quantification applications. Using statistical learning as an example, we demonstrate their advantage in training generative adversarial networks (GANs) for heavy-tailed, not-absolutely continuous sample distributions. We also show improved performance and stability over gradient-penalized Wasserstein GAN in image generation.

Keywords: f-divergences, Integral probability metrics, Wasserstein metric, Variational representations, GANs.

1. Introduction

Divergences and metrics provide a notion of 'distance' between multivariate probability distributions, thus allowing for comparison of models with one another and with data. Divergences are used in many theoretical and practical problems in mathematics, engineering, and the natural sciences, ranging from statistical physics, large deviations theory, uncertainty quantification and statistics to information theory, communication theory, and machine learning. In this work, we introduce and study what we term the (f, Γ)-divergences, denoted by D_f^Γ and defined by the variational expression

$$D_f^\Gamma(Q\|P) \equiv \sup_{g\in\Gamma} \left\{ E_Q[g] - \Lambda_f^P[g] \right\} , \tag{1}$$

$$\Lambda_f^P[g] \equiv \inf_{\nu\in\mathbb{R}} \left\{ \nu + E_P[f^*(g-\nu)] \right\} , \tag{2}$$

where Q and P are probability measures, f is a convex function with $f(1) = 0$, f^* denotes the Legendre Transform (LT) of f, and $\Gamma \subset \mathcal{M}_b(\Omega)$ is an appropriate function space.[1] The resemblance to the variational representation of the f-divergence is evident (see Equation 4 below), however, the additional optimization over shifts ν in (2), which is motivated by the Gibbs variational principle (Ben-Tal and Teboulle, 2007), will enable the derivation of many theoretical properties of the (f, Γ)-divergence. In the special case of the Kullback-Leibler (KL) divergence, $\Lambda_f^P[g]$ is exactly the cumulant generating function that arises in the Donsker-Varadhan variational formula (Dupuis and Ellis., 1997). We will show that the (f, Γ)-divergences are related to, interpolate between, and inherit key properties from both the f-divergences and the integral probability metrics (IPMs). To motivate the definition in (1), we first recall the definition and basic properties of f-divergences and IPMs.

The family of f-divergences includes among others the KL divergence (Kullback and Leibler, 1951), the total variation distance, the χ^2-divergence, the Hellinger distance, and the Jensen-Shannon (JS) divergence (Ali and Silvey, 1966; Csiszár, 1967). The f-divergence between two probability measures Q and P induced by a convex function f satisfying $f(1) = 0$ is defined by

$$D_f(Q\|P) \equiv E_P[f(dQ/dP)] . \tag{3}$$

This definition assumes absolute continuity between Q and P, $Q \ll P$, which in particular means that the support of Q is included in the support of P. The estimation of an f-divergence directly from (3) is challenging since it requires knowledge of the likelihood ratio (i.e., Radon-Nikodym derivative) dQ/dP, such as when working within a parametric family, or of a reasonable approximation to dQ/dP, usually through histogram binning, kernel density estimation (Wang et al., 2005; Kandasamy et al., 2015), or through k-nearest neighbor approximation (Wang et al., 2006). However, parametric methods greatly restrict the collection of allowed models, resulting in reduced expressivity, whereas non-parametric likelihood-ratio methods do not scale efficiently with the dimension of the data (Krishnamurthy et al., 2014). To address such challenges, statistical estimators which are based on variational representations of divergences have recently been introduced (Nguyen et al., 2010; Belghazi et al., 2018).

1. $\mathcal{M}_b(\Omega)$ denotes the set of all measurable and bounded real-valued functions on Ω.

Variational representation formulas for divergences, often referred to as dual formulations, convert divergence estimation into, in principle, an infinite-dimensional optimization problem over a function space. A typical example of a variational representation is the LT representation of the f-divergence between Q and P, given by (Broniatowski and Keziou, 2006; Nguyen et al., 2010)

$$D_f(Q\|P) = \sup_{g\in\mathcal{M}_b(\Omega)} \left\{E_Q[g] - E_P[f^*(g)]\right\}. \tag{4}$$

Such representations offer a useful mathematical tool to measure statistical similarity between data collections as well as to build, train, and compare complex probabilistic models. The main practical advantage of variational formulas is that an explicit form of the probability distributions or their likelihood ratio, dQ/dP, is not necessary. Only samples from both distributions are required since the difference of expected values in (4) can be approximated by statistical averages. In practice, the infinite-dimensional function space has to be approximated or even restricted. One of the first attempts was the restriction of the function space to a reproducing kernel Hilbert space (RKHS) and the corresponding kernel-based approximation in Nguyen et al. (2010). More recently, the optimization (4) has been approximated using flexible regression models and particularly by neural networks (Belghazi et al., 2018) and these techniques are widely used in the training of generative adversarial networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017; Nowozin et al., 2016; Gulrajani et al., 2017). Variational representations of divergences have also been used to quantify the model uncertainty in a probabilistic model (arising, e.g., from insufficient data and partial expert knowledge). For instance, applying the f-divergence formula (4) to $cg - \nu$, solving for $E_Q[g]$, and optimizing over $c > 0$, $\nu \in \mathbb{R}$ leads to the uncertainty quantification (UQ) bound (Chowdhary and Dupuis, 2013; Dupuis et al., 2016)

$$E_Q[g] \le \inf_{c>0}\left\{\frac{1}{c}\Lambda_f^P[cg] + \frac{1}{c}D_f(Q\|P)\right\}. \tag{5}$$

Similarly, one can obtain a corresponding lower bound for any quantity of interest $g \in \mathcal{M}_b(\Omega)$. The UQ inequality (5) bounds the uncertainty in the expectation of g under an alternative model Q in terms of expectations under the baseline model P and the discrepancy between Q and P (quantified via $D_f(Q\|P)$). Further discussion of the general connection between variational characterizations of divergences and UQ can be found in Glasserman and Xu (2014); Atar et al. (2015); Lam (2016); Breuer and Csiszár (2016); Gourgoulias et al. (2020); Dupuis et al. (2020); Dupuis and Mao (2019); Birrell et al. (2020).

Integral probability metrics are defined directly in terms of a variational formula (Müller, 1997; Sriperumbudur et al.), generalizing the Kantorovich-Rubinstein variational formula for the Wasserstein metric (Villani, 2008). More specifically, they are defined by maximizing the differences of respective expected values over a function space Γ,

$$W^\Gamma(Q,P) = \sup_{g\in\Gamma}\left\{E_Q[g] - E_P[g]\right\}, \tag{6}$$

and we refer to this object as the Γ-IPM. Despite the name, IPMs are not necessarily metrics in the mathematical sense unless further assumptions on Γ are made. This will not be an issue for us going forward, as we are not focused on the metric property; we will be

concerned with the divergence property, as defined in Section 2.1 below. Examples of IPMs include: the total variation metric, which is derived when the function space Γ is the unit ball in the space of bounded measurable functions; the Wasserstein, metric where Γ is the space of Lipschitz continuous functions with Lipschitz constant less than or equal to one; the Dudley metric, where the function space Γ is the unit ball in the space of bounded and Lipschitz continuous functions; and the maximum mean discrepancy (MMD), where Γ is the the unit ball in a RKHS, see also Müller (1997); Sriperumbudur et al.; Sriperumbudur et al. (2012). The definition of an IPM through the variational formula (6) leads to straightforward and unbiased statistical estimation algorithms (Sriperumbudur et al., 2012). Furthermore, the Wasserstein metric applied to generative adversarial networks (GANs) is known to substantially improve the stability of the training process (Arjovsky et al., 2017; Gulrajani et al., 2017), while MMD offers one of the most reliable two-sample tests for high dimensional statistical distributions (Gretton et al., 2012).

In summary, there are two fundamental mathematical ingredients involved in variational formulas for f-divergences and IPMs, with both families having their own strengths and weaknesses.

a) *The Objective Functional*: The objective functional in a variational representation is the quantity being maximized, namely $E_Q[g] - E_P[f^*(g)]$ for the f-divergences and $E_Q[g] - E_P[g]$ for the IPMs. The former depends on f and for appropriate f's it is strictly concave in g, while the latter is the same for all IPMs and is linear in g. Stronger convexity/concavity properties could result in improved statistical learning, estimation, and convergence performance. The ability to vary the objective functional by choosing f also allows one to tailor the divergence to the data source, e.g., for heavy tailed data. Finally, note that alternative objective functionals can yield the same divergence (Ben-Tal and Teboulle, 2007; Ruderman et al., 2012; Belghazi et al., 2018; Birrell et al., 2020), and their careful choice can have a substantial impact on their statistical estimation (Belghazi et al., 2018; Ruderman et al., 2012; Birrell et al., 2020).

b) *The Function Space*: This is the space over which the objective functional is optimized. In (4), it is the same function space for all f-divergences, namely $\mathcal{M}_b(\Omega)$, while the choice of function space Γ is what defines an IPM in (6). The choice of Γ has a profound impact on the properties of a divergence, e.g., the ability to meaningfully compare not-absolutely continuous distributions.

As we will show, the properties of the (f, Γ)-divergences can be tailored to the requirements of a particular problem through the choice of the objective functional (via f) and the function space Γ. The need for such a flexible family of divergences that combines the strengths of both f-divergences and IPMs is motivated by problems in machine learning and UQ, where properties of the data source or baseline model dictate the requirements on f and Γ, e.g., the f-divergence UQ bound (5) is unable to treat structurally different alternative models Q, which can easily be mutually singular with P, as $D_f(Q\|P) = \infty$ under a loss of absolute continuity; similar issues appear in GANs (Arjovsky et al., 2017).

Related approaches include the recent studies by Liu and Chaudhuri (2018); Farnia and Tse (2018); Miyato et al. (2018); Song and Ermon (2020); Husain et al. (2019); Dupuis and

Mao (2019); Glaser et al. (2021). In Miyato et al. (2018) the authors studied the use of spectral normalization to impose a Lipschitz constraint on the discriminator of a GAN; this is an example of (1) with a particular choice of function space. In Song and Ermon (2020), the authors proposed a class of objective functionals with an additional optimization layer, aiming to bridge the gap between the variational formulas for f-divergences and Wasserstein metrics and applied it to adversarial training of generative models. However, the paper does not provide a rigorous connection to the Wasserstein metric, since the function space appearing in their main Theorem 1 cannot include a Lipschitz constraint. This is in contrast to their practical implementation in Algorithm 1, which does employ a Lipschitz constraint. Our approach bridges this gap between theory and practice, as we are able to explicitly handle Lipschitz function spaces. Finally, our approach does not require the introduction of a third neural network, no matter what the choice of f-divergence may be. On the other hand, the authors in Dupuis and Mao (2019) developed a variational formula for general function spaces in the case of the KL divergence, providing a systematic and rigorous interpolation between KL divergence and IPMs. Definition (1) can be also viewed as a regularization of the classical f-divergences, and related objects have also been introduced and studied in Liu and Chaudhuri (2018); Husain et al. (2019); Farnia and Tse (2018); Glaser et al. (2021). While there is some overlap with several prior works, the aim of this paper is to provide a systematic and rigorous development of the (f, Γ)-divergences, focusing on a number of new properties that are potentially beneficial in learning and UQ applications. Specifically:

1. We derive conditions under which D_f^Γ has the divergence property, and thus provide a well-defined notion of 'distance' (Part 4 of Theorem 8 and Part 4 of Theorem 15). One key novelty is the introduction of the object (2) which is critical in the proof of this property.

2. We show that D_f^Γ interpolates between the f-divergence and Γ-IPM in the sense of infimal convolutions, including existence of an optimizer (Parts 1 and 2 of Theorem 15). Again, (2) plays a critical role here.

3. Using the infimal convolution formula, we derive a mass-redistribution/mass-transport interpretation of the (f, Γ)-divergences (Section 3).

4. We show that the family of (f, Γ)-divergences includes f-divergences and Γ-IPMs in suitable asymptotic limits (Theorem 17).

5. The relaxation of the hard constraint $g \in \Gamma$ in (1) to a soft-constraint penalty term is presented in Theorem 31. This is a generalization of the gradient penalty method for Wasserstein metrics (Gulrajani et al., 2017) to a much larger class of objective functionals and penalties and a key tool in designing numerically efficient implementations while still preserving the divergence property.

6. Relaxation of the condition $\Gamma \subset \mathcal{M}_b(\Omega)$ in (1), i.e., allowing Γ to contain appropriate unbounded functions, is addressed in Theorem 36. This is a necessary point when employing neural network estimation with unbounded activation functions.

7. We show that the (f, Γ)-divergences inherit several properties from both f-divergences and the IPMs. The primary advantage inherited from IPMs is the ability to compare

distributions which are not absolute continuous. The primary advantages inherited from the f-divergence are the strict concavity of the objective functional with respect to the test function, g, and the ability to compare heavy-tailed distributions (Section 6).

When combined, these advantages establish a divergence with better convergence and estimation properties. We numerically demonstrate these merits in the training of GANs. In Section 6.2, we show that the proposed divergence is capable of adversarial learning of lower dimensional sub-manifold distributions with heavy tails. In this example, both f-GAN (Nowozin et al., 2016) and Wasserstein GAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017) fail to converge or perform very poorly. Furthermore, in Section 6.3 we present improvements over WGAN-GP and WGAN with spectral-normalization (WGAN-SN) (Miyato et al., 2018), as measured by the inception score (Salimans et al., 2016) and FID score (Heusel et al., 2017) (two standard performance measures), in real data sets and particularly in CIFAR-10 (Krizhevsky, 2009) image generation. Interestingly, the training stability is significantly enhanced when using the proposed (f, Γ)-divergence, as compared to WGAN, which is evident from the fact that increasing the learning rate (i.e., stochastic gradient descent step size) eventually results in the collapse of WGAN but has comparatively little impact on our newly proposed method. We conjecture that this is due to the strict concavity of the objective functional of the (f, Γ)-divergence. We refer to these new proposed GANs which are based on (f, Γ)-divergences as (f, Γ)-GANs.

The organization of the paper is as follows. The key properties of the (f, Γ)-divergences are presented in Section 2. The mass-redistribution/mass-transport interpretation of the (f, Γ)-divergences is discussed in Section 3. Section 4 develops a general theory of soft-constraint penalization. Section 5 provides conditions under which the function space Γ can be expanded to contain unbounded functions. The application of the (f, Γ)-divergences in adversarial generative modeling is presented in Section 6. We conclude the paper and discuss plans for future work in Section 7. Finally, detailed proofs can be found in the appendices.

2. Construction and Properties of the (f, Γ)-Divergences

In this section, we will derive the divergence property for the (f, Γ)-divergences and show that they interpolate between f-divergences and IPMs as it is described in our main result (Theorem 15). First we introduce our notation and recall some important properties of the f-divergences.

2.1 Notation

For the remainder of the paper $(\Omega, \mathcal{M})$ will denote a measurable space, $\mathcal{M}(\Omega)$ will be the set of all measurable real-valued functions on Ω, $\mathcal{M}_b(\Omega)$ will denote the subspace of bounded measurable functions, $\mathcal{P}(\Omega)$ will denote the space of probability measures on $(\Omega, \mathcal{M})$, and $M(\Omega)$ will be the set of finite signed measures on $(\Omega, \mathcal{M})$. A subset $\Psi \subset \mathcal{M}_b(\Omega)$ will be called $\mathcal{P}(\Omega)$**-determining** if for all $Q, P \in \mathcal{P}(\Omega)$, $\int \psi dQ = \int \psi dP$ for all $\psi \in \Psi$ implies $Q = P$. The integral (expectation) of g with respect to $P \in \mathcal{P}(\Omega)$ will also be written as $E_P[g]$. We say that a map $D : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to [0, \infty]$ has the **divergence property**

Notation	Description	Reference
$(\Omega, \mathcal{M})$	Measurable space	Section 2.1
(S, d)	Metric space	Section 2.1
$M(\Omega)$ & $M(S)$	Spaces of finite signed measures	Section 2.1
$\mathcal{P}(\Omega)$ & $\mathcal{P}(S)$	Spaces of probability measures	Section 2.1
$\mathcal{M}(\Omega)$ & $\mathcal{M}_b(\Omega)$	Spaces of measurable real-valued functions	Section 2.1
$C(S)$ & $C_b(S)$	Spaces of continuous real-valued functions	Section 2.1
$\mathrm{Lip}(S)$ & $\mathrm{Lip}_b(S)$	Spaces of Lipschitz continuous functions	Section 2.1
P, Q	Probability distributions/measures	Section 2.1
f	Convex function on $\mathbb{R}$	Definition 2
$\mathcal{F}_1(a, b)$	Set of convex functions	Definition 2
D_f	f-Divergence	Eq. (7)
Λ_f^P	Generalized cumulant generating function	Eq. (9)
Γ	Test function space	Definition 5
D_f^Γ	(f, Γ)-Divergence	Eq. (15)
W^Γ	Γ-Integral probability metric	Eq. (16)
W^ρ	Gradient-penalty Wasserstein divergence	Eq. (38)
D_α^L	Lipschitz α-divergence	Eq. (43) - (44)

Table 1: List of main symbols used throughout the manuscript.

if $D(Q, P) = 0$ if and only if $Q = P$; such maps provide a notion of 'distance' between probability measures.

Remark 1 *We emphasize that despite the standard (but potentially confusing) terminology, not all f-divergences have the divergence property; see Section 2.2 below for further information. Going forward, we will continue to distinguish between what we call a divergence and the divergence property.*

(S, d) will denote a complete separable metric space (i.e., a Polish space), $C(S)$ will denote the space of continuous real-valued functions on S, and $C_b(S)$ will be the subspace of bounded continuous functions. $\mathrm{Lip}(S)$ will denote the space of Lipschitz functions on S, $\mathrm{Lip}_b(S)$ the subspace of bounded Lipschitz functions, and for $L > 0$ we let $\mathrm{Lip}_b^L(S)$ denote the subspace consisting of bounded L-Lipschitz functions (i.e., functions having Lipschitz constant L). $\mathcal{P}(S)$ will denote the space of Borel probability measures on S equipped with the Prokhorov metric, thus making $\mathcal{P}(S)$ a Polish space. Recall that the Prokhorov metric topology on $\mathcal{P}(S)$ is the same as the weak topology induced by the set of functions $\pi_g : P \mapsto E_P[g]$, $g \in C_b(S)$. For $\mu \in M(S)$ (finite signed Borel measures on S) we define $\tau_\mu : C_b(S) \to \mathbb{R}$ by $\tau_\mu(g) = \int g d\mu$ and we let $\mathcal{T} = \{\tau_\mu : \mu \in M(S)\}$. $\mathcal{T}$ is a separating vector space of linear functionals on $C_b(S)$. We equip $C_b(S)$ with the weak topology from $\mathcal{T}$ (i.e., the weakest topology on $C_b(S)$ for which every $\tau \in \mathcal{T}$ is continuous), which makes $C_b(S)$ a locally convex topological vector space with dual space $C_b(S)^* = \mathcal{T}$ (Rudin, 2006,

Theorem 3.10). We will let $\overline{\mathbb{R}} \equiv \mathbb{R} \cup \{-\infty, \infty\}$ denote the extended reals. Given a function $h : \mathbb{R} \to \overline{\mathbb{R}}$, its Legendre transform is defined by $h^*(y) \equiv \sup_{x\in\mathbb{R}}\{yx - h(x)\}$. Recall that if $h : \mathbb{R} \to (-\infty, \infty]$ is convex and lower semicontinuous (LSC) then $(h^*)^* = h$ (Bot et al., 2009, Theorem 2.3.5). Also recall that if h is convex and finite on (a, b) then the left and right derivatives, which we denote by $h'_-(x)$ and $h'_+(x)$ respectively, exist for all $x \in (a, b)$ (Roberts and Varberg, 1974, Chapter 1). We will denote the closure of a set A by $\overline{A}$ and its interior by A^o. Finally, we include in Table 1 a list of important notations, some of which are defined elsewhere in the manuscript, with corresponding references.

2.2 Background on f-Divergences

The f-divergences are constructed using functions of the following form:

Definition 2 *For a, b with $-\infty \le a < 1 < b \le \infty$ we define $\mathcal{F}_1(a, b)$ to be the set of convex functions $f : (a, b) \to \mathbb{R}$ with $f(1) = 0$. For $f \in \mathcal{F}_1(a, b)$, if b is finite we extend the definition of f by $f(b) \equiv \lim_{x\nearrow b} f(x)$. Similarly, if a is finite we define $f(a) \equiv \lim_{x\searrow a} f(x)$ (convexity implies these limits exist in $(-\infty, \infty]$). Finally, extend f to $x \notin [a, b]$ by $f(x) = \infty$. The resulting function $f : \mathbb{R} \to (-\infty, \infty]$ is convex and LSC.*

The f-divergences are then defined as follows:

Definition 3 *For $f \in \mathcal{F}_1(a, b)$ and $Q, P \in \mathcal{P}(\Omega)$ the corresponding f-divergence is defined by*

$$D_f(Q\|P) \equiv \begin{cases} E_P[f(dQ/dP)], & Q \ll P \\ \infty, & Q \not\ll P\,. \end{cases} \tag{7}$$

A number of important properties of f-divergences are collected in Appendix B. An f-divergence defines a notion of 'distance' between probability measures, as is made precise by the following divergence property: $D_f(Q\|P) \ge 0$ for all $f \in \mathcal{F}_1(a, b)$ and if f is furthermore strictly convex at 1 (i.e., f is not affine on any neighborhood of 1) then $D_f(Q\|P) = 0$ if and only if $Q = P$. However, the f-divergences are generally not probability metrics. Our primary examples will be the KL divergence and the family of α-divergences, which are constructed from the following functions:

$$f_{KL}(x) \equiv x\log(x) \in \mathcal{F}_1(0, \infty)\,, \quad f_\alpha(x) = \frac{x^\alpha - 1}{\alpha(\alpha - 1)} \in \mathcal{F}_1(0, \infty)\,, \text{where } \alpha > 0,\ \alpha \ne 1\,. \tag{8}$$

See Nowozin et al. (2016, Table 1) for further examples.

Key to our work are a pair of variational formulas that relate the f-divergence to the functional

$$\Lambda_f^P[g] \equiv \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g - \nu)]\}\,, \quad g \in \mathcal{M}_b(\Omega)\,. \tag{9}$$

As we will see, Λ_f^P takes the place of the cumulant generating function when one generalizes from the KL divergence to f-divergences. The first of the following formulas expresses D_f as an infinite-dimensional convex conjugate of Λ_f^P and the second is the dual variational formula.

1. Let $f \in \mathcal{F}_1(a,b)$ and $Q, P \in \mathcal{P}(\Omega)$. Then,

$$D_f(Q\|P) = \sup_{g \in \mathcal{M}_b(\Omega)} \{E_Q[g] - E_P[f^*(g)]\} \tag{10}$$

$$= \sup_{g \in \mathcal{M}_b(\Omega)} \{E_Q[g] - \Lambda_f^P[g]\}, \tag{11}$$

   where the second equality follows from (9) and (10) due to the invariance of $\mathcal{M}_b(\Omega)$ under the shift map $g \mapsto g - \nu$ for $\nu \in \mathbb{R}$; see also Proposition 50.

2. Let $f \in \mathcal{F}_1(a,b)$ with $a \geq 0$, $P \in \mathcal{P}(\Omega)$, and $g \in \mathcal{M}_b(\Omega)$. Then we can rewrite $\Lambda_f^P[g]$ as

$$\Lambda_f^P[g] = \sup_{Q \in \mathcal{P}(\Omega): D_f(Q\|P) < \infty} \{E_Q[g] - D_f(Q\|P)\}. \tag{12}$$

Remark 4 *f-divergences can alternatively be defined in terms of the densities of Q and P with respect to some common dominating measure (Liese and Vajda, 2006). This definition agrees with Eq.* (7) *when $Q \ll P$ but in some cases the definition in Liese and Vajda (2006) leads to a finite value even when $Q \not\ll P$. In this paper, we use the definition* (7) *because it satisfies the variational formula* (10)*, even when $Q \not\ll P$ (see the proof of Proposition 50), as well as the dual formula* (12)*.*

When $f = f_{KL}$ it is straightforward to show that Λ_f^P becomes the cumulant generating function,

$$\Lambda_{f_{KL}}^P[g] = \log E_P[e^g], \tag{13}$$

and Eq. (11) becomes the Donsker-Varadhan variational formula (Dupuis and Ellis., 1997, Appendix C.2). Subsequently, Eq. (12) becomes the Gibbs variational formula (Dupuis and Ellis., 1997, Proposition 1.4.2). For this reason, we will call (12) the Gibbs variational formula for f-divergences. Versions of Eq. (10) were proven in Broniatowski and Keziou (2006); Nguyen et al. (2010); we provide an elementary proof in Theorem 50 of Appendix B for completeness. Eq. (11) is implicitly found in Ruderman et al. (2012, Theorem 1); see Birrell et al. (2020) for further discussion of this relationship. More specifically, Ruderman et al. (2012); Birrell et al. (2020) show that when $a \geq 0$ the representation in (10) arises from convex duality over the space of finite positive measures while (11) arises from convex duality over the space of probability measures. On a metric space S, the optimizations in Equations (10) and (11) can be restricted to $C_b(S)$ via the application of Lusin's Theorem (see Corollary 51). The dual formula (12) was proven in Ben-Tal and Teboulle (2007) and is also implicitly contained in Ruderman et al. (2012, Equation 5) (we will require a generalization that also covers the case $a < 0$; see Proposition 57). Under appropriate assumptions (Broniatowski and Keziou, 2006, Theorem 4.4) the optimizer of (10) is given by

$$g_* = f'(dQ/dP). \tag{14}$$

The definition in (7) does not depend on the value of $f(x)$ for $x < 0$ and it is invariant under the transformation $f \mapsto f_c$ where $f_c(x) = f(x) + c(x-1)$, $c \in \mathbb{R}$. However, the objective functionals in the variational formulas (10) and (11) can depend on these choices due to

the presence of f^*. They both depend on the definition of $f(x)$ for $x < 0$. The identity $f_c^*(y) = f^*(y - c) + c$ implies that the objective functional in (10) depends on the choice of c but the objective functional in Eq. (11) does not. Substituting f_c into Eq. (10) and then taking the supremum over $c \in \mathbb{R}$ is another way to derive Eq. (11), thus providing additional motivation for the introduction of Λ_f^P.

2.3 Definition and General Properties of the (f, Γ)-Divergences

Motivated by Eq. (11) - (12), by working with subsets of test functions $\Gamma \subset \mathcal{M}_b(\Omega)$ we can construct a new family of so-called (f, Γ)-divergences whose convex conjugates at $g \in \Gamma$ equal $\Lambda_f^P[g]$ and that have variational characterizations akin to Eq. (11). This is an extension of the ideas in Dupuis and Mao (2019), which studied generalizations of the KL-divergence. *The identification of Λ_f^P as the proper replacement for the cumulant generating function is the key new insight required to extend from the KL case to general f.* Specifically, we make the following definition:

Definition 5 *Let $f \in \mathcal{F}_1(a, b)$ and $\Gamma \subset \mathcal{M}_b(\Omega)$ be nonempty. For $Q, P \in \mathcal{P}(\Omega)$ we define the (f, Γ)-divergence by*

$$D_f^\Gamma(Q\|P) \equiv \sup_{g \in \Gamma} \left\{ E_Q[g] - \Lambda_f^P[g] \right\} , \tag{15}$$

where Λ_f^P was defined in Eq. (9), and we define the Γ-IPM by

$$W^\Gamma(Q, P) \equiv \sup_{g \in \Gamma} \{E_Q[g] - E_P[g]\} . \tag{16}$$

When we want to emphasize the distinction between $D_f(Q\|P)$ and $D_f^\Gamma(Q\|P)$ we will refer to the former as a classical f-divergence. When f corresponds to the KL-divergence (see Equation 8) we write $R(Q\|P)$ and $R^\Gamma(Q\|P)$ in place of $D_f(Q\|P)$ and $D_f^\Gamma(Q\|P)$, respectively.

The definition (15) is an infinite-dimensional convex conjugate, akin to Eq. (11). From (11), we see that $D_f = D_f^\Gamma$ when $\Gamma = \mathcal{M}_b(\Omega)$ or, on a metric space S (and for appropriate f's), when $\Gamma = C_b(S)$ (see Corollary 51 and Remark 53). The W^Γ's are generalizations of the classical Wasserstein metric on a metric space, which is obtained by setting $\Gamma = \mathrm{Lip}_b^1(S)$. Neither W^Γ nor D_f^Γ necessarily have the divergence property, however, our main results present conditions which do imply the divergence property. As we will see, the use of Λ_f^P in (15) is crucial in our proof of the divergence property (see Theorem 8), as well as in our derivation of the infimal convolution formula (see Theorem 15).

One can alternatively write the (f, Γ)-divergence as

$$D_f^\Gamma(Q\|P) = \sup_{g \in \Gamma, \nu \in \mathbb{R}} \{E_Q[g - \nu] - E_P[f^*(g - \nu)]\} . \tag{17}$$

This formulation is useful when computing a numerical approximation to $D_f^\Gamma(Q\|P)$. It shows that Λ_f^P in (15) does not need to be computed separately; one can formulate the computation as a single optimization problem, incorporating one additional 1-dimensional

parameter. In addition, if Γ is closed under the shift transformations $g \mapsto g - \nu$, $\nu \in \mathbb{R}$ then one can write

$$D_f^\Gamma(Q\|P) = \sup_{g\in\Gamma} \{E_Q[g] - E_P[f^*(g)]\} \,, \tag{18}$$

thus arriving at the objects defined in Liu and Chaudhuri (2018); Husain et al. (2019); Farnia and Tse (2018). In the KL case, one can simplify Eq. (15) by using (13),

$$R^\Gamma(Q\|P) = \sup_{g\in\Gamma} \{E_Q[g] - \log E_P[e^g]\} \,, \tag{19}$$

which results in the special case studied in Dupuis and Mao (2019).

Several of our results will require us to work on a metric space (see Section 2.4), but first we present several properties that hold more generally. In the following theorem we derive a dual variational formula to (15), which shows that if $g \in \Gamma$ then Eq. (12) holds with D_f replaced by D_f^Γ. This lends further credence to the definition (15) and its use of Λ_f^P.

Theorem 6 *Let $f \in \mathcal{F}_1(a, b)$ where $a \geq 0$, $P \in \mathcal{P}(\Omega)$, and $\Gamma \subset \mathcal{M}_b(\Omega)$ be nonempty. For $g \in \Gamma$ we have*

$$(D_f^\Gamma)^*(g; P) \equiv \sup_{Q\in\mathcal{P}(\Omega)} \{E_Q[g] - D_f^\Gamma(Q\|P)\} = \Lambda_f^P[g] \,. \tag{20}$$

Remark 7 *We refer to Theorem 70 in Appendix C for the proof. While most cases of interest like Eq. (8) have $a \geq 0$, we also cover the case $a < 0$ in Theorem 70.*

Theorem 6 establishes D_f^Γ as a natural generalization of D_f when Γ is used as the test-function space, generalizing the dual formula (12) for f-divegences obtained in Ben-Tal and Teboulle (2007); Ruderman et al. (2012). Next we show that the D_f^Γ is bounded above by both D_f and W_Γ. This fact allows the (f, Γ)-divergences to inherit many useful properties from both f-divergences and IPMs; see the examples in Section 6. We also give conditions under which D_f^Γ has the divergence property and thus provides a notion of 'distance' between probability measures. This, along with Theorem 15 below, constitute the main theoretical results of this paper. The proof of Theorem 8 can be found in Theorem 71 of Appendix C.

Theorem 8 *Let $f \in \mathcal{F}_1(a, b)$, $\Gamma \subset \mathcal{M}_b(\Omega)$ be nonempty, and $Q, P \in \mathcal{P}(\Omega)$.*

1. $$D_f^\Gamma(Q\|P) \leq \inf_{\eta\in\mathcal{P}(\Omega)} \{D_f(\eta\|P) + W^\Gamma(Q, \eta)\} \,. \tag{21}$$

   In particular, $D_f^\Gamma(Q\|P) \leq \min\{D_f(Q\|P), W^\Gamma(Q, P)\}$.

2. *The map $(Q, P) \in \mathcal{P}(S) \times \mathcal{P}(S) \mapsto D_f^\Gamma(Q\|P)$ is convex.*

3. *If there exists $c_0 \in \Gamma \cap \mathbb{R}$ then $D_f^\Gamma(Q\|P) \geq 0$.*

4. *Suppose f and Γ satisfy the following:*

(a) There exist a nonempty set $\Psi \subset \Gamma$ with the following properties:

i. Ψ is $\mathcal{P}(\Omega)$-determining.

ii. For all $\psi \in \Psi$ there exists $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ such that $c_0 + \epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.

(b) f is strictly convex on a neighborhood of 1.

(c) f^* is finite and C^1 on a neighborhood of $\nu_0 \equiv f'_+(1)$.

Then:

(i) D_f^Γ has the divergence property.

(ii) W^Γ has the divergence property.

Remark 9 *Under stronger assumptions one can show that Eq. (21) is in fact an equality; see Theorem 15 below.*

Remark 10 *Assumptions 4(b) and 4(c) hold, for instance, if f is strictly convex on (a, b) and $\nu_0 \in \{f^* < \infty\}^o$; see Theorem 26.3 in Rockafellar (1970).*

Eq. (21) implies the following upper bound on D_f^Γ:

Corollary 11 (Upper Bounds) *Let $\mathcal{U} \subset \mathcal{P}(\Omega)$. Then*

$$D_f^\Gamma(Q\|P) \leq \inf_{\eta \in \mathcal{U}} \{D_f(\eta\|P) + W^\Gamma(Q, \eta)\}.$$

For instance, $\mathcal{U}$ could be a pushforward family, i.e., the distributions of $h_\theta(X)$, $\theta \in \Theta$ where h_θ are Ω-valued measurable maps and X is a random quantity. Such families are used in GANs; see Section 6.

Examples of $P(\Omega)$-determining sets:

1. Exponentials, $e^{c\cdot x}$, $c \in \mathbb{R}^n$, i.e., the moment generating function; see Section 30 in Billingsley (2012).
2. The set of 1-Lipschitz functions, g, on a metric space with $\|g\|_\infty \leq 1$. This follows from the Portmanteau Theorem; see, e.g., Theorem 2.1 in Billingsley (2013).
3. The unit ball of a reproducing kernel Hilbert space (RKHS), under appropriate assumptions; see Sriperumbudur et al. (2011).
4. The set of ReLU neural networks. This follows from the universal approximation theorem (Cybenko, 1989) and also applies to other activation functions, e.g., sigmoid.
5. The set of ReLU neural networks with spectral normalization (Miyato et al., 2018).

Several of these classes of functions have been used in existing methods; see Table 2 below. Our examples in Section 6 will use Lipschitz functions and ReLU neural networks, including spectral normalization in Section 6.3.1.

Remark 12 *Note that it is a well-known result that polynomials do not constitute a $\mathcal{P}(\Omega)$-determining set; there exist distinct measures that agree on all moments.*

Remark 13 *Depending on the domain, several of the above examples of $P(\Omega)$-determining sets consist of unbounded functions. To fit them into our framework it generally suffices to work with truncated versions of these functions; we refer to Section 5 for a detailed discussion.*

2.4 (f,Γ)-Divergences on Polish Spaces

When working on a Polish space, S, and under further assumptions on f and Γ, we are able to show that D_f^Γ interpolates between the classical f-divergence, D_f, and the Γ-IPM, W^Γ. At various points, we will require f and Γ to have the following properties:

Definition 14 *We will call $f \in \mathcal{F}_1(a,b)$* **admissible** *if $\lim_{y\to-\infty} f^*(y) < \infty$ (note that this limit always exists by convexity) and $\{f^* < \infty\} = \mathbb{R}$. If f is also strictly convex at 1 then we will call f* **strictly admissible**. *We will call $\Gamma \subset C_b(S)$* **admissible** *if $0 \in \Gamma$, Γ is convex, and Γ is closed in the weak topology generated by the maps τ_μ, $\mu \in M(S)$ (see Section 2.1). Γ will be called* **strictly admissible** *if it also satisfies the following property: There exists a $\mathcal{P}(S)$-determining set $\Psi \subset C_b(S)$ such that for all $\psi \in \Psi$ there exists $c \in \mathbb{R}$, $\epsilon > 0$ such that $c \pm \epsilon\psi \in \Gamma$.*

Our main result, Theorem 15, will require admissibility of both f and Γ. The functions f_{KL} and f_α, $\alpha > 1$, defined in Eq. (8), are strictly admissible but f_α, $\alpha \in (0,1)$ is not admissible (however, Theorem 8 above does apply to f_α for $0 < \alpha < 1$). The admissibility requirements that Γ be convex and closed will let us express D_f^Γ as the infinite-dimensional convex conjugate of a convex and LSC functional. This will allow us to analyze D_f^Γ using tools from convex analysis. Strict admissibility will be key in proving the divergence property for both W^Γ and D_f^Γ.

Examples of strictly admissible Γ:

1. $\Gamma = C_b(S)$, which leads to the classical f-divergences.
2. $\Gamma = \mathrm{Lip}_b^1(S)$, i.e., all bounded 1-Lipschitz functions, which leads to generalizations of the Wasserstein metric.
3. $\Gamma = \{g \in C_b(S) : |g| \le 1\}$, which leads to generalizations of the total variation metric.
4. $\Gamma = \{g \in \mathrm{Lip}_b^1(S) : |g| \le 1\}$, which leads to generalizations of the Dudley metric.
5. $\Gamma = \{g \in X : \|g\|_X \le 1\}$, the unit ball in a RKHS $X \subset C_b(S)$ (under appropriate assumptions given in Lemma 77). This yields a generalization of MMD and is also related to the recent KL-MMD interpolation method in Glaser et al. (2021); the latter employs a soft constraint rather than working on the RKHS unit ball and is based on the representation (10) instead of (11).

Note that the first two examples are shift invariant (hence Equation 18 is applicable) while the latter three are not.

We are now ready to present the second key theorem in this paper, where we derive the infimal convolution representation of D_f^Γ and provide alternative (to Theorem 8) conditions that ensure D_f^Γ possesses the divergence property. The proof can be found in Appendix C, Theorem 74.

Theorem 15 *Suppose f and Γ are admissible. For $Q, P \in \mathcal{P}(S)$ let $D_f^\Gamma(Q\|P)$ be defined by (15) and let $W^\Gamma(Q,P)$ be defined as in (16). These have the following properties:*

1. *Infimal Convolution Formula:*

$$D_f^\Gamma(Q\|P) = \inf_{\eta\in\mathcal{P}(S)}\{D_f(\eta\|P) + W^\Gamma(Q,\eta)\}\,. \tag{22}$$

In particular, $0 \le D_f^\Gamma(Q\|P) \le \min\{D_f(Q\|P), W^\Gamma(Q,P)\}$.

2. *If* $D_f^\Gamma(Q\|P) < \infty$ *then there exists* $\eta_* \in \mathcal{P}(S)$ *such that*

$$D_f^\Gamma(Q\|P) = D_f(\eta_*\|P) + W^\Gamma(Q,\eta_*)\,. \tag{23}$$

If f *is strictly convex then there is a unique such* η_*.

3. *Divergence Property for* W^Γ*: If* Γ *is strictly admissible then* W^Γ *has the divergence property.*

4. *Divergence Property for* D_f^Γ*: If* f *and* Γ *are both strictly admissible then* D_f^Γ *has the divergence property.*

Remark 16 *If* $a \ge 0$ *in Definition 2 then* f^* *is nondecreasing and so the condition* $\lim_{y\to-\infty} f^*(y) < \infty$ *is satisfied; see Lemma 46. In many cases, the divergence property for* D_f^Γ *still holds even if one or both of the conditions* $\lim_{y\to-\infty} f^*(y) < \infty$, $\{f^* < \infty\} = \mathbb{R}$ *are violated and also under relaxed conditions on* Γ*; this was shown in Theorem 8.*

The infimal convolution formula (22) - (23) gives one precise sense in which the (f,Γ)-divergence variationally interpolates between the Γ-IPM, W^Γ, and the classical f-divergence, D_f. It is a generalization of the results in Farnia and Tse (2018); Dupuis and Mao (2019), the former assuming compactly supported measures and the latter covering the KL case.

2.5 Additional Properties

The following theorem details the behavior of D_f^Γ in a pair of limiting regimes and further illustrates the manner in which D_f^Γ interpolates between D_f and W^Γ. These results again require (strict) admissibility (see Definition 14).

Theorem 17 *Let* $Q, P \in \mathcal{P}(S)$ *and* Γ*,* f *both be admissible. Then for all* $c > 0$ *the set* $\Gamma_c \equiv \{cg : g \in \Gamma\}$ *is admissible and we have the following two limiting formulas.*

1. *If* Γ *is strictly admissible then the sets* Γ_L *are strictly admissible for all* $L > 0$ *and*

$$\lim_{L\to\infty} D_f^{\Gamma_L}(Q\|P) = D_f(Q\|P)\,.$$

2. *If* f *is strictly admissible then*

$$\lim_{\delta\searrow 0} \frac{1}{\delta} D_f^{\Gamma_\delta}(Q\|P) = W^\Gamma(Q,P)\,.$$

The proof of Theorem 17 is very similar to that of the corresponding results in the KL case (Dupuis and Mao, 2019, Proposition 5.1 and 5.2). For completeness, we include its proof in Appendix C (Theorem 79).

Theorem 8 implies the following convergence and continuity properties (see Theorem 80 in Appendix C for the proof):

Theorem 18 *Let $f \in \mathcal{F}_1(a,b)$ and $\Gamma \subset \mathcal{M}_b(\Omega)$. Then:*

1. *If there exists $c_0 \in \Gamma \cap \mathbb{R}$ then $W^\Gamma(Q_n,P) \to 0 \implies D_f^\Gamma(Q_n\|P) \to 0$ and $D_f(Q_n\|P) \to 0 \implies D_f^\Gamma(Q_n\|P) \to 0$, and similarly if one permutes the order of Q_n and P.*
2. *Suppose f and Γ satisfy the following:*
   (a) *There exist a nonempty set $\Psi \subset \Gamma$ with the following properties:*
      i. *Ψ is $\mathcal{P}(\Omega)$-determining.*
      ii. *For all $\psi \in \Psi$ there exists $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ such that $c_0 + \epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.*
   (b) *f is strictly convex on a neighborhood of 1.*
   (c) *f^* is finite and C^1 on a neighborhood of $\nu_0 \equiv f'_+(1)$.*

   Let $P, Q_n \in \mathcal{P}(\Omega)$, $n \in \mathbb{Z}_+$. If $D_f^\Gamma(Q_n\|P) \to 0$ or $D_f^\Gamma(P\|Q_n) \to 0$ then $E_{Q_n}[\psi] \to E_P[\psi]$ for all $\psi \in \Psi$.
3. *On a metric space S, if f is admissible then the map $(Q,P) \in \mathcal{P}(S)\times\mathcal{P}(S) \mapsto D_f^\Gamma(Q\|P)$ is lower semicontinuous.*

Corollary 19 *Under the assumptions of Part 2 of Theorem 18 we have the following: If $\Gamma = Lip_b^1(S)$ where S is a compact metric space then one can take $\Psi = \Gamma$ and thereby conclude that $D_f^\Gamma(Q_n\|P) \to 0$ iff $D_f^\Gamma(P\|Q_n) \to 0$ iff $Q_n \to P$ in distribution iff $W^\Gamma(Q_n,P) \to 0$.*

Remark 20 *Corollary 19 follows from the equivalence between weak convergence and convergence in the Wasserstein metric on compact spaces; see Theorem 2 in Arjovsky et al. (2017) for this further relations between convergence in the Wasserstein metric and other notions of convergence.*

Finally, we derive a data processing inequality for (f,Γ)-divergences (see Theorem 81 in Appendix C for the proof). This result applies to general measurable spaces. We will need the following notation: Let $(N,\mathcal{N})$ be another measurable space and K be a probability kernel from Ω to N. Given $P \in \mathcal{P}(\Omega)$ we denote the composition of P with K by $P \otimes K$ (a probability measure on $\Omega \times N$) and we denote the marginal distribution on N by $K[P]$. Given $g \in \mathcal{M}_b(\Omega \times N)$ we let $K[g]$ denote the bounded measurable function on Ω given by $x \to \int g(x,y)K_x(dy)$.

Theorem 21 (Data Processing Inequality) *Let $f \in \mathcal{F}_1(a,b)$, $Q,P \in \mathcal{P}(\Omega)$, and K be a probability kernel from $(\Omega,\mathcal{M})$ to $(N,\mathcal{N})$.*

1. *Let $\Gamma \subset \mathcal{M}_b(N)$ be nonempty. Then*
$$D_f^\Gamma\left(K[Q]\|K[P]\right) \le D_f^{K[\Gamma]}(Q\|P)\,. \tag{24}$$
2. *Let $\Gamma \subset \mathcal{M}_b(\Omega \times N)$ be nonempty. Then*
$$D_f^\Gamma\left(Q \otimes K\|P \otimes K\right) \le D_f^{K[\Gamma]}(Q\|P)\,. \tag{25}$$

Remark 22 *In Eq. (24) we use the obvious embedding of $\mathcal{M}_b(N) \subset \mathcal{M}_b(\Omega \times N)$ to define $K[\Gamma] \equiv \{K[g] : g \in \Gamma\}$.*

We end this section by referring the reader to Table 2, which lists related works and connections to our general framework.

Extension of & connections to related work			
Related Paper	Function Space Γ	Objective Functional	Relevant Theorems
(Goodfellow et al., 2014)	Neural networks	JS divergence using (10)	Theorem 8
(Nowozin et al., 2016)	Neural networks	f-divergence using (10)	Theorem 8
(Belghazi et al., 2018)	Neural networks	KL-div. using (10) & (11)	Theorem 8
(Miyato et al., 2018)	Neural networks & spectral normalization	IPM (16) or f-divergence (10)	Theorem 8
(Arjovsky et al., 2017)	$\mathrm{Lip}_b^1(S)$	IPM (16)	Theorem 15
(Gulrajani et al., 2017)	$\mathrm{Lip}_b(S)$	IPM (16) & gradient penalty	Theorems 15 & 31
(Song and Ermon, 2020, Algorithm 1)	$\mathrm{Lip}_b^1(S)$	KL divergence using (10)	Theorem 15
(Nguyen et al., 2010)	RKHS	KL, f-divergence using (10)	Theorem 15
(Gretton et al., 2012)	Unit ball in RKHS	IPM (16)	Theorem 15
(Glaser et al., 2021)	RKHS	KL-div. using (10) & RKHS norm penalty	Theorems 15 & 31
(Dupuis and Mao, 2019)	convex & closed Γ	KL-divergence	Theorem 15

Table 2: Summarizing how our main theorems extend or relate to certain existing methods. Our theory either applies directly to the cited methods or motivates the construction of closely related interpolation and/or regularization methods that are based on (11).

3. Mass Redistribution/Transport Interpretation of (f,Γ)-Divergences

The bound

$$D_f^\Gamma(Q\|P) \le W^\Gamma(Q,P)\,, \tag{26}$$

which follows from Part 1 of either Theorem 8 or Theorem 15, makes it clear that $D_f^\Gamma(Q\|P)$ can be finite and informative even if $Q \not\ll P$. For instance, if $\Gamma = \mathrm{Lip}_b^1(S)$ then W^Γ is the classical Wasserstein metric, and this can be finite even for mutually singular Q and P. It is well-known that the Wasserstein metric can be understood in terms of mass transport (Villani, 2008). Generalizing this idea, the variational formula (23) allows us to interpret the (f,Γ)-divergences in terms of a two-stage mass-redistribution/mass-transport process:

1. First the 'mass' distribution, P, is redistributed to form an intermediate measure, η_*. This has cost $D_f(\eta_*\|P)$, which depends on the relative amount of mass moved from or added to each point, but is insensitive to the distance that the mass is moved. However, the support of η_* cannot be enlarged or shifted outside the support of P during its construction, otherwise the cost would be infinite.

2. Next, the mass is transported from η_* to Q with a cost $W^\Gamma(Q,\eta_*)$ that depends on the distance the mass must be moved. In this step, the support of η_* could be drastically different from the support of Q, if necessary.

The optimizing η_* achieves the optimal balance between the cost of redistributing mass in step 1 and the cost of transporting mass in step 2.

Remark 23 *When $\Gamma \neq Lip_b^1(S)$, D_f^Γ is still characterized by the above two-stage procedure, with the only difference being that the interpretation of W^Γ may differ.*

In this section we derive a characterization of the solution to the infimal convolution problem (22) in the case where $f \in \mathcal{F}_1(a,b)$ with $a \ge 0$ and will use this to provide further insight into the mass-redistribution/mass-transport interpretation. A key step will be to first obtain existence and uniqueness results regarding the dual optimization problem (12) for the classical f-divergences. The proof is found in Appendix C, Theorem 82.

Theorem 24 *Let $P \in \mathcal{P}(\Omega)$, $g \in \mathcal{M}_b(\Omega)$, and $f \in \mathcal{F}_1(a,b)$ be admissible with $a \ge 0$. If f is strictly convex on (a,b) then there exists $\nu_* \in \mathbb{R}$ such that*

$$dQ_* \equiv (f^*)'(g-\nu_*)dP$$

is a probability measure and

$$\sup_{Q\in\mathcal{P}(\Omega)}\{E_Q[g] - D_f(Q\|P)\} = E_{Q_*}[g] - D_f(Q_*\|P) = \nu_* + E_P[f^*(g-\nu_*)] = \Lambda_f^P[g]\,.$$

Moreover, Q_* is the unique solution to the optimization problem

$$\sup_{Q\in\mathcal{P}(\Omega)}\{E_Q[g] - D_f(Q\|P)\}\,. \tag{27}$$

Theorem 24 (specifically, the generalization found in Theorem 82) allows us to derive in Theorem 25 a characterization of the solution, η_*, to the infimal convolution problem (22). First we present a formal calculation; a precise statement of the result can be found in Theorem 25 and a rigorous proof is given in Theorem 83 of Appendix C. This result generalizes Theorem 4.12 in Dupuis and Mao (2019), which considered the KL case: First assume (g_*, ν_*) is a maximizer of (15), and assume η_* solves (22). Then

$$\begin{aligned} D_f^\Gamma(Q\|P) &= E_Q[g_*] - (\nu_* + E_P[f^*(g_* - \nu_*)]) \\ &= E_Q[g_*] - E_{\eta_*}[g_*] + E_{\eta_*}[g_*] - (\nu_* + E_P[f^*(g_* - \nu_*)]) \\ &\leq W^\Gamma(Q, \eta_*) + D_f(\eta_*\|P) = D_f^\Gamma(Q\|P)\,. \end{aligned} \tag{28}$$

Therefore, as the inequalities become equalities, we have

$$W^\Gamma(Q, \eta_*) = E_Q[g_*] - E_{\eta_*}[g_*]$$

and

$$D_f(\eta_*\|P) = E_{\eta_*}[g_*] - (\nu_* + E_P[f^*(g_* - \nu_*)])\,. \tag{29}$$

Note that this also implies $E_P[(f^*)'(g_* - \nu_*)] = 1$ and

$$\begin{aligned} d\eta_* &= (f^*)'(g_* - \nu_*)dP\,, \\ g_* &= f'(d\eta_*/dP) + \nu_* \quad P\text{-a.s.} \end{aligned}$$

In particular, in the KL case (Dupuis and Mao, 2019, Remark 4.11), one has

$$g_* = \log(d\eta_*/dP) + c_0 \quad P\text{-a.s.}$$

for some $c_0 \in \mathbb{R}$ and, if $Q \ll P$, this leads to

$$R^\Gamma(Q\|P) = E_Q[\log(d\eta_*/dP)]\,, \tag{30}$$

which has an obvious similarity to the formula for the classical KL divergence.

Theorem 25 *Let $\Gamma \subset C_b(S)$ be admissible and $f \in \mathcal{F}_1(a, b)$ be admissible, where $a \geq 0$ and f^* is C^1. Fix $Q, P \in \mathcal{P}(S)$ and suppose we have $g_* \in \Gamma$ and $\nu_* \in \mathbb{R}$ that satisfy the following:*

1. *$f((f^*)'(g_* - \nu_*)) \in L^1(P)$,*
2. *$E_P[(f^*)'(g_* - \nu_*)] = 1$,*
3. *$W^\Gamma(Q, \eta_*) = E_Q[g_*] - E_{\eta_*}[g_*]$, where $d\eta_* \equiv (f^*)'(g_* - \nu_*)dP$.*

Then $\eta_* \in \mathcal{P}(S)$ solves the infimal convolution problem (22) and

$$D_f^\Gamma(Q\|P) = E_Q[g_*] - (\nu_* + E_P[f^*(g_* - \nu_*)])\,. \tag{31}$$

If f is strictly convex then η_* is the unique solution to the infimal convolution problem.

Remark 26 *In the context of MMD, g_* is called the witness function (Gretton et al., 2012). In the KL case, the existence of g_* can be proven under appropriate compactness assumptions (Dupuis and Mao, 2019, Theorem 4.8).*

Remark 27 *Eq.* (23) *from Theorem 15 makes it clear that $D_f^\Gamma(Q\|P) < D_f(Q\|P)$ in 'most' cases. An exception to this occurs when Eq.* (10) *has an optimizer g_* with $g_* \in \Gamma$. In such cases we have $D_f^\Gamma(Q\|P) = D_f(Q\|P)$, the supremum* (15) *will also be achieved at g_* since Eq.* (31) *holds with $\nu_* = 0$, and the solution to the infimal convolution problem is $\eta_* = Q$.*

In general, the task of computing the intermediate measure η_* in (23) is difficult, though a naive approach could proceed as follows:

1. Approximate $\eta \in \mathcal{P}(S)$ by a neural network family $h_\theta(X)$, where X is some random noise source (as in the generator of a GAN; see Section 6); in this step we are using Corollary 11 to construct an upper bound.
2. Approximate $D_f(\eta\|P)$ and $W^\Gamma(Q,\eta)$ via their variational formulas (10) or (11) and (16) respectively, with the function spaces being approximated via neural network families (as in the discriminator of a GAN; again, see Section 6).
3. Solve the resulting min-max problem (22) via a stochastic-gradient-descent method to approximate η_* (and also g_*).

We did not explore the effectiveness of this naive method here, as it is tangential to the goals of this paper; we leave the computation of η_* for a future work. Nevertheless, the following subsection presents a simple example that provides useful intuition.

3.1 Example: Dirac Masses

Here we consider a simple example involving Dirac masses where the (f,Γ)-divergence can be explicitly computed using Theorem 25. This example further illustrates the two-stage mass-redistribution/mass-transport interpretation of the infimal convolution formula (23) and demonstrates how the location and distribution of probability mass impacts the result; see Figure 1. Further explicit examples in the KL case can be found in Dupuis and Mao (2019).

Let $0 = x_1 < x_2 < x_3$ and define the uniform distributions

$$P = \frac{1}{2}\delta_{x_1} + \frac{1}{2}\delta_{x_2}\,, \quad Q = \frac{1}{3}\delta_{x_1} + \frac{1}{3}\delta_{x_2} + \frac{1}{3}\delta_{x_3}\,. \tag{32}$$

Note that $Q \not\ll P$ and so $D_f(Q\|P) = \infty$; we will see that the (f,Γ)-divergences can be finite. Specifically, we will compute the $(f_\alpha, \mathrm{Lip}_b^1(\mathbb{R}))$-divergence for $\alpha > 1$ via Theorem 25. To do this we must find $g_* \in \mathrm{Lip}_b^1(\mathbb{R})$ and $\nu_* \in \mathbb{R}$ such that

$$\frac{1}{2}(f_\alpha^*)'(g_*(x_1) - \nu_*) + \frac{1}{2}(f_\alpha^*)'(g_*(x_2) - \nu_*) = 1\,, \tag{33}$$

$$g_* \in \operatorname*{argmax}_{g \in \mathrm{Lip}_b^1(\mathbb{R})} \left\{ \frac{1}{3}(g(x_1) + g(x_2) + g(x_3)) - \frac{1}{2}\left(g(x_1)(f_\alpha^*)'(g_*(x_1) - \nu_*) \right.\right. \tag{34}$$

$$\left.\left. + g(x_2)(f_\alpha^*)'(g_*(x_2) - \nu_*) \right) \right\},$$

where

$$(f_\alpha^*)'(y) = (\alpha-1)^{1/(\alpha-1)} y^{1/(\alpha-1)} 1_{y>0}$$

(see Eq. (60)); Eq. (33) is a simplification of Assumption 2 from Theorem 25 while Eq. (34) corresponds to Assumption 3. The solution to the infimal convolution problem then has the form

$$d\eta_* = \frac{1}{2}(f_\alpha^*)'(g_*(x_1)-\nu_*)\delta_{x_1} + \frac{1}{2}(f_\alpha^*)'(g_*(x_2)-\nu_*)\delta_{x_2}\,. \tag{35}$$

We will now outline how one solves for ν_* and g_*. Without loss of generality we can assume $g_*(x_1) = 0$ (the objective functional for W^Γ is invariant under constant shifts and at the same time, shifting g_* in η_* can be achieved by redefining ν_*). The only dependence on $g(x_3)$ in Eq. (34) is in the $g(x_3)/3$ term, hence the optimal solution has $g(x_3) = x_3 - x_2 + g(x_2)$. Therefore we need to solve

$$\begin{aligned} &\frac{1}{2}(f_\alpha^*)'(-\nu_*) + \frac{1}{2}(f_\alpha^*)'(g_*(x_2)-\nu_*) = 1\,, \\ &g_*(x_2) \in \operatorname*{argmax}_{g(x_2)\in[-x_2,x_2]} \left\{ \frac{1}{3}(x_3-x_2) + \left(\frac{2}{3} - \frac{1}{2}(f_\alpha^*)'(g_*(x_2)-\nu_*)\right) g(x_2) \right\} \end{aligned} \tag{36}$$

for ν_* and $g_*(x_2)$. The solution to this is obtained as follows:

1. Let $\nu_*(g_2)$ be the unique solution to $\frac{1}{2}(f_\alpha^*)'(-\nu_*) + \frac{1}{2}(f_\alpha^*)'(g_2-\nu_*) = 1$; the two terms on the left hand side will be used to obtain the redistributed weights in η_*.

2. Take $g_{*,2}$ such that $\frac{1}{2}(f_\alpha^*)'(g_{*,2} - \nu_*(g_{*,2})) = 2/3$; this is inspired by the second line in Eq. (36).

3. If $0 < x_2 < g_{*,2}$ then the solution to Eq. (36) is obtained at $\nu_* = \nu_*(x_2)$ and

$$g_*(x) = \begin{cases} 0, & x < 0 \\ x, & x \in [0, x_3) \\ x_3, & x \geq x_3\,. \end{cases}$$

   In this case, the optimal solution has $1/3 < \eta_*(x_2) < 2/3$, i.e., some amount of mass is redistributed from $x_1 = 0$ to x_2 when forming η_* and then mass is transported from both x_1 and x_2 to x_3 to form Q.

4. If $x_2 \geq g_{*,2}$ then the solution to Eq. (36) is obtained at $\nu_* = \nu_*(g_{*,2})$ and

$$g_*(x) = \begin{cases} 0, & x < 0 \\ \frac{g_{*,2}}{x_2} x, & x \in [0, x_2) \\ x - x_2 + g_{*,2}, & x \in [x_2, x_3) \\ x_3 - x_2 + g_{*,2}, & x \geq x_3\,. \end{cases}$$

   In this case, x_2 is sufficiently far away from $x_1 = 0$ that the optimal solution, η_*, is obtained by first redistributing mass from $x_1 = 0$ to x_2 so that $\eta_*(x_1) = 1/3$, $\eta_*(x_2) = 2/3$. In the second step, mass is transported solely from x_2 to x_3 in order to form Q.

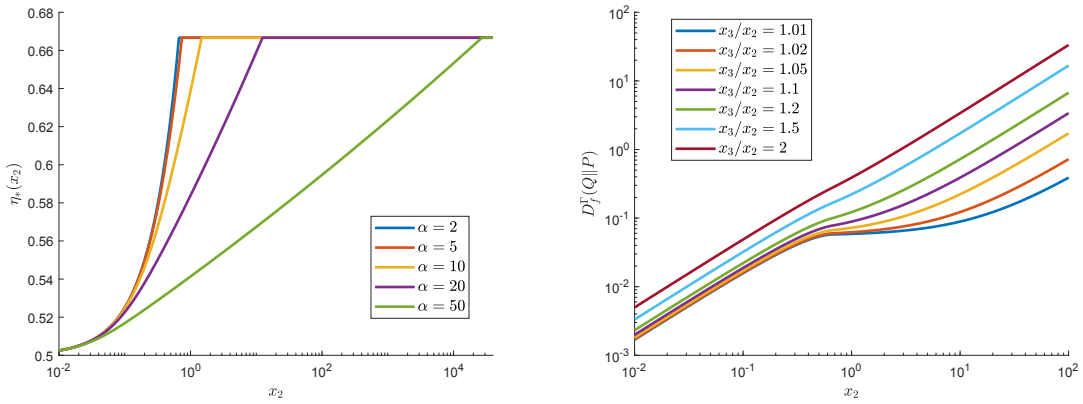


Figure 1: Solution of the infimal convolution problem (23) for $D^\Gamma_{f_\alpha}(Q\|P)$, where $\Gamma = \mathrm{Lip}_b^1(\mathbb{R})$ and Q and P are given by Eq. (32). The left panel shows the mass $\eta_*(x_2)$ as a function of x_2. For each value of α there is a transition point where all of the mass required by Q at x_3 is first redistributed to x_2 when forming η_*, resulting in $\eta_*(x_2) = 2/3$. Note that the amount of mass moved to x_2 in the redistribution step does not depend on the distance of x_3 from x_2, only on the distance of x_2 from $x_1 = 0$. The right panel shows $D^\Gamma_{f_2}(Q\|P)$ as a function of x_2 and for several different values of the ratio x_3/x_2.

This completes the construction of η_* from Eq. (35). The value of the $(f_\alpha, \mathrm{Lip}_b^1(\mathbb{R}))$-divergence can then be computed via Eq. (31). The computation of $g_{*,2}$ and $\nu_*(g_{*,2})$ from steps 1 and 2 must be done numerically and so we illustrate the solution graphically in Figure 1 by plotting $\eta_*(x_2)$ as a function of x_2 for a number of α's. This shows how the mass must be redistributed when forming η_* from P. The above calculations reveal an interesting transition; when x_2 is not close [2] to x_1 then the mass is transferred solely from x_2 after it has been redistributed from x_1. However, when x_1 and x_2 are close enough then redistributing all the necessary mass from x_1 to x_2 is not optimal and it is cheaper to transport probability mass from both x_1 and x_2 to x_3. The transition between these cases corresponds to the point where x_2 crosses above $g_{*,2}$ (which depends on α) and hence $\eta_*(x_2)$ saturates at the value $2/3$.

4. Soft Constraints and the Divergence Property

For computational purposes, it is often advantageous to replace the hard (i.e., strict) constraint $g \in \Gamma$ with a soft constraint in the form of a penalty term, V, subtracted from the objective functional; by a penalty term, we mean V 'activates' (i.e., is nonzero) when the constraint $g \in \Gamma$ is violated. In this way we can construct a new divergence D_f^V with $D_f^\Gamma \leq D_f^V \leq D_f$ (we let the context distinguish between cases where the superscript denotes

2. Here, 'closeness' depends not only on the distance between the two points but also on f.

a constraint space and cases where it denotes a penalty term); see Theorem 31 for the main result of this section.

Of particular interest is the case $\Gamma = \mathrm{Lip}_b^1(\mathbb{R}^n)$ (we equip $\mathbb{R}^n$ with the Euclidean metric), where the 1-Lipschtiz constraint can be relaxed to a one-sided gradient penalty term, thus defining objects such as

$$D_f^\rho(Q\|P) = \sup_{g\in \mathrm{Lip}_b(\mathbb{R}^n)} \left\{ E_Q[g] - \Lambda_f^P[g] - \lambda \int \max\{0, \|\nabla g\|^2 - 1\} d\rho_{Q,P} \right\}, \tag{37}$$

where $\lambda > 0$ is the strength of the penalty term and $\rho_{Q,P}$ is a positive measure (often depending on Q and P). Here we are relying on Rademacher's theorem (see Theorem 5.8.6 in Evans, 2010): L-Lipschitz functions on $\mathbb{R}^n$ are differentiable Lebesgue-a.e. and the norm of the gradient is bounded by L. The penalty term in Eq. (37) will therefore be activated only when g is not 1-Lipschitz.

Divergences with soft Lipschitz constraints were first applied to Wasserstein GAN (Gulrajani et al., 2017) with great success, but the theoretical properties of such objects have not been explored; specifically, it has not been shown that they satisfy the divergence property. Here we show in great generality that the relaxation of a hard constraint to a soft constraint preserves the divergence property, and therefore objects such as (37) still provide a well-defined notion of 'distance' between probability measures. The basic requirement is that the penalty term, which we denote by V, vanishes on the constraint space Γ.

Lemma 28 *Let $(\Omega, \mathcal{M})$ be a measurable space, $\Gamma \subset \widetilde{\Gamma} \subset \mathcal{M}(\Omega)$, $H : \widetilde{\Gamma} \times \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \overline{\mathbb{R}}$, and $V : \widetilde{\Gamma} \times \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to [0, \infty]$ with $V|_{\Gamma\times\mathcal{P}(\Omega)\times\mathcal{P}(\Omega)} = 0$. Define*

$$D^\Gamma(Q\|P) = \sup_{g\in\Gamma} H[g; Q, P], \quad D^{\widetilde{\Gamma}}(Q\|P) = \sup_{g\in\widetilde{\Gamma}} H[g; Q, P],$$
$$D^V(Q\|P) = \sup_{g\in\widetilde{\Gamma}} \{H[g; Q, P] - V[g; Q, P]\},$$

where $\infty - \infty \equiv -\infty$. If D^Γ and $D^{\widetilde{\Gamma}}$ both have the divergence property then so does D^V.

Remark 29 *The convention $\infty - \infty \equiv -\infty$ is simply a convenient rigorous shorthand for restricting the supremum to those g's for which this generally undefined operation does not occur.*

Remark 30 *More generally, if the supremum $\sup_{g\in\Gamma} H[g; Q, P]$ is achieved at $g_* \in \Gamma$ (depending on Q, P) then the requirement $V|_{\Gamma\times\mathcal{P}(\Omega)\times\mathcal{P}(\Omega)} = 0$ can be relaxed to $V[g_*; Q, P] = 0$ for all Q, P.*

Proof Using $\Gamma \subset \widetilde{\Gamma}$, $V \geq 0$, and $V|_\Gamma = 0$ we have $D^\Gamma \leq D^V \leq D^{\widetilde{\Gamma}}$. D^Γ satisfies the divergence property, hence is non-negative. Therefore $D^V \geq 0$. $D^{\widetilde{\Gamma}}$ has the divergence property, hence if $Q = P$ then $0 = D^{\widetilde{\Gamma}}(Q\|P) \geq D^V(Q\|P) \geq 0$. Therefore $D^V(Q\|P) = 0$. Finally, if $D^V(Q\|P) = 0$ then $D^\Gamma(Q\|P) = 0$ and hence the divergence property for D^Γ implies $Q = P$. ■

Using Theorem 15 and Corollary 68, we can apply Lemma 28 to the (f, Γ)-divergences and thereby conclude the following:

Theorem 31 *Let f and $\Gamma \subset C_b(S)$ be strictly admissible. Let $\Gamma \subset \widetilde{\Gamma} \subset \mathcal{M}(S)$ and $V : \widetilde{\Gamma} \times \mathcal{P}(S) \times \mathcal{P}(S) \to [0,\infty]$ with $V|_{\Gamma\times\mathcal{P}(S)\times\mathcal{P}(S)} = 0$. For $Q, P \in \mathcal{P}(S)$ define*

$$D_f^V(Q\|P) \equiv \sup_{g\in\widetilde{\Gamma}} \left\{ \left(E_Q[g] - \Lambda_f^P[g]\right) - V[g;Q,P] \right\} ,$$

where $\infty - \infty \equiv -\infty$, $-\infty + \infty \equiv -\infty$. Then D_f^V has the divergence property and $D_f^\Gamma \leq D_f^V \leq D_f$.

Proof Combine Lemma 28 with Part 4 of Theorem 15 and Theorem 65 below; the latter shows that the variational formula for D_f also holds when using the test-function space $\mathcal{M}(\Omega)$. ■

4.1 Soft-Lipschitz Constraints on $\mathbb{R}^n$: One-Sided Versus Two-Sided Penalties

The gradient penalty term in Eq. (37) is one-sided, meaning that it penalizes $\|\nabla g\| > 1$ but not $\|\nabla g\| \leq 1$. This is consistent with the hard constraint that the Lipschitz constant be less than or equal to 1. The first use of soft Lipschitz penalties in Gulrajani et al. (2017), which considered the Wasserstein metric, also used a two-sided gradient penalty,

$$W^\rho(Q,P) = \sup_{g\in \mathrm{Lip}_b(\mathbb{R}^n)} \left\{ E_Q[g] - E_P[g] - \lambda \int (\|\nabla g\| - 1)^2 d\rho_{Q,P} \right\} , \tag{38}$$

which penalizes $\|\nabla g\| \neq 1$. An intuitively reasonable requirement to impose on any soft constraint is that it vanish on the exact optimizer (if one exists) of the original strictly-constrained optimization problem. The justification for a two-sided gradient penalty in the Wasserstein case rests on Proposition 1 in Gulrajani et al. (2017), which shows that the exact optimizer of the Kantorovich-Rubinstein variational formula for the classical Wasserstein metric has gradient with norm 1 a.e. As the two-sided gradient penalty vanishes on such functions, the object (38) will still possess the divergence property (see Remark 30). However, two-sided gradient penalties are not appropriate constraint-relaxations of the (f,Γ)-divergences, as the gradient of the exact optimizer generally does not have norm 1 a.e. We demonstrate this via the following simple counterexample: Let $\Gamma = \mathrm{Lip}_b^1(\mathbb{R}^n)$, $P \in \mathcal{P}(\mathbb{R}^n)$, and define Q by $dQ/dP = Z^{-1} e^{-\min\{\|x\|,1\}/2}$. The optimizer of the variational formula defined in (14) is given for the classical KL divergence by

$$g_* = \log(dQ/dP) + 1 = -\min\{\|x\|, 1\}/2 + 1 + \log(Z^{-1}) ,$$

which is bounded and 1/2-Lipschitz, and so $g_* \in \Gamma$. Therefore it is straightforward to see that g_* is also the optimizer for $R^\Gamma(Q\|P)$ and it satisfies $\|\nabla g_*\| \leq 1/2$ a.e. This proves that the 2-sided penalty does not vanish on g_*. Similar counterexamples can be constructed using Eq. (14) for other choices of f.

5. Extension of the (f,Γ)-Divergence Variational Formula to Unbounded Functions

The assumption that all of the test functions $g \in \Gamma$ are bounded can be very restrictive in practice. In this section we provide general conditions under which the test-function space

can be expanded to include (possibly) unbounded functions without changing the value of D_f^Γ. This fact will be used in the numerical examples in Section 6 below. The main result in this Section in Theorem 36.

The key step in the extension to unbounded g's is the following lower bound.

Lemma 32 *Let f, Γ be admissible and, in addition, suppose f^* is bounded below. Fix $Q, P \in \mathcal{P}(S)$. If $g \in L^1(Q)$ and there exists $g_n \in \Gamma$, a measurable set A, and $C \in \mathbb{R}$ with $g_n \to g$ pointwise, $|g_n| \leq |g|$ for all n, and $g_n \leq g1_A + C1_{A^c}$ for all n, then*

$$D_f^\Gamma(Q\|P) \geq E_Q[g] - \Lambda_f^P[g]\,.$$

Remark 33 *The additional assumption that f^* is bounded below is satisfied in many cases of interest, e.g., the KL divergence and α-divergences for $\alpha > 1$.*

Proof We need to show that

$$D_f^\Gamma(Q\|P) \geq E_Q[g] - (\nu + E_P[f^*(g-\nu)])$$

for all $\nu \in \mathbb{R}$. Note that we have assumed f^* is bounded below by some $D \in \mathbb{R}$, hence $E_P[f^*(g-\nu)]$ exists in $(-\infty, \infty]$. If $E_P[f^*(g-\nu)] = \infty$ then the claim is trivial, so for the remainder of this proof we suppose $f^*(g-\nu) \in L^1(P)$.

The assumptions on g allow us to use the dominated convergence theorem to conclude $E_Q[g_n] \to E_Q[g]$. Continuity of f^* implies $f^*(g_n - \nu) \to f^*(g-\nu)$. The admissibility assumption implies $\lim_{y\to-\infty} f^*(y) < \infty$. Using this together with Lemma 44 we see that f^* is nondecreasing, hence

$$D \leq f^*(g_n - \nu) \leq f^*(g-\nu)1_A + f^*(C-\nu)1_{A^c} \in L^1(P)\,.$$

Therefore the dominated convergence theorem implies $E_P[f^*(g_n-\nu)] \to E_P[f^*(g-\nu)]$. We have $g_n \in \Gamma$, hence Eq. (15) implies

$$\begin{aligned} D_f^\Gamma(Q\|P) \geq& \lim_{n\to\infty} (E_Q[g_n] - (\nu + E_P[f^*(g_n-\nu)])) \\ =& E_Q[g] - (\nu + E_P[f^*(g-\nu)])\,. \end{aligned}$$

This completes the proof. ■

Using Lemma 32, one can augment Γ by including any functions that satisfy the stated assumptions; this will not change the value of the supremum in (15). Rather than formulating a general result of this type, we consider one of the most useful special cases, the set of Lipschitz functions. Other cases can be treated similarly.

Lemma 34 *Let $c : S \times S \to [0,\infty]$, $L \in (0,\infty)$, and define*

$$Lip_b^L(S,c) = \{g \in C_b(S) : |g(x) - g(y)| \leq Lc(x,y) for\ all x, y \in S\}\,. \tag{39}$$

If $c = d$ (the metric on S) then we use our earlier notation, $Lip_b^L(S)$, in place of $Lip_b^L(S,d)$.

The set $Lip_b^L(S,c)$ is admissible and if $d \leq Kc$ for some $K \in (0,\infty)$ then $Lip_b^L(S,c)$ is strictly admissible.

Proof Convexity is trivial. Weak convergence in $C_b(S)$ implies pointwise convergence (take $\mu = \delta_x$, $x \in S$), hence $\mathrm{Lip}_b^L(S,c)$ is closed. Finally, if $d \le Kc$ then strict admissibility follows from the fact that $\mathrm{Lip}_b^{L/K}(S)$ is $\mathcal{P}(S)$-determining and $\mathrm{Lip}_b^{L/K}(S) \subset \mathrm{Lip}_b^L(S,c)$. ■

Remark 35 *For $L > 0$ we have $Lip_b^L(S,c) = \{Lg : g \in Lip_b^1(S,c)\}$ and so (under appropriate assumptions) Theorem 17 implies the following limiting formulas:*

$$\lim_{L\to\infty} D_f^{Lip_b^L(S,c)}(Q\|P) = D_f(Q\|P)\,,$$
$$\lim_{L\searrow 0} L^{-1} D_f^{Lip_b^L(S,c)}(Q\|P) = W^{Lip_b^1(S,c)}(Q,P)\,.$$

Using Lemma 32 we can show that the boundedness constraint can be dropped in the formula for D_f^Γ when $\Gamma = \mathrm{Lip}_b^L(S,c)$; we exploit this fact in the numerical examples in Section 6 below.

Theorem 36 *Let $c : S \times S \to [0,\infty]$, $L \in (0,\infty)$, and define*

$$Lip^L(S,c) = \{g \in C(S) : |g(x) - g(y)| \le Lc(x,y) for\ all x,y \in S\}\,.$$

Let f be admissible such that f^* is bounded below. Then for $Q, P \in \mathcal{P}(S)$ we have

$$D_f^{Lip_b^L(S,c)}(Q\|P) = \sup_{g \in Lip^L(S,c)\cap L^1(Q)} \left\{E_Q[g] - \Lambda_f^P[g]\right\}\,. \tag{40}$$

Proof Lemma 34 shows that $\Gamma \equiv \mathrm{Lip}_b^L(S,c)$ satisfies the conditions of Theorem 15. Fix $g \in \mathrm{Lip}^L(S,c) \cap L^1(Q)$. For $n \in \mathbb{Z}^+$ define $g_n = n1_{g>n} + g1_{-n\le g\le n} - n1_{g<-n}$. It is easy to see that $g_n \in \mathrm{Lip}_b^L(S,c)$, $g_n \to g$, $|g_n| \le |g|$, $g_n \le g1_{g\ge 0}$. Therefore Lemma 32 implies

$$D_f^\Gamma(Q\|P) \ge E_Q[g] - \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}\,. \tag{41}$$

One inequality in Eq. (40) follows from taking the supremum over all $g \in \mathrm{Lip}^L(S,c) \cap L^1(Q)$ in Eq. (41) and the reverse follows from the fact that $\mathrm{Lip}_b^L(S,c) \subset \mathrm{Lip}^L(S,c) \cap L^1(Q)$. ■

6. (f,Γ)-GANs

Generative adversarial networks constitute a class of methods for 'learning' a probability distribution, Q, via a two-player game between a discriminator and a generator (both neural networks) (Goodfellow et al., 2014; Nowozin et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017; Pantazis et al., 2020). Mathematically, most GANs can be formulated as divergence minimization problems for a divergence, D, that has a variational characterization $D(Q\|P) = \sup_{g\in\Gamma} H[g;Q,P]$. The goal is then to solve the following optimization problem:

$$\inf_{\theta\in\Theta} D(Q\|P_\theta) = \inf_{\theta\in\Theta}\sup_{g\in\Gamma} H[g;Q,P_\theta]\,. \tag{42}$$

Here, g is called the discriminator and P_θ is the distribution of $h_\theta(X)$, where X is a random noise source and h_θ, $\theta \in \Theta$ is a neural network family (the generator). The min-max problem (42) can be interpreted as two-player zero-sum game. GANs based on the Wasserstein-metric have been very successful (Arjovsky et al., 2017; Gulrajani et al., 2017) and GANs based on the classical f-divergences have also been explored (Nowozin et al., 2016). Here we show that GANs based on the D_f^Γ divergences, which generalize and interpolate between the above two extremes, inherit desirable properties from both IPM-GANs (e.g., Wasserstein GAN) and f-GANs. Specifically, we focus on the following:

1. (f, Γ)-GANs can perform well when applied to heavy-tailed distributions. This property is inherited from the classical f-divergences.

2. (f, Γ)-GANs can perform well even when there is a lack of absolute continuity. This property is inherited from the Γ-IPMs.

We will specifically focus on the cases where $f = f_\alpha$, $\alpha > 1$, (see Equation 8) and $\Gamma = \mathrm{Lip}_b^L(\mathbb{R}^n)$ (see Lemma 34). We call the corresponding (f, Γ)-divergences the Lipschitz α-divergences and will denote them by D_α^L. As Γ is closed under shifts, we can express these divergences in one of two ways (see Equation 18):

$$D_\alpha^L(Q\|P) = \sup_{g \in \mathrm{Lip}_b^L(\mathbb{R}^n)} \{E_Q[g] - \Lambda_{f_\alpha}^P[g]\} \tag{43}$$

$$= \sup_{g \in \mathrm{Lip}_b^L(\mathbb{R}^n)} \{E_Q[g] - E_P[f_\alpha^*(g)]\}. \tag{44}$$

The formula for f_α^* can be found in Eq. (60) below.

Remark 37 *Formally taking the $\alpha \to \infty$ limit of* (44) *we arrive at what we call the Lipschitz ∞-divergence:*

$$D_\infty^L(Q\|P) = \sup_{g \in Lip_b^L(\mathbb{R}^n)} \{E_Q[g] - E_P[\max\{g, 0\}]\}.$$

It is straightforward to show that $D_\infty^L(Q\|P) = LW(Q, P)$, where W is classical Wasserstein metric

$$W(Q, P) = \sup_{g \in Lip_b^1(\mathbb{R}^n)} \{E_Q[g] - E_P[g]\},$$

though they are expressed in terms of different objective functionals (and hence their performance can differ in practice).

In numerical computations it can be inconvenient to restrict one's attention to bounded discriminators only. Fortunately, as shown in Theorem 36 above, the equality (15) remains true when Γ is expanded to include many unbounded g's. This fact justifies our use of unbounded discriminators (i.e., unbounded activation functions) in the following computations.

As our baseline method we take the two-sided gradient-penalized Wasserstein GAN (WGAN-GP) from Gulrajani et al. (2017),

WGAN-GP

$$\inf_{\theta} \sup_{g\in \mathrm{Lip}(\mathbb{R}^n)} \left\{ E_Q[g] - E_{P_\theta}[g] - \lambda \int (\|\nabla g\|/L - 1)^2 d\rho_\theta \right\}, \tag{45}$$

where $\lambda > 0$ is the strength of the penalty regularization. Here, and below, we have relaxed the Lipschitz constraint to a gradient penalty (two-sided for WGAN-GP and one-sided otherwise; see Section 4 for further discussion). We approximate the supremum over g by the supremum over a neural network family (the discriminator network). Again, the family of measures P_θ are the distributions of $X_\theta = h_\theta(X)$ where h_θ is the generator neural network, parameterized by $\theta \in \Theta$, and we let X be a Gaussian noise source. Finally, we let $\rho_\theta \sim TX_\theta + (1-T)Y$ where X_θ, $Y \sim Q$, $T \sim \mathrm{Unif}([0,1])$ are all independent (this choice of ρ_θ was used in Gulrajani et al., 2017).

We compare WGAN-GP to the Lipschitz α-GANs and Lipschitz KL-GAN, which are based on (44) and (43) respectively.

Lipschitz α-GAN

$$\inf_{\theta\in\Theta} \sup_{g\in \mathrm{Lip}(\mathbb{R}^n)} \left\{ E_Q[g] - E_{P_\theta}[f_\alpha^*(g)] - \lambda \int \max\{0, \|\nabla g\|^2/L^2 - 1\} d\rho_\theta \right\}. \tag{46}$$

When we want to make the values of α and/or L explicit we will refer to these as the D_α^L-GANs. By swapping Q and P_θ one obtains another family of GANs, which we call the reverse Lipschitz α-GANs (when clarity is needed, Equation 46 will be called a forward GAN). We note that forward and reverse GANs can have very different properties Goodfellow (2016).

In the case of the KL-divergence one can evaluate the optimization over ν in (43) (see Equation 19), leading to the following:

Lipschitz KL-GAN

$$\inf_{\theta\in\Theta} \sup_{g\in \mathrm{Lip}(\mathbb{R}^n)} \left\{ E_Q[g] - \log E_{P_\theta}[e^g] - \lambda \int \max\{0, \|\nabla g\|^2/L^2 - 1\} d\rho_\theta \right\}. \tag{47}$$

Remark 38 *For numerical purposes the GAN* (47)*, obtained using the representation* (11)*, performs significantly better than the GAN obtained from* (10)*. This is due to the numerical issues inherent in computing* $E_P[f_{KL}^*(g)] = E_P[\exp(g-1)]$*, as compared to computing the cumulant generating function* $\log E_P[\exp(g)]$*; see also Belghazi et al. (2018). We also refer to Birrell et al. (2020) for a more general perspective on finding tighter variational representations of divergences.*

6.1 Statistical Estimation of (f,Γ)-Divergences

In numerical computations, we approximate the (f,Γ)-divergence by replacing expectations under Q and P in (15) or (17) with their m-sample empirical means using i.i.d. samples from Q and P respectively, i.e., we estimate

$$E[D_f^\Gamma(Q_m\|P_m)] = E\left[\sup_{g\in\Gamma,\nu\in\mathbb{R}} \{E_{Q_m}[g-\nu] - E_{P_m}[f^*(g-\nu)]\} \right]. \tag{48}$$

Note that, at fixed g and ν, the objective functional on the right-hand-side of (48) is an unbiased estimator of the (f,Γ)-divergence objective functional. Including the optimization over g and ν we obtain a biased estimator which is an upper bound on D_f^Γ, as shown in the lemma below.

Lemma 39 *Let $f \in \mathcal{F}_1(a,b)$, $\Gamma \subset \mathcal{M}_b(\Omega)$ be nonempty, $Q, P \in \mathcal{P}(\Omega)$, and Q_m, P_m be empirical distributions constructed from m i.i.d. samples from Q and P respectively. Then*

$$E[D_f^\Gamma(Q_m\|P_m)] \geq D_f^\Gamma(Q\|P)\,.$$

Proof Using (17) we can compute

$$\begin{aligned} E[D_f^\Gamma(Q_m\|P_m)] =& E\left[\sup_{g\in\Gamma,\nu\in\mathbb{R}}\{E_{Q_m}[g-\nu]-E_{P_m}[f^*(g-\nu)]\}\right] \\ \geq& \sup_{g\in\Gamma,\nu\in\mathbb{R}} E\left[E_{Q_m}[g-\nu]-E_{P_m}[f^*(g-\nu)]\right] \\ =& \sup_{g\in\Gamma,\nu\in\mathbb{R}}\{E_Q[g-\nu]-E_P[f^*(g-\nu)]\} = D_f^\Gamma(Q\|P)\,. \end{aligned}$$

■

Remark 40 *As noted above, in the KL case one can evaluate the optimization over ν in* (48)*. This results in a biased objective functional due to the presence of the logarithm outside the expectation in* (19)*. This same issue was addressed earlier in Belghazi et al. (2018), e.g., by using sufficiently large minibatch sizes or an exponential moving average. This concern is not present in the objective functional for* (48) *or* (46)*.*

In the following GAN examples we work with Lipschitz functions and approximate the optimization over $\mathrm{Lip}(\mathbb{R}^n)$ by the optimization over some neural network family g_ϕ, $\phi\in\Phi$, and estimate the expectations using the m-sample empirical measures Q_m, $P_{m,\theta}$, $\rho_{m,\theta}$, e.g., we approximate the Lipschitz α-GAN (46) by

$$\inf_{\theta\in\Theta}\sup_{\phi\in\Phi}\left\{E_{Q_m}[g_\phi]-E_{P_{m,\theta}}[f_\alpha^*(g_\phi)]-\lambda\int\max\{0,\|\nabla g_\phi\|^2/L^2-1\}d\rho_{m,\theta}\right\}. \tag{49}$$

Various neural network architectures are known to be universal approximators (Hornik et al., 1989; Cybenko, 1989; Pinkus, 1999; Lu et al., 2017; Kidger and Lyons, 2020). Approximating the supremum over $g\in\mathrm{Lip}(\mathbb{R}^n)$ by the supremum over a finite-dimensional neural network family essentially results in a lower bound on the original, intended divergence. In the case of KL and Rényi divergences, such an approximation scheme is known to lead to consistent estimators as the sample size and network complexity grows (see Belghazi et al., 2018 and Birrell et al., 2021 respectively). Investigating the analogous consistency result for the (f,Γ)-divergence estimator is one avenue for future work.

6.2 (f,Γ)-GANs for Non-Absolutely-Continuous Heavy-Tailed Distributions

We mentioned above that the f-divergences are better suited to heavy-tailed distributions, as compared to the Wasserstein metric. Before demonstrating this in the context of GANs we provide a simple explicit example. Let $dQ = x^{-2}1_{x\geq 1}dx$ and $dP = (1+\delta)x^{-(2+\delta)}1_{x\geq 1}dx$ for $\delta > 0$, i.e., the tail of P decays faster than that of Q. For $\alpha > 1$ we can use Eq. (7) to compute

$$D_{f_\alpha}(Q\|P) = \frac{1}{\alpha(\alpha-1)(1+\delta)^{\alpha-1}}\int_1^\infty x^{\delta\alpha-(2+\delta)}dx - \frac{1}{\alpha(\alpha-1)}\,,$$

and so $D_{f_\alpha}(Q\|P) < \infty$ for all $\delta \in (0, 1/(\alpha-1))$. On the other hand, we can use the formula for the Wasserstein metric on $\mathcal{P}(\mathbb{R})$ from Vallender (1974) to compute

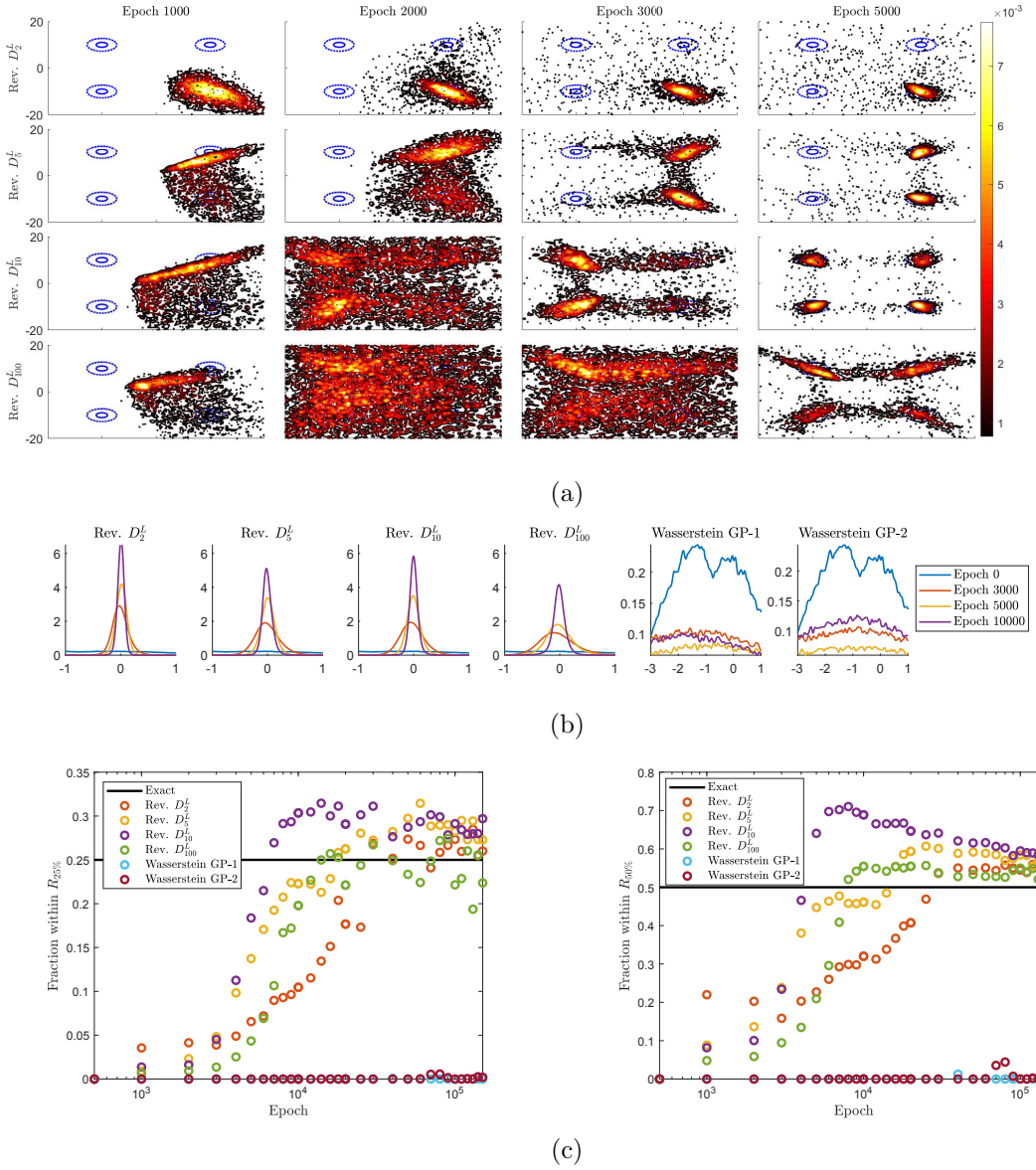
$$W(Q,P) = \int_{-\infty}^\infty |F_Q(t) - F_P(t)|dt = \int_1^\infty t^{-1} - t^{-(1+\delta)}dt = \infty \tag{50}$$

for all $\delta > 0$ (F_P and F_Q denote the cumulative distribution functions).

This calculation suggests that Lipschitz α-GANs may succeed for heavy-tailed distributions, even when WGAN-GP fails to converge. On the other hand the Wasserstein metric can be finite and informative even when Q and P are non-absolutely continuous, unlike the classical f-divergences (7). The (f,Γ)-divergences inherit both of these strengths from the Wasserstein and f-divergences (see Part 1 of Theorem 15), thus allowing for the training of GANs with heavy-tailed data and in the absence of absolute continuity. We demonstrate this via the following example, where both the WGAN-GP and classical f-GAN (i.e., without gradient penalty) fail to converge but the (f,Γ)-GANs succeed.

Here the data source, Q, is a mixture of four 2-dimensional t-distributions with 0.5 degrees of freedom, embedded in a plane in 12-dimensional space; note that this is a heavy-tailed distribution, as the mean does not exist; this suggests that WGAN will have difficulty learning this distribution. The generator uses a 10-dimensional noise source and so the generator and data source are generally not absolutely continuous with respect to one another (the former has support equal to the full 12-dimensional space while the latter is supported on a 2-dimensional plane). This suggests one cannot use the classical f-GAN (Nowozin et al., 2016), i.e., without gradient penalty (we confirmed that they perform very poorly on this problem). The (f,Γ)-GANs allow us to address both of the above difficulties; heavy tails can be accommodated by an appropriate choice of f and the lack of absolute continuity is addressed by using a 1-Lipschitz constraint (as in the Wasserstein metric). In this example we used gradient-penalty parameter values $\lambda = 10$, $L = 1$; Wasserstein GAN was run with both 1-sided and 2-sided gradient penalties (GP-1 and GP-2 respectively). In all cases the generator and discriminator have three fully connected hidden layers of 64, 32, and 16 nodes respectively, and with ReLU activation functions. The generator uses a 10-dimensional Gaussian noise source. Each SGD iteration was performed with a minibatch size of 1000 and 5 discriminator iterations were performed for every one generator iteration. Computations were done in TensorFlow and we used the RMSProp SGD algorithm with a learning rate of 2×10^{-4}.

In Figure 2 below we show generator samples for Wasserstein GAN, as in Eq. (45) and Gulrajani et al. (2017), and for various reverse Lipschitz α-GANs (46). Specifically, Panel



(a)

(b)

(c)

Figure 2: Generator samples and their statistical behavior from Wasserstein and reverse Lipschitz α-GAN methods. The data set consists of 5000 samples from a mixture of four 2-dimensional t-distributions with 0.5 degrees of freedom that are embedded in a plane in 12-dimensional space. Panel (a) shows the projection onto the 2-dimensional support plane (each column shows the result after a given number of training epochs); the solid and dashed blue ovals mark the 25% and 50% probability regions, respectively, of the data source while the heat-map shows the generator samples. Panel (b) shows the generator distribution, projected onto components orthogonal to the support plane. Values concentrated around zero indicate convergence to the sub-manifold. Panel (c) shows the fraction of generator samples, projected onto the 2D support plane of the measure, that are within the 25% and 50% probability regions.

(a) shows the projection onto the 2-dimensional support plane of Q (the heat-map shows samples from the generator and the data source, Q, is illustrated by the blue ovals) and Panel (b) shows the generator distribution, projected onto components orthogonal to the support plane. Panel (a) does not show WGAN-GP samples, as WGAN-GP failed to converge in this example; this is demonstrated in Panel (b) wherein we see that the Lipschitz α-GAN samples concentrate near the support plane (at 0) while the WGAN-GP samples spread out away from the support plane. The classical f-GAN without gradient penalty (Nowozin et al., 2016), which we don't show here, similarly failed to converge; this is unsurprising due to the lack of absolute continuity. Again, we can see that WGAN-GP fails to converge, while the Lipschitz α-GANs perform well. Some α's perform significantly better than others, making it an important hyperparameter to tune in this case. Results from a second set of runs, using a larger sample set, are shown in Figure 5 in Appendix E; the conclusions are similar. Forward Lipschitz GANs and forward Lipschitz KL-GANs all experienced blow-up and so they are not shown here. This behavior is reasonable when one considers the fact that Q is heavy tailed, while P_θ is not (it is generated by pushing forward Gaussian noise by Lipschitz functions), and so $D_f(Q\|P_\theta) = \infty$, while $D_f(P_\theta\|Q) < \infty$ (see Equation 7). As we have already demonstrated the inability of the Wasserstein metric to compare heavy-tailed distributions (see Equation 50), it is reasonable to conjecture that the finiteness of D_{f_α} is key in determining the success of the D_α^L-GAN. Interestingly, the Lipschitz constraint also appears to be key to the convergence of the method, something one would not anticipate solely based on finiteness of the corresponding divergences. We illustrate this with Figure 6 in Appendix E, where we apply the same method to the mixture of four 2-dimensional t-distributions, but without the high-dimensional embedding. In this case, the classical f-divergence is finite, however we find that the classical f-GAN fails to converge –WGAN also fails– but the (f, Γ)-GANs succeed. The theoretical understanding of this behavior is an interesting question, but we will not pursue it further here.

6.3 Strict Convexity and Enhanced Stability of (f, Γ)-GANs

Even in the absence of heavy tails, we find that the Lipschitz α-GANs can outperform WGAN-GP, as measured both by accuracy on quantities of interest as well as improved stability. The improved stability can be motivated by a simple (formal) calculation of the Hessian of the objective functional in Eq. (18),

$$H_f[g; Q, P] \equiv E_Q[g] - E_P[f^*(g)] \tag{51}$$

(see Appendix D for an analysis of the objective functional in the non shift-invariant case (15)). Let $g_0 \in \Gamma$ and perturb in some direction ψ, i.e., take a line segment $g_\epsilon = g_0 + \epsilon\psi \in \Gamma$. Then

$$\frac{d^2}{d\epsilon^2}|_{\epsilon=0} H_f[g_\epsilon; Q, P] = -E_P[(f^*)''(g_0)\psi^2]\,. \tag{52}$$

Convexity of f^* implies $(f^*)'' \geq 0$. If we have $(f^*)'' > 0$ then (52) implies the objective functional is strictly concave at g_0 in all directions, ψ, are nonzero on a set of positive P-probability. This strict concavity implies that the maximization problem (15) is a strictly convex optimization problem and suggests that numerical computation of $D_f^\Gamma(Q\|P)$ via

(15) may generally be more stable than computation of the Γ-IPM (16), as the latter uses a linear objective functional. Indeed, in Daskalakis et al. (2018) the authors demonstrated that gradient descent/ascent dynamics (used for training of GANs) oscillate without converging to the optimum for the Wasserstein-GAN loss function (16) in the special case where Γ consists of a parametric family of linear functions. In this case, more sophisticated algorithms such as training with optimism (Daskalakis et al., 2018) or two-step extra-gradient approaches (Mokhtari et al., 2020) were required to guarantee convergence. Here our (f, Γ) interpolation replaces the optimization of a linear objective functional in the case of the Γ-IPM (16) with the strictly concave problem (15). In the case of a linear discriminator space Γ, we obtain a complete theoretical justification based on the concavity calculation (52). In particular, we consider (51) where

$$H_f[g_\phi; Q, P] = E_Q[g_\phi] - E_P[f^*(g_\phi)]\,, \tag{53}$$

and where we assume that $\Gamma = \{g = g_\phi(x) : \phi = (\phi_1, \phi_2, ..., \phi_k) \in D\}$ is a parametric linear family (D is a closed, convex subset of $\mathbb{R}^k$), i.e., for any constants a_0, a_1 and any parameter values ϕ_0, ϕ_1 we have

$$g_{a_0\phi_0 + a_1\phi_1}(x) = a_0 g_{\phi_0}(x) + a_1 g_{\phi_1}(x)\,. \tag{54}$$

Using (54) and considering (52) for any $g = g_{\phi_0}$ and $\psi(x) = g_{\phi_1}$, $g_\epsilon = g_{\phi_0} + \epsilon g_{\phi_1}$, we readily have

$$\phi_1^\intercal \nabla_\phi^2 H_f[g_{\phi_0}; Q, P]\phi_1 = \frac{d^2}{d\epsilon^2}|_{\epsilon=0} H_f[g_\epsilon; Q, P] = -E_P[(f^*)''(g_{\phi_0}) g_{\phi_1}^2]\,, \tag{55}$$

provided all expected values are finite. As in (52), this analysis implies the strict concavity of (53) with respect to the linear parameterization ϕ. Thus our analysis covers linear spaces Γ such as linear combinations of splines or reproducing kernel Hilbert spaces (RKHS). However, when Γ is a family of neural networks then the g_ϕ's are not linear in ϕ and the above analysis does not apply. We will not pursue the theoretical analysis of this important case here but instead we will carry out an empirical study that explores the improved stability that (local) strict concavity would imply.

In Figure 3 we demonstrate both improved performance and improved stability of the Lipschitz α-GANs, as compared to WGAN-GP, on the CIFAR-10 data set (Krizhevsky, 2009), which consists of 32x32 RGB images from 10 classes. We use the same ResNet neural network architecture as in Gulrajani et al. (2017, Appendix F) and focus on evaluating the benefits of simply modifying the objective functional. We employ the adaptive learning rate Adam Optimizer method (Kingma and Ba, 2014) using the hyperparameter values shown in Algorithm 1 of Gulrajani et al. (2017) (note that in Gulrajani et al. (2017), α denotes the learning rate parameter and should not be confused with our use for α-divergences). We show the inception score as a function of the number of training epochs; the inception score (Salimans et al., 2016) is a commonly used performance measure for evaluating the diversity of images produced by a GAN. It uses a pre-trained classifier to estimate the number of distinct classes produced by the generator and so, when applied to CIFAR-10, values closer to 10 are considered better. In the legends we also show the final FID score achieved by each method. FID score is a performance measure that computes a distance between feature vectors of a classification model when applied to the original data, as compared to

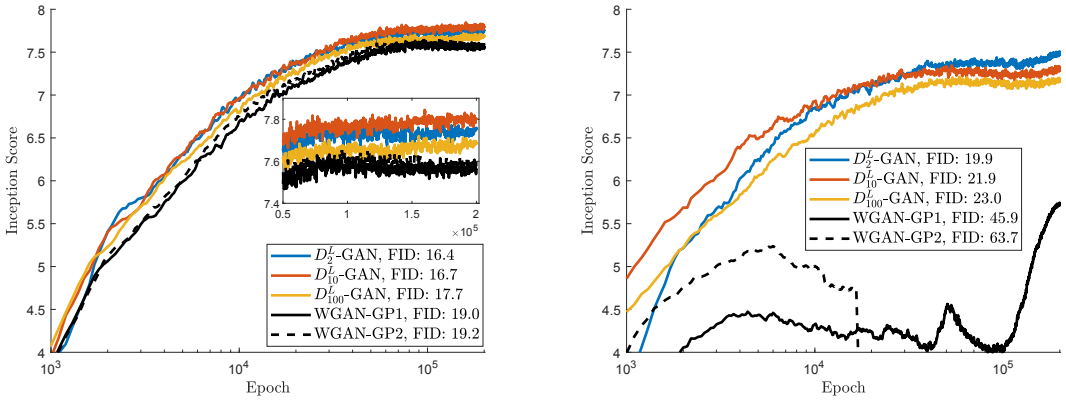


Figure 3: Comparison between Lipschitz α-GAN and WGAN-GP (both 1 and 2-sided) on the CIFAR-10 data set. Here we plot the inception score as a function of the number of training epochs (moving average over the last 5 data points, with results averaged over 5 runs). We also show the averaged final FID score in the legend, computed using 50000 samples from both Q and P. The neural network architecture is as in Appendix F of Gulrajani et al. (2017); in particular, it employs residual blocks. The left panel used an initial learning rate of 0.0002, the same as in Gulrajani et al. (2017), while in the right panel we used an initial learning rate of 0.001. Here, and in other similar tests, we find the Lipschitz α-GANs to be significantly more stable and require less tuning of the learning rate; in particular, none of the WGAN-GP2 runs shown in the right panel were able to complete successfully.

the generated samples (Heusel et al., 2017); a lower FID score is better. In the left panel of Figure 3 we show the results using an initial learning rate of 0.0002; we find a small improvement in inception score and substantial improvement in FID score when using the Lipschitz α-GANs, as compared to WGAN-GP (either 1 or 2-sided). In this example we find the performance to be relatively insensitive to the value of α.

In addition to the performance improvement, we find the Lipschitz α-GANs to be far less sensitive to the choice of learning rate. In the right panel of Figure 3 we show results using an initial learning rate of 0.001; here we observe significant degradation of the performance of WGAN-GP, but only a slight impact on the Lipschitz α-GANs. We conjecture that this increased stability is due to the strong concavity of the (f,Γ)-divergence objective functionals. Regarding increased stability, these numerical findings, the analysis for a general (non-parameterized) function space Γ in (52), as well as for the linear parametric case (55) provide only preliminary indications for the conjecture; a dedicated analysis for general parameterized Γ's that will include nonconvex parametric families such as neural networks is clearly necessary but we will not pursue it further here.

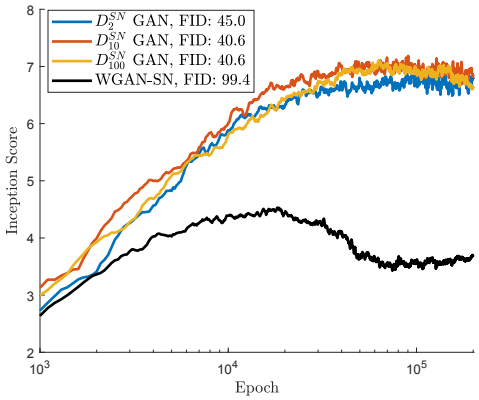


Figure 4: A comparison between Lipschitz α-GAN and WGAN, both using spectral normalization (SN) to enforce Lipschitz constraints. We used an initial learning rate of 0.0001 and otherwise employed same ResNet architecture and hyperparameters as in Figure 3. In particular, we did not attempt to further optimize the architecture when changing from a gradient penalty to SN. None of the methods performed as well as their gradient penalty counterparts from Figure 3, but note the especially poor performance of WGAN-SN. This suggests a further robustness of our methods to the use of sub-optimal architecture and hyperparameters.

6.3.1 Enhanced Stability and Spectral Normalization

In Miyato et al. (2018) the authors showed that spectral normalization, which directly controls the Lipschitz constant of each layer of a neural network by setting the largest singular value of its weight matrix to 1, provides enhanced stability as compared to WGAN-GP and at a lower computational cost (see Figures 1 and 2 in Miyato et al., 2018). Their method, which uses the Jensen-Shannon divergence, is equivalent to Eq. (18) (i.e., they do not include an optimization over shifts as in Equation 17) with a change of variables $g = \log(D)$ and using a function space Γ that consists of a neural network family with spectral normalization. In this example we use a spectral normalization function space in our method (17); this falls under the purview of Theorem 8 (see Table 2). We provide empirical evidence that the improved stability they observed is at least partially due to the strict concavity of the objective functional. Specifically, we find that WGAN with spectral normalization fails to inherit this improved stability and even fails to outperform WGAN-GP. Our results demonstrate that combining spectral normalization with other (strictly convex) objective functionals can enhance stability, similar both to what was observed in Miyato et al. (2018) and also to what we found in Figure 3. Here we again study the case $f = f_\alpha$, denoting these methods by D_α^{SN}; results are shown in Figure 4.

7. Conclusion

We have provided a systematic and rigorous exploration of the properties of the (f, Γ)-divergences, as defined in Eq. (1). This work was motivated by the need for a flexible collection

of novel divergences that combine key properties from f-divergences and Wasserstein metrics, such as the ability to work with heavy tails and with not-absolutely continuous distributions. A large list of proposed GANs fall under the presented mathematical framework (see Table 2), unifying to a considerable extent the loss formulation of GANs. We have illustrated the utility of the (f,Γ)-divergences in the training of GANs, showing both an increased domain of applicability and improved convergence stability. The theoretical results allow for a wide range of choices on f and Γ. We have shown that there are families of distributions that are better suited for (f,Γ)-divergence over either f-divergence or Γ-IPM. A more systematic exploration on the selection of proper f and Γ will add practical value from a practitioner's perspective, but further and more elaborate experimentation is required, along with a need for new theoretical insights. In the future we intend to further study the stability, the related statistical estimation theory, and explore these new divergences in additional challenging settings such as high-dimensional time-series generation, extreme events prediction, mutual information estimation, and uncertainty quantification for heavy-tailed distributions and in the absence of absolute continuity.

Acknowledgments

The authors are grateful to Dipjyoti Paul for providing code to build upon and for helping us to run simulations. The authors also want to acknowledge the anonymous referees for valuable comments, suggestions and insights. The research of J. B. was partially supported by NSF TRIPODS CISE-1934846 and by the Air Force Office of Scientific Research (AFOSR) under the grant FA-9550-21-1-0354. The research of M.K., and L. R.-B. was partially supported by by the National Science Foundation (NSF) under the grants TRIPODS CISE-1934846 and DMS-2008970, and by the Air Force Office of Scientific Research (AFOSR) under the grants FA-9550-18-1-0214 and FA-9550-21-1-0354. The research of P.D. was supported in part by the National Science Foundation (NSF) under the grant DMS-1904992 and by the Air Force Office of Scientific Research (AFOSR) under the grants FA-9550-18-1-0214 and FA-9550-21-1-0354. This work was performed in part using high performance computing equipment obtained under a grant from the Collaborative R&D Fund managed by the Massachusetts Technology Collaborative.

Appendix A. Properties of the Legendre Transform of $f \in \mathcal{F}_1(a,b)$

Here we collect a number of important properties regarding the LT of function in $\mathcal{F}_1(a,b)$. Recall that we use the same notation for the convex LSC extension $f:\mathbb{R}\to(-\infty,\infty]$ (see Definition 2). First we state an important continuity result (Rockafellar, 1970, Theorem 10.1).

Lemma 41 *Let $f \in \mathcal{F}_1(a,b)$. Then $f^*(y) = \sup_{x\in(a,b)}\{yx - f(x)\}$ and f^* is continuous on $\overline{\{f^* < \infty\}}$.*

Lemma 42 *Let $f \in \mathcal{F}_1(a,b)$. Then $f^*(y) \geq y$ for all $y \in \mathbb{R}$.*

Proof $f^*(y) = \sup_{x\in\mathbb{R}}\{yx - f(x)\} \geq 1\cdot y - f(1) = y.$ ■

Lemma 43 *If $f \in \mathcal{F}_1(a,b)$ is superlinear, i.e., $\lim_{x\to\pm\infty} f(x)/|x| = \infty$, then $\{f^* < \infty\} = \mathbb{R}$.*

Proof Suppose $y \in \mathbb{R}$ with $f^*(y) = \infty$, i.e., $\sup_{x\in(a,b)}\{yx - f(x)\} = \infty$. Then there exists $x_n \in (a,b)$ with $yx_n - f(x_n) \to \infty$. We can take a subsequence x_{n_j} with $x_{n_j} \to x \in [a,b]$. If x is finite then continuity of f on $\overline{(a,b)}$ allows us to compute $f(x) = -\infty$, a contradiction. If x is infinite then we write

$$|x_{n_j}|(\mathrm{sgn}(x_{n_j})y - f(x_{n_j})/|x_{n_j}|) \to \infty\,,$$

which contradicts $f(x_{n_j})/|x_{n_j}| \to \infty$. This completes the proof. ∎

Lemma 44 *Let $f \in \mathcal{F}_1(a,b)$ and suppose there exists $x_n \in \mathbb{R}$ with $x_n \to -\infty$ and $f^*(x_n)$ uniformly bounded above. Then f^* is nondecreasing.*

Proof Suppose not. Then there exists $y_1 < y_2$ with $f^*(y_1) > f^*(y_2)$ (in particular, $f^*(y_2) < \infty$). Take N such that for $n \geq N$ we have $x_n < y_1$. For $n \geq N$ let $t_n = (y_1 - x_n)/(y_2 - x_n)$. Then $t_n \in (0,1)$, $1 - t_n = (y_2 - y_1)/(y_2 - x_n) \to 0$ as $n \to \infty$ and $t_n y_2 + (1-t_n)x_n = y_1$. Hence

$$f^*(y_1) \leq t_n f^*(y_2) + (1-t_n)f^*(x_n)\,.$$

Note that this implies $f^*(y_1) < \infty$. We therefore have

$$f^*(x_n) \geq \frac{f^*(y_1) - f^*(y_2)}{1-t_n} + f^*(y_2) \to \infty$$

as $n \to \infty$. This is a contradiction and so we are done. ∎

Lemma 45 *Let $f \in \mathcal{F}_1(a,b)$. Then one of the following holds:*

1. *f^* is bounded below.*
2. *The set $I = \{y : f^*(y) < \infty\}$ is of the form $I = (-\infty, d)$ or $I = (-\infty, d]$ for some $d \in (-\infty, \infty]$ and f^* is nondecreasing.*

In addition, if f^* is not bounded below then there exists $b \leq 0$ such that $f^*|_{(-\infty,b]} \leq 0$ and $f^*|_{(b,\infty)} \geq 0$.

Proof Suppose f^* is not bounded below. Take $y_n \in I$ with $f^*(y_n) \to -\infty$. We know $f^*(y) \geq y$, hence $y_n \to -\infty$. Let $d = \sup I$. The set I is convex, therefore $(y_n, d) \subset I$ for all n. Hence $(-\infty, d) \subset I \subset (-\infty, d]$.

Now suppose we have $x_1 < x_2$ with $f^*(x_1) > f^*(x_2)$. With y_n as above, take n such that $y_n < x_1$ and $f^*(y_n) < f^*(x_2)$ and let $t = (x_2 - x_1)/(x_2 - y_n) \in (0,1)$. f^* is convex, hence

$$\begin{aligned} f^*(x_1) =& f^*(ty_n + (1-t)x_2) \leq t f^*(y_n) + (1-t)f^*(x_2) \\ \leq& t f^*(x_2) + (1-t)f^*(x_2) = f^*(x_2) < f^*(x_1)\,. \end{aligned}$$

This is a contradiction, therefore f^* is nondecreasing.

If f^* is not bounded below then let $b = \sup\{x : x \le 0, f^*(x) \le 0\}$. The properties $f^*|_{(-\infty,b]} \le 0$ and $f^*|_{(b,\infty)} \ge 0$ follow from the fact that $f^*(0) \ge 0$ and f^* is non-decreasing and is continuous on $\overline{I}$ (see Lemma 41). ■

Lemma 46 *Let $f \in \mathcal{F}_1(a,b)$ with $a \ge 0$. Then f^* is nondecreasing and $\{f^* < \infty\} = (-\infty, d)$ for $d \in (-\infty, \infty]$ or $\{f^* < \infty\} = (-\infty, d]$ for $d \in \mathbb{R}$.*

Proof Let $y_1 < y_2$. Then $xy_1 - f(x) \le xy_2 - f(x)$ for all $x \ge 0$. $f|_{(-\infty,0)} = \infty$, hence

$$f^*(y_1) = \sup_{x\in\mathbb{R}}\{y_1x - f(x)\} = \sup_{x\ge 0}\{y_1x - f(x)\} \le \sup_{x\ge 0}\{y_2x - f(x)\} = f^*(y_2)\,.$$

■

Next we give several results pertaining to the derivative of a convex function and its LT. A key tool will be the following decomposition of a convex function into an affine part and a remainder which can be found in Liese and Vajda (2006):

Lemma 47 *Let $f \in \mathcal{F}_1(a,b)$. For $x, y \in (a,b)$ we have*

$$f(y) = f(x) + f'_+(x)(y-x) + R_f(x,y)\,, \tag{56}$$

where $R_f \ge 0$, $R_f(x,x) = 0$, and if z is between x and y then $R_f(x,z) \le R_f(x,y)$.

Using this we can derive an explicit formula for f^* and prove several useful identities.

Lemma 48 *Let $f \in \mathcal{F}_1(a,b)$ and $y \in \{f^* < \infty\}^o$. Then*

$$f^*(y) = y(f^*)'_+(y) - f((f^*)'_+(y))\,.$$

Proof By assumption, $I \equiv \{f^* < \infty\}$ has nonempty interior and so

$$f(x) = \sup_{z\in I^o}\{zx - f^*(z)\}$$

for all x. Applying Lemma 47 to the convex function f^* on the interval I^o gives

$$f^*(z) = f^*(y) + (f^*)'_+(y)(z-y) + R_{f^*}(y,z)$$

for all $z \in I^o$. The assumption $y \in I$ implies $(f^*)'_+(y)$ exists and is finite. Hence

$$\begin{aligned} f((f^*)'_+(y)) &= \sup_{z\in I^o}\{z(f^*)'_+(y) - f^*(z)\} \\ &= \sup_{z\in I^o}\{z(f^*)'_+(y) - f^*(y) - (f^*)'_+(y)(z-y) - R_{f^*}(y,z)\} \\ &= (f^*)'_+(y)y - f^*(y) - \inf_{z\in I^o} R_{f^*}(y,z) = (f^*)'_+(y)y - f^*(y)\,. \end{aligned}$$

■

Lemma 49 *Let $f \in \mathcal{F}_1(a,b)$ and define $\nu_0 = f'_+(1)$. Then:*

1. *$f^*(\nu_0) = \nu_0$.*
2. *If $\nu_0 \in \{f^* < \infty\}^o$ and f is strictly convex on a neighborhood of 1 then $(f^*)'_+(\nu_0) = 1$.*

Proof

1. Using Lemma 47 we can compute

$$\begin{aligned} f^*(\nu_0) = &\sup_{x\in(a,b)} \{\nu_0 x - f(x)\} = \sup_{x\in(a,b)} \{\nu_0 x - (\nu_0(x-1) + R_f(1,x))\} \\ =&\nu_0 - \inf_{x\in(a,b)} R_f(1,x) = \nu_0\,. \end{aligned}$$

2. Using Lemma 48 along with Part 1 of this lemma we can write

$$f((f^*)'_+(\nu_0)) = \nu_0(f^*)'_+(\nu_0) - f^*(\nu_0) = \nu_0((f^*)'_+(\nu_0) - 1)\,. \tag{57}$$

In particular, we see that $(f^*)'_+(\nu_0) \in \{f < \infty\} \subset \overline{(a,b)}$. Using Lemma 47 we can write

$$f(x) = f'_+(1)(x-1) + R_f(1,x) \tag{58}$$

for $x \in (a,b)$. Therefore

$$f((f^*)'_+(\nu_0)) = \nu_0((f^*)'_+(\nu_0) - 1) + R_f(1,(f^*)'_+(\nu_0))$$

(this holds even if $(f^*)'_+(\nu_0)$ equals either a or b, as can be seen by taking limits in (58) and using the continuous extension of $R_f(1,\cdot)$). Combining this with Eq. (57) we find

$$R_f(1,(f^*)'_+(\nu_0)) = 0\,.$$

Using Lemma 47 (and taking limits if necessary) we obtain $0 \le R_f(1,z) \le R_f(1,(f^*)'_+(\nu_0))$ for all z between 1 and $(f^*)'_+(\nu_0)$, and hence $R_f(1,z) = 0$ for all such z. If $(f^*)'_+(\nu_0) \neq 1$ then this, combined with Eq. (56), implies that f is affine on the (non-trivial) line segment from 1 to $(f^*)'_+(\nu_0)$, which would contradict the assumption that f is strictly convex on a neighborhood of 1. Therefore we conclude that $(f^*)'_+(\nu_0) = 1$.

■

Finally, we will need formulas for f^*_{KL} and f^*_α from Eq. (8):

$$f^*_{KL}(y) = \exp(y-1), \tag{59}$$

$$f^*_\alpha(y) = \begin{cases} \alpha^{-1}(\alpha-1)^{\alpha/(\alpha-1)} y^{\alpha/(\alpha-1)} 1_{y>0} + \frac{1}{\alpha(\alpha-1)}, & \alpha > 1 \\ \infty 1_{y\ge 0} + \left(\alpha^{-1}(1-\alpha)^{-\alpha/(1-\alpha)} |y|^{-\alpha/(1-\alpha)} - \frac{1}{\alpha(1-\alpha)}\right) 1_{y<0}, & \alpha \in (0,1)\,. \end{cases} \tag{60}$$

Note that f_{KL} and f_α for $\alpha > 1$ are all strictly admissible but f_α is not admissible if $\alpha \in (0,1)$ (see Definition 14). Theorem 15 applies to f_{KL} and to f_α, $\alpha > 1$ while Theorem 8 applies to the case $\alpha \in (0,1)$.

Appendix B. Properties of the Classical f-Divergences

Here we collect a number of important properties of the classical f-divergences; see Definition 3. Perhaps most important is the following variational characterization. Versions of this were proven in Broniatowski and Keziou (2006) and Nguyen et al. (2010).

Proposition 50 *Let $f \in \mathcal{F}_1(a,b)$ and P, Q be probability measures on $(\Omega, \mathcal{M})$. Then*

$$\begin{aligned} D_f(Q\|P) &= \sup_{g\in\mathcal{M}_b(\Omega)} \{E_Q[g] - E_P[f^*(g)]\} \\ &= \sup_{g\in\mathcal{M}_b(\Omega)} \{E_Q[g] - \Lambda_f^P[g]\}. \end{aligned} \tag{61}$$

Proof Lemma 42 implies that $f^*(y) \geq y$ and so for $g \in \mathcal{M}_b(\Omega)$ we have $E_P[f^*(g)] \geq E_P[g] > -\infty$. Therefore the objective functional in Eq. (61) is well defined. Fix $y_0 \in \mathbb{R}$ with $f^*(y_0) \in \mathbb{R}$ and first consider the case $Q \not\ll P$. Then there exists a measurable set A with $P(A) = 0$ and $Q(A) > 0$. If we define $g = R1_A + y_0 1_{A^c}$ then g is bounded, measurable, and

$$E_Q[g] - E_P[f^*(g)] = RQ(A) + y_0 Q(A^c) - f^*(y_0)P(A^c)$$

for all R. Hence

$$\sup_{g\in\mathcal{M}_b(\Omega)} \{E_Q[g] - E_P[f^*(g)]\} \geq \lim_{R\to\infty} (RQ(A) + y_0 Q(A^c) - f^*(y_0)P(A^c)) = \infty\,.$$

Therefore $\sup_{g\in\mathcal{M}_b(\Omega)}\{E_Q[g] - E_P[f^*(g)]\} = \infty = D_f(Q\|P)$.

Now suppose $Q \ll P$: f is convex and LSC on $\mathbb{R}$, hence we can use convex duality to compute

$$\begin{aligned} E_Q[g] - E_P[f^*(g)] =& E_P[gdQ/dP - f^*(g)] \leq E_P[\sup_{y\in\mathbb{R}}\{ydQ/dP - f^*(y)\}] \\ =& E_P[(f^*)^*(dQ/dP)] = E_P[f(dQ/dP)] = D_f(Q\|P) \end{aligned}$$

for all $g \in \mathcal{M}_b(\Omega)$. Therefore it suffices to show $\sup_{g\in\mathcal{M}_b(\Omega)}\{E_Q[g] - E_P[f^*(g)]\} \geq D_f(Q\|P)$. Let $I \equiv \{f^* < \infty\}$. This is a nonempty interval in $\mathbb{R}$, hence we can find compact intervals $I_n \subset I_{n+1} \subset I$ with $\cup_n I_n = I$. f^* is convex and LSC on $\mathbb{R}$, hence it is continuous on $\overline{I}$. In particular, $y \to yx - f^*(y)$ is continuous on the compact set I_n. Therefore there exists measurable $g_n : \Omega \to I_n$ such that $|g_n dQ/dP - f^*(g_n) - \sup_{y\in I_n}\{ydQ/dP - f^*(y)\}| < 1/n$. The functions g_n are also bounded since $\text{Range}(g_n) \subset I_n$, a compact subset of $\mathbb{R}$, hence

$$\begin{aligned} &\sup_{g\in\mathcal{M}_b(\Omega)} \{E_Q[g] - E_P[f^*(g)]\} \geq E_Q[g_n] - E_P[f^*(g_n)] = E_P[dQ/dP g_n - f^*(g_n)] \\ \geq& E_P[\sup_{y\in I_n}\{ydQ/dP - f^*(y)\}] - 1/n \end{aligned}$$

for all n. Therefore

$$\sup_{g\in\mathcal{M}_b(\Omega)} \{E_Q[g] - E_P[f^*(g)]\} \geq \liminf_{n\to\infty} E_P[\sup_{y\in I_n}\{ydQ/dP - f^*(y)\}]\,.$$

Fix a large enough N such that $y_0 \in I_N$. Then for $n \geq N$ we have $\sup_{y\in I_n}\{ydQ/dP - f^*(y)\} \geq y_0 dQ/dP - f^*(y_0) \in L^1(P)$ (recall that $f^*(y_0)$ is finite). Therefore we can use Fatou's Lemma to compute

$$\begin{aligned}\liminf_{n\to\infty} E_P[\sup_{y\in I_n}\{ydQ/dP - f^*(y)\}] \geq & E_P[\liminf_{n\to\infty}\sup_{y\in I_n}\{ydQ/dP - f^*(y)\}] \\ = & E_P[\sup_{y\in I}\{ydQ/dP - f^*(y)\}] \\ = & E_P[(f^*)^*(dQ/dP)] = D_f(Q\|P)\,.\end{aligned}$$

This completes the proof of the first equality. To prove the second we compute

$$\begin{aligned}\sup_{g\in\mathcal{M}_b(\Omega)}\{E_Q[g] - \Lambda_f^P[g]\} &= \sup_{g\in\mathcal{M}_b(\Omega)}\{E_Q[g] - \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}\} \\ &= \sup_{g\in\mathcal{M}_b(\Omega),\nu\in\mathbb{R}}\{E_Q[g-\nu] - E_P[f^*(g-\nu)]\} \\ &= \sup_{g\in\mathcal{M}_b(\Omega)}\{E_Q[g] - E_P[f^*(g)]\}\,,\end{aligned}$$

where in the last line we used the fact that the map $\mathbb{R}\times\mathcal{M}_b(\Omega)\to\mathcal{M}_b(\Omega)$, $(\nu, g)\mapsto g-\nu$ is surjective. ■

On a metric space, and assuming f^* is everywhere finite, one can restrict the optimization in (61) to the set of bounded continuous functions.

Corollary 51 *Let $f\in\mathcal{F}_1(a,b)$, S be a metric space, and $Q, P\in\mathcal{P}(S)$. If $\{f^*<\infty\} = \mathbb{R}$ then*

$$D_f(Q\|P) = \sup_{g\in C_b(S)}\{E_Q[g] - E_P[f^*(g)]\}\,. \tag{62}$$

In particular, $(Q,P)\mapsto D_f(Q\|P)$ is lower semicontinuous.

Proof To prove Eq. (62) we start with Eq. (61) and use the extension of Lusin's theorem found in Appendix D of Dudley (2014) (which applies to an arbitrary metric space) to approximate bounded measurable functions with bounded continuous functions. The assumption $\{f^*<\infty\}$ implies $f^*(g)\in C_b(S)$ for all $g\in C_b(S)$ and so $(Q,P)\mapsto E_Q[g] - E_P[f^*(g)]$ is continuous. The supremum over g is therefore lower semicontinuous. ■

One can further restrict the optimization to Lipschitz functions.

Corollary 52 *Let $f\in\mathcal{F}_1(a,b)$, S be a metric space, $Q,P\in\mathcal{P}(S)$. If $\{f^*<\infty\} = \mathbb{R}$ then*

$$D_f(Q\|P) = \sup_{g\in Lip_b(S)}\{E_Q[g] - E_P[f^*(g)]\}\,, \tag{63}$$

where $Lip_b(S)$ denotes the set of real-valued bounded Lipschitz functions on S.

Proof The result follows from Corollary 51, together with the fact that every $g\in C_b(S)$ is the pointwise limit of Lipschitz functions, g_n, with $\|g_n\|_\infty \leq \|g\|_\infty$ (see Box 1.5 on page 6 of Santambrogio, 2015). ■

Remark 53 *Due to the invariance of the spaces $\mathcal{M}_b(S)$, $C_b(S)$, and $Lip_b(S)$ under shifting by a constant, one can replace $E_P[f^*(g)]$ in any of Eq. (61), Eq. (62), or Eq. (63) by $\inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}$ without changing the left-hand-side.*

Lemma 54 *Let $f \in \mathcal{F}_1(a,b)$, S be a Polish space, and $P \in \mathcal{P}(S)$. If $\{f^* < \infty\} = \mathbb{R}$ then the map $\mathcal{P}(S) \to [0,\infty]$, $Q \mapsto D_f(Q\|P)$ has compact sublevel sets.*

Proof Let $c \in \mathbb{R}$ and consider the sublevel set $L_c = \{Q : D_f(Q\|P) \leq c\}$. If $c < 0$ then $L_c = \emptyset$ and the claim is trivial, hence let $c \geq 0$. Corollary 51 implies that the map $\mathcal{P}(S) \times \mathcal{P}(S) \to [0,\infty]$, $(Q,P) \mapsto D_f(Q\|P)$ is LSC, therefore L_c is closed. By Prokhorov's theorem, if L_c is tight then L_c is precompact which will complete the proof: S is Polish, hence P is tight. Therefore for every $\delta > 0$ there exists a compact set K such that $P(K^c) \leq \delta$. Given $\epsilon > 0$ choose $d > 0$ large enough that $(c + f^*(0))/d \leq \epsilon/2$ and choose $\delta > 0$ such that $\frac{1}{d}(f^*(d) - f^*(0))\delta \leq \epsilon/2$. Then for any $Q \in L_c$, letting $g = d1_{K^c}$ (a bounded measurable function) and using the variational formula (61) we find

$$c \geq D_f(Q\|P) \geq E_Q[g] - E_P[f^*(g)] = dQ(K^c) - f^*(0) - (f^*(d) - f^*(0))P(K^c)\,.$$

Hence

$$Q(K^c) \leq (f^*(0) + c)/d + d^{-1}(f^*(d) - f^*(0))P(K^c) \leq \epsilon\,.$$

Therefore we conclude that L_c is tight. This completes the proof. ∎

In light of Corollary 51 and Lemma 54, it is useful to have simple conditions that ensure $\{f^* < \infty\} = \mathbb{R}$; see Lemma 43 above for such a result.

Next we show that $D_f(Q\|P)$ is strictly convex in Q when f is strictly convex.

Lemma 55 *Let $f \in \mathcal{F}_1(a,b)$ be strictly convex and $P \in \mathcal{P}(\Omega)$. Then $Q \mapsto D_f(Q\|P)$ is strictly convex on $\{Q : D_f(Q\|P) < \infty\}$.*

Proof First note that strict convexity of f on (a,b) implies strict convexity of the convex LSC extension (also denoted by f) on $\{f < \infty\}$. Fix distinct $Q_0, Q_1 \in \{D_f(Q\|P) < \infty\}$ and $t \in (0,1)$, and define $Q_t = tQ_1 + (1-t)Q_0$. Convexity of $Q \mapsto D_f(Q\|P)$ (which follows from Equation 61) implies $D_f(Q_t\|P) < \infty$.

Define $F = \{f(dQ_1/dP) < \infty \text{and} f(dQ_0/dP) < \infty\}$ and $G = \{dQ_1/dP \neq dQ_0/dP\}$. Finiteness of $D_f(Q_i\|P)$ implies $P(F) = 1$ and $Q_0 \neq Q_1$ implies $P(F \cap G) > 0$. We can write

$$\begin{aligned} &tD_f(Q_1\|P) + (1-t)D_f(Q_0\|P) - D_f(Q_t\|P) \\ =&E_P[1_F t f(dQ_1/dP) + 1_F(1-t)f(dQ_0/dP) - 1_F f(dQ_t/dP)]\,, \end{aligned}$$

where convexity of f implies the integrand is non-negative and strict convexity of f implies the integrand is positive on $F \cap G$. Therefore we can bound it below by integrating only over $F \cap G$, a set of positive measure. Hence the expectation is positive and we have proven the claim. ∎

The following lemma is the key step in the proof of the Gibbs variational principle for f-divergences in Proposition 57 below.

Lemma 56 *Let $f \in \mathcal{F}_1(a,b)$, P be a probability measure on $(\Omega, \mathcal{M})$, and $g \in \mathcal{M}_b(\Omega)$. Then*

$$\begin{aligned} E_P[f^*(g)] &= \sup_{h \in \mathcal{M}_b(\Omega): E_P[f(h)]<\infty} \{E_P[gh] - E_P[f(h)]\} \\ &= \sup_{h \in \mathcal{M}_{cr}(\Omega,(a,b))} \{E_P[gh] - E_P[f(h)]\}, \end{aligned} \tag{64}$$

where $\mathcal{M}_{cr}(\Omega,(a,b))$ denotes the set of measurable functions on Ω whose range is contained in a compact subset of (a,b).

Proof f is convex and $f(1) = 0$, hence $f(x) \geq f'_+(1)(x-1)$. This implies $E_P[f(h)]$ exists in $(-\infty, \infty]$. Therefore the right-hand-sides of (64) are well defined and the terms inside the suprema are finite for all h's satisfying the indicated conditions. The left-hand-side is well defined in $(-\infty, \infty]$ since $f^*(g) \geq g$ (see Lemma 42). For $h \in \mathcal{M}_b(\Omega)$ with $E_P[f(h)] < \infty$ we have

$$E_P[gh] - E_P[f(h)] = E_P[gh - f(h)] \leq E_P[\sup_{x \in \mathbb{R}}\{gx - f(x)\}] = E_P[f^*(g)]$$

and hence

$$\sup_{h \in \mathcal{M}_b(\Omega): E_P[f(h)]<\infty} \{E_P[gh] - E_P[f(h)]\} \leq E_P[f^*(g)]. \tag{65}$$

For the other direction, let $a < a_n < 1 < b_n < b$ with $a_n \searrow a$, $b_n \nearrow b$. Then

$$f^*(g) = \sup_n \sup_{x \in [a_n, b_n]} \{gx - f(x)\} = \lim_n \sup_{x \in [a_n, b_n]} \{gx - f(x)\}.$$

By letting $x = 1$ we see that $\sup_{x \in [a_n,b_n]}\{gx - f(x)\} \geq g \in L^1(P)$, and therefore Fatou's Lemma implies

$$E_P[f^*(g)] = E_P[\lim_n \sup_{x \in [a_n, b_n]} \{gx - f(x)\}] \leq \liminf_n E_P[\sup_{x \in [a_n, b_n]} \{gx - f(x)\}].$$

Using continuity of $gx - f(x)$ on $x \in (a,b)$ and finiteness of $\sup_{x \in [a_n,b_n]}\{gx - f(x)\}$ we see that there exists measurable $h_n : \Omega \to [a_n, b_n]$ such that

$$\sup_{x \in [a_n, b_n]} \{gx - f(x)\} \leq \frac{1}{n} + gh_n - f(h_n).$$

$h_n \in \mathcal{M}_{cr}(\Omega,(a,b))$, hence

$$E_P[\sup_{x \in [a_n, b_n]} \{gx - f(x)\}] \leq \frac{1}{n} + \sup_{h \in \mathcal{M}_{cr}(\Omega,(a,b))} \{E_P[gh] - E_P[f(h)]\}.$$

Therefore

$$E_P[f^*(g)] \leq \liminf_n E_P[\sup_{x \in [a_n, b_n]} \{gx - f(x)\}] \leq \sup_{h \in \mathcal{M}_{cr}(\Omega,(a,b))} \{E_P[gh] - E_P[f(h)]\}.$$

We have $\mathcal{M}_{cr}(\Omega,(a,b)) \subset \{h \in \mathcal{M}_b(\Omega) : E_P[f(h)] < \infty\}$ and so

$$E_P[f^*(g)] \leq \sup_{h\in\mathcal{M}_{cr}(\Omega,(a,b))} \{E_P[gh] - E_P[f(h)]\} \leq \sup_{h\in\mathcal{M}_b(\Omega):E_P[f(h)]<\infty} \{E_P[gh] - E_P[f(h)]\}\,.$$

When combined with Eq. (65), this completes the proof. ■

We can now prove the Gibbs variational formula for f-divergences in full generality; note that Corollary 58, which covers the case where $a \geq 0$, as proven in Ben-Tal and Teboulle (2007), but to the best of our knowledge the case (66), which covers $a < 0$, is new.

Proposition 57 *Let $f \in \mathcal{F}_1(a,b)$, $P \in \mathcal{P}(\Omega)$, and $g \in \mathcal{M}_b(\Omega)$. Then*

$$\sup_{h\in\mathcal{M}_b(\Omega):E_P[h]=1,E_P[f(h)]<\infty} \{E_P[gh] - E_P[f(h)]\} = \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g-\nu)]\}\,. \tag{66}$$

Corollary 58 *If $a \geq 0$ then (66) can be written as*

$$\sup_{Q\in\mathcal{P}(\Omega):D_f(Q\|P)<\infty} \{E_Q[g] - D_f(Q\|P)\} = \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g-\nu)]\}\,. \tag{67}$$

Remark 59 *Eq. (66) is an optimization over signed measures, $d\mu = hdP$, of net 'charge' 1.*

Remark 60 *The most commonly used case where $a < 0$ is the χ^2-divergence, which corresponds to the choice $f(x) = x^2 - 1$, $a = -\infty$, $b = \infty$.*

Proof Define the convex set $X = \{h \in \mathcal{M}_b(\Omega) : E_P[f(h)] < \infty\}$ (note that convexity of f implies $E_P[f(h)] > -\infty$ for all $h \in \mathcal{M}_b(\Omega)$), define the convex function $F : X \to \mathbb{R}$ by $F[h] = E_P[f(h)] - E_P[gh]$, and define the affine function $H : \mathcal{M}_b(\Omega) \to \mathbb{R}$ by $H[h] = 1 - E_P[h]$. These satisfy the Slater conditions (see Theorem 8.3.1 and Problem 8.7 in Luenberger, 1997) and so we have strong duality:

$$\begin{aligned} &\sup_{h\in\mathcal{M}_b(\Omega):E_P[h]=1,E_P[f(h)]<\infty} \{E_P[gh] - E_P[f(h)]\} \\ &= \inf_{\nu\in\mathbb{R}} \{\nu + \sup_{h\in\mathcal{M}_b(\Omega):E_P[f(h)]<\infty} \{E_P[(g-\nu)h] - E_P[f(h)]\} \\ &= \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g-\nu)]\}\,, \end{aligned} \tag{68}$$

where we used Lemma 56 to obtain the last line. If $a \geq 0$ then $E_P[f(h)] = \infty$ unless $h \geq 0$ P-a.s. and so the supremum in (68) can be restricted to non-negative h. Defining $dQ = hdP$ we can then rewrite (68) as the supremum over $Q \in \mathcal{P}(\Omega)$ with $D_f(Q\|P) < \infty$. ■

B.1 Variational Characterization over Unbounded g's

In many cases it is useful to extend the variational formula (61) to unbounded g's. In this section we give several such results. First we prove a pair of lemmas that ensure certain expectations are finite. The first of these can also be found in Lemma 2 of Birrell et al. (2020).

Lemma 61 *Let $f \in \mathcal{F}_1(a,b)$ and $Q, P \in \mathcal{P}(\Omega)$ with $D_f(Q\|P) < \infty$. If $g \in L^1(Q)$ then $E_P[f^*(g)^-] < \infty$.*

Remark 62 *Recall that we use $g^\pm$ to denote the positive and negative parts of a function g (so that $g^\pm \geq 0$ and $g = g^+ - g^-$).*

Proof Fix $g \in L^1(Q)$. The result it trivial if f^* is bounded below, so suppose not. Lemma 45 therefore implies that $I = \{y : f^*(y) < \infty\}$ is of the form $I = (-\infty, d)$ or $I = (-\infty, d]$ for some $d \in (-\infty, \infty]$, f^* is nondecreasing, and there exists $b \in \mathbb{R}$ such that $f^* \leq 0$ on $(-\infty, b]$ and $f^* \geq 0$ on (b, ∞). Define $g_b = g1_{g \leq b} + b1_{g > b}$, so that $g_b \leq b$ and $g_b \in L^1(Q)$. We have

$$\begin{aligned} &E_P[f^*(g)^-] = E_P[1_{g\leq b} f^*(g)^-] = E_P[1_{g\leq b} f^*(g_b)^-] \\ \leq &E_P[f^*(g_b)^-] = E_P[-f^*(g_b)]\,, \end{aligned}$$

hence

$$E_P[f^*(g_b)] \leq -E_P[f^*(g)^-]\,.$$

Let $g_{b,n} = -n1_{g_b < -n} + g_b 1_{g_b \geq -n}$. The $g_{b,n}$ are bounded, therefore we can use (61) to obtain

$$E_Q[g_{b,n}] - E_P[f^*(g_{b,n})] \leq D_f(Q\|P)\,,$$

where $E_P[f^*(g_{b,n})] \in (-\infty, \infty]$. This implies

$$E_Q[g_{b,n}] \leq D_f(Q\|P) + E_P[f^*(g_{b,n})]\,.$$

$g_{b,n} \to g_b$ pointwise, $|g_{b,n}| \leq |g_b|$, and $g_b \in L^1(Q)$, so we use the dominated convergence theorem to compute

$$E_Q[g_b] \leq \liminf_n (D_f(Q\|P) + E_P[f^*(g_{b,n})]) = D_f(Q\|P) + \liminf_n E_P[f^*(g_{b,n})]\,. \tag{69}$$

(The assumption that $D_f(Q\|P) < \infty$ implies the right-hand-side of (69) is well-defined). We have $g_{b,n+1} \leq g_{b,n}$, hence $f^*(g_{b,n+1}) \leq f^*(g_{b,n})$. The function f^* is continuous on $(-\infty, b]$ and for N large enough we have $g_{b,n} \leq b$ for all $n \geq N$, hence

$$0 \leq -f^*(g_{b,n}) \nearrow -f^*(g_b)$$

as $n \to \infty$. Therefore the monotone convergence theorem gives $\lim_n E_P[f^*(g_{b,n})] = E_P[f^*(g_b)]$ and we find

$$-\infty < E_Q[g_b] \leq D_f(Q\|P) + E_P[f^*(g_b)] \leq D_f(Q\|P) - E_P[f^*(g)^-]\,.$$

This proves $E_P[f^*(g)^-] < \infty$, as claimed. ∎

Lemma 63 *Let $f \in \mathcal{F}_1(a,b)$, $P \in \mathcal{P}(\Omega)$, and $g \in \mathcal{M}(\Omega)$. Suppose $E_P[f^*(cg - \nu)^+] < \infty$ for some $\nu \in \mathbb{R}$ and $c > 0$. Then for all $Q \in \mathcal{P}(\Omega)$ with $D_f(Q\|P) < \infty$ we have $E_Q[g^+] < \infty$.*

Proof Fix d for which $f^*(d)$ is finite and define $g_n = g1_{g\in[0,n)} + (d+\nu)/c1_{g\notin[0,n)} \in \mathcal{M}_b(\Omega)$. Hence $cg_n - \nu \in L^1(Q)$ and the variational formula (61) gives

$$D_f(Q\|P) \geq E_Q[cg_n - \nu] - E_P[f^*(cg_n - \nu)]\,,$$

where $E_P[f^*(cg_n - \nu)]$ is defined in $(-\infty, \infty]$. Hence

$$E_Q[cg_n] - D_f(Q\|P) \leq \nu + E_P[f^*(cg_n - \nu)]\,.$$

We can bound

$$f^*(cg_n - \nu) = f^*(cg - \nu)1_{g\in[0,n)} + f^*(d)1_{g\notin[0,n)} \leq f^*(cg - \nu)^+ + |f^*(d)|\,,$$

and so

$$E_P[f^*(cg_n - \nu)] \leq E_P[f^*(cg - \nu)^+] + |f^*(d)|\,.$$

Therefore

$$E_Q[cg_n] - D_f(Q\|P) \leq \nu + E_P[f^*(cg - \nu)^+] + |f^*(d)| < \infty$$

for all n. Taking $n \to \infty$ we obtain

$$\liminf_n E_Q[g_n] \leq c^{-1}(\nu + E_P[f^*(cg - \nu)^+] + |f^*(d)| + D_f(Q\|P)) < \infty\,.$$

The functions g_n are uniformly bounded below, therefore Fatou's Lemma implies

$$E_Q[\liminf_n g_n] \leq \liminf_n E_Q[g_n] \leq c^{-1}(\nu + E_P[f^*(cg - \nu)^+] + |f^*(d)| + D_f(Q\|P)) < \infty\,.$$

We have $g_n \to g1_{g\geq 0} + c^{-1}(d+\nu)1_{g<0}$ pointwise, hence

$$E_Q[g^+] + c^{-1}(d+\nu)Q(g<0) = E_Q[g^+ + c^{-1}(d+\nu)1_{g<0}] < \infty\,.$$

This implies $E_Q[g^+] < \infty$, as claimed. ■

We can now prove variational characterizations of D_f that allow for g to be unbounded. The following is found in Theorem 2 of Birrell et al. (2020).

Proposition 64 *Let $f \in \mathcal{F}_1(a,b)$ and P, Q be probability measures on $(\Omega, \mathcal{M})$. If f^* is bounded below or $D_f(Q\|P) < \infty$ then*

$$D_f(Q\|P) = \sup_{g\in L^1(Q)} \{E_Q[g] - E_P[f^*(g)]\}\,, \tag{70}$$

where $E_P[f^*(g)]$ exists in $(-\infty, \infty]$.

Proof If f^* is bounded below then $E_P[f^*(g)^-] < \infty$ for all $g \in L^1(Q)$. If $D_f(Q\|P) < \infty$ then Lemma 61 also implies $E_P[f^*(g)^-] < \infty$. So in either case we find that $E_P[f^*(g)]$ is defined in $(-\infty, \infty]$. In particular, the objective functional in (70) is well defined. Due to the variational

characterization (61), to prove (70) it suffices to prove $E_Q[g] - E_P[f^*(g)] \leq D_f(Q\|P)$ for all $g \in L^1(Q)$.

Fix $g \in L^1(Q)$. If $D_f(Q\|P) = \infty$ or $E_P[f^*(g)] = \infty$ then the required bound is trivial, so suppose not. In this case we have $f^*(g) < \infty$ P-a.s., i.e., g maps into $I \equiv \{f^* < \infty\}$ P-a.s. We are in the case where $D_f(Q\|P) < \infty$ so $Q \ll P$, hence g maps into I Q-a.s. as well. Therefore, by redefining g on a measure zero set (under both Q and P) we can assume that $\text{Range}(g) \subset I$. In summary, we have now reduced the problem to showing that $E_Q[g] - E_P[f^*(g)] \leq D_f(Q\|P)$ in the case where $g \in L^1(Q)$, $f^*(g) \in L^1(P)$, $D_f(Q\|P) < \infty$, $\text{Range}(g) \subset I$: Fix $y_0 \in I$ and define $g_n = y_0 1_{g<-n} + g1_{-n\leq g\leq n} + y_0 1_{g>n} \in \mathcal{M}_b(\Omega)$. Eq. (61) implies

$$D_f(Q\|P) \geq E_Q[g_n] - E_P[f^*(g_n)]\,.$$

$g_n \to g$ pointwise, and $|g_n| \leq |g| + |y_0| \in L^1(Q)$, hence the dominated convergence theorem gives $E_Q[g_n] \to E_Q[g]$. $\text{Range}(g_n), \text{Range}(g) \subset I$ and f^* is continuous on I, hence $f^*(g_n) \to f^*(g)$ pointwise. We have

$$\begin{aligned} |f^*(g_n)| =& |f^*(g_n)|1_{g<-n} + |f^*(g_n)|1_{-n\leq g\leq n} + |f^*(g_n)|1_{g>n} \\ \leq& |f^*(y_0)| + |f^*(g)| \in L^1(P)\,. \end{aligned}$$

Therefore the dominated convergence theorem implies $E_P[f^*(g_n)] \to E_P[f^*(g)]$. Hence

$$D_f(Q\|P) \geq \lim_{n\to\infty} (E_Q[g_n] - E_P[f^*(g_n)]) = E_Q[g] - E_P[f^*(g)]\,.$$

This completes the proof. ∎

In some cases it is convenient to define conventions regarding infinities that result in a variational formula involving the supremum over all measurable g:

Theorem 65 *Let $f \in \mathcal{F}_1(a,b)$, $Q, P \in \mathcal{P}(\Omega)$, and $\mathcal{M}_b(\Omega) \subset \Gamma \subset \mathcal{M}(\Omega)$. Then*

$$D_f(Q\|P) = \sup_{g\in\Gamma}\{E_Q[g] - E_P[f^*(g)]\}\,, \tag{71}$$

where we define $\infty - \infty \equiv -\infty$, $-\infty + \infty \equiv -\infty$.

Remark 66 *Our convention regarding infinities ensures that $\int g d\eta \equiv \int g^+ d\eta - \int g^- d\eta$ is defined in $\overline{\mathbb{R}}$ for all $\eta \in \mathcal{P}(\Omega)$, $g \in \mathcal{M}(\Omega)$.*

Remark 67 *The above conventions regarding infinities can be viewed as a convenient but rigorous shorthand for restricting the optimization in (71) to those $g \in \Gamma$ for which such infinitities do not occur. However, there is more content to Theorem 65 than this simple convention. For instance, the equality Eq. (71) implies that if $D_f(Q\|P) < \infty$ and $g \in \Gamma$ with $E_Q[g] = \infty$ then it must also be the case that $E_P[f^*(g)] = \infty$.*

Proof From (61) we have

$$\begin{aligned} D_f(Q\|P) &= \sup_{g\in\mathcal{M}_b(\Omega)}\{E_Q[g]-E_P[f^*(g)]\} \\ &\le \sup_{g\in\Gamma}\{E_Q[g]-E_P[f^*(g)]\} \\ &\le \sup_{g\in\mathcal{M}(\Omega)}\{E_Q[g]-E_P[f^*(g)]\}\,. \end{aligned} \tag{72}$$

If $D_f(Q\|P)=\infty$ then the above inequalities are all equalities and we are done. In the case where $D_f(Q\|P)<\infty$, Proposition 64 implies

$$D_f(Q\|P)\ge E_Q[g]-E_P[f^*(g)] \tag{73}$$

for all $g\in L^1(Q)$. In light of Eq. (72), if we can show (73) holds for all $g\in\mathcal{M}(\Omega)$ then we are done. If $g^-\notin L^1(Q)$ then (73) is a trivial consequence of our conventions regarding infinities. This leaves only the case where $g\in\mathcal{M}(\Omega)$ with $g^+\notin L^1(Q)$ and $g^-\in L^1(Q)$.

First we show that $E_P[f^*(g)^-]<\infty$ in this case: If f^* is bounded below this is trivial so suppose not. Therefore Lemma 45 implies f^* is nondecreasing and there exists $b\le 0$ such that $f^*|_{(-\infty,b]}\le 0$ and $f^*|_{(b,\infty)}\ge 0$. Define $g_n=g1_{g\le n}+b1_{g>n}$ for $n\in\mathbb{Z}^+$. $g_n\in L^1(Q)$ and $f^*(g)^-1_{g\ge n}=0$, hence $E_P[f^*(g_n)]\in(-\infty,\infty]$ and

$$\infty > E_P[f^*(g_n)^-]=E_P[f^*(g)^-1_{g\le n}+f^*(b)^-1_{g>n}]=E_P[f^*(g)^-]+f^*(b)^-E_P[1_{g>n}]\,.$$

Therefore $E_P[f^*(g)^-]<\infty$ as claimed. Lemma 63 together with $D_f(Q\|P)<\infty$ and $g^+\notin L^1(Q)$ implies $E_P[f^*(g)^+]=\infty$. Therefore we conclude that $E_Q[g]-E_P[f^*(g)]=\infty-\infty=-\infty<D_f(Q\|P)$. This completes the proof. ■

The following alternative form is also useful:

Corollary 68 *Let $f\in\mathcal{F}_1(a,b)$ and $Q,P\in\mathcal{P}(\Omega)$. Then*

$$D_f(Q\|P)=\sup_{g\in\mathcal{M}(\Omega)}\{E_Q[g]-\inf_{\nu\in\mathbb{R}}\{\nu+E_P[f^*(g-\nu)]\}\}\,,$$

where we define $\infty-\infty\equiv-\infty$, $-\infty+\infty\equiv-\infty$.

Proof The bound

$$\sup_{g\in\mathcal{M}(\Omega)}\{E_Q[g]-\inf_{\nu\in\mathbb{R}}\{\nu+E_P[f^*(g-\nu)]\}\}\ge D_f(Q\|P)$$

is an obvious consequence of Theorem 65. To conclude the reverse inequality, we need to show that

$$E_Q[g]-\inf_{\nu\in\mathbb{R}}\{\nu+E_P[f^*(g-\nu)]\}\le D_f(Q\|P) \tag{74}$$

for all $g\in\mathcal{M}(\Omega)$. If $E_Q[g]=\infty$ and $\inf_{\nu\in\mathbb{R}}\{\nu+E_P[f^*(g-\nu)]\}\ne\infty$ then $E_Q[g^-]<\infty$ and there exists ν_0 with $E_P[f^*(g-\nu_0)]<\infty$. Hence Theorem 65 implies

$$D_f(Q\|P)\ge E_Q[g-\nu_0]-E_P[f^*(g-\nu_0)]=\infty\,.$$

Therefore the claim holds in this case. Otherwise, we are in the case where $\inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\} = \infty$ or $E_Q[g] = -\infty$ or $E_Q[g] \in \mathbb{R}$. The first two of these immediately imply (74), due to our conventions regarding infinities. Finally, if $E_Q[g] \in \mathbb{R}$ then

$$\begin{aligned} E_Q[g] - \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\} &= \sup_{\nu\in\mathbb{R}}\{E_Q[g-\nu] - E_P[f^*(g-\nu)]\} \\ &\leq \sup_{g\in\mathcal{M}(\Omega)}\{E_Q[g] - E_P[f^*(g)]\} = D_f(Q\|P)\,. \end{aligned}$$

This completes the proof. ■

Appendix C. Proofs of Properties of (f,Γ)-Divergences

In this appendix we prove the (f,Γ)-divergence properties from Section 2; some results will be proven in greater generality than were stated earlier. Our method of proof will require us to work with finite signed measures (at least during the intermediate steps), and not just with probability measures. For that reason we provide a more general definition of the (f,Γ)-divergences here:

Definition 69 *Let $f \in \mathcal{F}_1(a,b)$ and $\Gamma \subset \mathcal{M}_b(\Omega)$ be nonempty. For $P \in \mathcal{P}(\Omega)$ and $\mu \in M(\Omega)$ we define the (f,Γ)-divergence by*

$$D_f^\Gamma(\mu\|P) = \sup_{g\in\Gamma}\left\{\int g d\mu - \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}\right\} \tag{75}$$

and for $\mu,\kappa \in M(\Omega)$ we define the Γ-IPM by

$$W^\Gamma(\mu,\kappa) = \sup_{g\in\Gamma}\left\{\int g d\mu - \int g d\kappa\right\}. \tag{76}$$

We start with the proof of the dual variational formula from Theorem 6, which we restate below. In addition, we treat the case $a < 0$, which requires the more general definition (75).

Theorem 70 *Let $f \in \mathcal{F}_1(a,b)$, $P \in \mathcal{P}(\Omega)$, and $\Gamma \subset \mathcal{M}_b(\Omega)$ be nonempty. For $g \in \Gamma$ we have*

$$(D_f^\Gamma)^*(g;P) \equiv \sup_{\mu\in M_1(\Omega)}\left\{\int g d\mu - D_f^\Gamma(\mu\|P)\right\} = \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}\,, \tag{77}$$

where $M_1(\Omega) \equiv \{\mu \in M(\Omega) : \mu(\Omega) = 1\}$. If $a \geq 0$ then

$$(D_f^\Gamma)^*(g;P) \equiv \sup_{Q\in\mathcal{P}(\Omega)}\{E_Q[g] - D_f^\Gamma(Q\|P)\} = \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}\,. \tag{78}$$

Proof Let $g \in \Gamma \subset \mathcal{M}_b(\Omega)$. Using the definition of D_f^Γ along with Proposition 57 we have

$$\begin{aligned}
&\sup_{\mu\in M_1(\Omega)} \left\{ \int g d\mu - D_f^\Gamma(\mu\|P) \right\} \\
&= \sup_{\mu\in M_1(\Omega)} \left\{ \int g d\mu - \sup_{\widetilde{g}\in\Gamma} \left\{ \int \widetilde{g} d\mu - \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(\widetilde{g}-\nu)]\} \right\} \right\} \\
&\leq \sup_{\mu\in M_1(\Omega)} \left\{ \int g d\mu - \left(\int g d\mu - \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g-\nu)]\} \right) \right\} \\
&= \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g-\nu)]\} \\
&= \sup_{h\in\mathcal{M}_b(\Omega): E_P[h]=1, E_P[f(h)]<\infty} \left\{ \int g dP_h - E_P[f(h)] \right\},
\end{aligned}$$

where $dP_h = h dP$. Noting that $P_h \in M_1(\Omega)$ and using $(f^*)^* = f$ we obtain the bound

$$D_f^\Gamma(P_h\|P) = \sup_{g\in\Gamma} \sup_{\nu\in\mathbb{R}} E_P[(g-\nu)h - f^*(g-\nu)] \leq E_P[f(h)]\,.$$

Hence

$$\begin{aligned}
&\sup_{h\in\mathcal{M}_b(\Omega): E_P[h]=1, E_P[f(h)]<\infty} \left\{ \int g dP_h - E_P[f(h)] \right\} \\
&\leq \sup_{h\in\mathcal{M}_b(\Omega): E_P[h]=1, E_P[f(h)]<\infty} \left\{ \int g dP_h - D_f^\Gamma(P_h\|P) \right\} \\
&\leq \sup_{\mu\in M_1(\Omega)} \left\{ \int g d\mu - D_f^\Gamma(\mu\|P) \right\}.
\end{aligned}$$

Combining these we arrive at Eq. (77). If $a \geq 0$ then we can use Eq. (77), the bound $D_f^\Gamma(Q\|P) \leq D_f(Q\|P)$, and then Eq. (67) to obtain

$$\begin{aligned}
\inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g-\nu)]\} &\geq \sup_{Q\in\mathcal{P}(\Omega)} \{E_Q[g] - D_f^\Gamma(Q\|P)\} \\
&\geq \sup_{Q\in\mathcal{P}(\Omega)} \{E_Q[g] - D_f(Q\|P)\} \\
&= \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g-\nu)]\}\,,
\end{aligned}$$

which proves (78). ■

Next we prove that the (f,Γ)-divergences are bounded above by the classical f-divergence and Γ-IPM and also derive the convexity and divergence properties from Theorem 8.

Theorem 71 *Let $f \in \mathcal{F}_1(a,b)$, $\Gamma \subset \mathcal{M}_b(\Omega)$ be nonempty, and $Q, P \in \mathcal{P}(\Omega)$.*

1.

$$D_f^\Gamma(Q\|P) \leq \inf_{\eta\in\mathcal{P}(\Omega)} \{D_f(\eta\|P) + W^\Gamma(Q,\eta)\}\,. \tag{79}$$

In particular, $D_f^\Gamma(Q\|P) \leq \min\{D_f(Q\|P), W^\Gamma(Q,P)\}$.

2. *The map* $(Q,P) \in \mathcal{P}(S)\times\mathcal{P}(S) \mapsto D_f^\Gamma(Q\|P)$ *is convex.*

3. *If there exists* $c_0 \in \Gamma \cap \mathbb{R}$ *then* $D_f^\Gamma(Q\|P) \geq 0$.

4. *Suppose* f *and* Γ *satisfy the following:*

   (a) *There exist a nonempty set* $\Psi \subset \Gamma$ *with the following properties:*

      i. Ψ *is* $\mathcal{P}(\Omega)$*-determining.*

      ii. *For all* $\psi \in \Psi$ *there exists* $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ *such that* $c_0 + \epsilon\psi \in \Gamma$ *for all* $|\epsilon| < \epsilon_0$.

   (b) f *is strictly convex on a neighborhood of* 1.

   (c) f^* *is finite and* C^1 *on a neighborhood of* $\nu_0 \equiv f'_+(1)$.

   Then:

   (i) D_f^Γ *has the divergence property.*

   (ii) W^Γ *has the divergence property.*

Proof

1. Let f_0 be the restriction of f to the interval $(\max\{a,0\}, b)$, so that $f_0 \in \mathcal{F}_1(\max\{a,0\},b)$. Note that f and f_0 agree on $[0,\infty)$ and so $D_f = D_{f_0}$ (see Equation 7). The definition of the Legendre transform implies $f_0^* \leq f^*$ and so $\Lambda_{f_0}^P \leq \Lambda_f^P$ (see Equation 9). We can now compute

$$\begin{aligned}
\inf_{\eta\in\mathcal{P}(\Omega)}\{D_f(\eta\|P) + W^\Gamma(Q,\eta)\} &= \inf_{\eta\in\mathcal{P}(\Omega)}\{D_{f_0}(\eta\|P) + W^\Gamma(Q,\eta)\} \\
&= \inf_{\eta\in\mathcal{P}(\Omega)}\{\sup_{g\in\Gamma}\{D_{f_0}(\eta\|P) + E_Q[g] - E_\eta[g]\}\} \\
&\geq \sup_{g\in\Gamma}\{\inf_{\eta\in\mathcal{P}(\Omega)}\{D_{f_0}(\eta\|P) + E_Q[g] - E_\eta[g]\}\} \\
&= \sup_{g\in\Gamma}\{E_Q[g] - \Lambda_{f_0}^P[g]\} \\
&\geq \sup_{g\in\Gamma}\{E_Q[g] - \Lambda_f^P[g]\} = D_f^\Gamma(Q\|P)\,,
\end{aligned}$$

   where we used Eq. (12) to obtain the second-to-last line.

   Remark 72 *We emphasize that* Λ_f^P *naturally appears when working with the infimal convolution of an* f*-divergence and a* Γ*-IPM, due to the identity* (12).

2. Convexity of D_f^Γ on $\mathcal{P}(S)\times\mathcal{P}(S)$ follows from (17), which shows that $D_f^\Gamma(Q\|P)$ is the supremum of functions that are affine in (Q,P).

3. Lemma 49 implies $f^*(\nu_0) = \nu_0$, where $\nu_0 \equiv f'_+(1)$. By assumption, $c_0 \in \Gamma\cap\mathbb{R}$, hence bounding (17) below by its value at $g = c_0$, $\nu = c_0 - \nu_0$ we find

$$\begin{aligned}
D_f^\Gamma(Q\|P) &\geq E_Q[c_0 - (c_0 - \nu_0)] - E_P[f^*(c_0 - (c_0 - \nu_0))] \\
&= \nu_0 - f^*(\nu_0) = 0\,.
\end{aligned}$$

Remark 73 *Note the importance of the infimum over ν in Λ_f^P, which allowed us to obtain a lower bound at an appropriate value of ν. This same technique will be used several times below and highlights the importance of employing Λ_f^P in the definition* (15) *of D_f^Γ. See also Remark 75.*

4. Now suppose f and Γ satisfy 2.a - 2.c Lemma 49 implies that

$$f^*(\nu_0) = \nu_0\,, \quad (f^*)'(\nu_0) = 1\,. \tag{80}$$

Assumption 2.a.ii implies there exists $c_0 \in \Gamma \cap \mathbb{R}$ hence Part 3 of this theorem implies $D_f^\Gamma(Q\|P) \geq 0$. From Eq. (79) we have $D_f^\Gamma(Q\|P) \leq D_f(Q\|P)$. Combining this with the non-negativity of D_f^Γ and the fact that $D_f(P\|P) = 0$ we see that if $Q = P$ then $D_f^\Gamma(Q\|P) = 0$.

Next assume $D_f^\Gamma(Q\|P) = 0$: From assumption 2.a.ii, given $\psi \in \Psi$ there exists $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ such that $g_\epsilon \equiv c_0 + \epsilon\psi \in \Gamma \cap \mathcal{M}_b(\Omega)$ for all $|\epsilon| < \epsilon_0$. Therefore

$$\begin{aligned} 0 =& D_f^\Gamma(Q\|P) \geq E_Q[g_\epsilon] - \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g_\epsilon - \nu)]\} \\ \geq& E_Q[g_\epsilon] - (c_0 - \nu_0 + E_P[f^*(\nu_0 + \epsilon\psi)]) \\ =& (\nu_0 + \epsilon E_Q[\psi]) - E_P[f^*(\nu_0 + \epsilon\psi)]) \equiv h(\epsilon)\,. \end{aligned}$$

By assumption 2.c, there exists $\delta > 0$ such that f^* is finite and C^1 on the ball of radius δ centered at ν_0, denoted by $B_\delta(\nu_0)$. ψ is bounded, therefore we can find $C > 0$ with $|\psi| \leq C$. For $|\epsilon| < \min\{\epsilon_0, \delta/(2C)\}$ we have $\mathrm{Range}(\nu_0 + \epsilon\psi) \subset B_{\delta/2}(\nu_0)$. On $B_{\delta/2}(\nu_0)$, f^* is C^1 and f^*, $(f^*)'$ are both bounded. Hence the dominated convergence theorem implies h is C^1 on $|\epsilon| < \min\{\epsilon_0, \delta/C\}$ and $h'(\epsilon) = E_Q[\psi] - E_P[(f^*)'(\nu_0 + \epsilon\psi)\psi]$. Evaluating this at $\epsilon = 0$ and using Eq. (80) we find

$$h'(0) = E_Q[\psi] - E_P[(f^*)'(\nu_0)\psi] = E_Q[\psi] - E_P[\psi]\,.$$

Again using (80) we can also compute $h(0) = \nu_0 - E_P[f^*(\nu_0)] = 0$. Combining these facts with the bound $h(\epsilon) \leq 0$ we can conclude that $h'(0) = 0$ and hence $E_Q[\psi] = E_P[\psi]$ for all $\psi \in \Psi$. By assumption 2.a.i, Ψ is $\mathcal{P}(\Omega)$-determining and so $Q = P$. This completes the proof of the divergence property for D_f^Γ.

The divergence property for W^Γ then follows from the divergence property for D_f^Γ together with the bound $D_f^\Gamma(Q\|P) \leq W^\Gamma(Q,P)$ and the definition (16).

■

Next we prove the infimal convolution formula, (22), as well as the other properties from Theorem 15, again in somewhat greater generality.

Theorem 74 *Suppose f and Γ are admissible. For $P \in \mathcal{P}(S)$, $\mu \in M(S)$ let $D_f^\Gamma(\mu\|P)$ be defined by* (75) *and for $\mu, \kappa \in M(S)$ let $W^\Gamma(\mu,\kappa)$ be defined as in* (76)*. These have the following properties:*

1. *Infimal Convolution Formula:*

$$D_f^\Gamma(\mu\|P) = \inf_{\eta\in\mathcal{P}(S)} \{D_f(\eta\|P) + W^\Gamma(\mu,\eta)\}\,. \tag{81}$$

In particular, $D_f^\Gamma(\mu\|P) \le W^\Gamma(\mu,P)$ *and if* $Q\in\mathcal{P}(S)$ *then* $D_f^\Gamma(Q\|P)\le D_f(Q\|P)$.

2. *If* $D_f^\Gamma(\mu\|P)<\infty$ *then there exists* $\eta_*\in\mathcal{P}(S)$ *such that*

$$D_f^\Gamma(\mu\|P) = D_f(\eta_*\|P) + W^\Gamma(\mu,\eta_*)\,. \tag{82}$$

If f *is strictly convex then there is a unique such* η_*.

3. *Divergence Property for* W^Γ*:* $W^\Gamma\ge 0$ *and* $W^\Gamma(\mu,\mu)=0$ *for all* $\mu\in M(S)$. *If* Γ *is strictly admissible then for all* $Q,P\in\mathcal{P}(S)$ *we have* $W^\Gamma(Q,P)=0$ *if and only if* $Q=P$.

4. *Divergence Property for* D_f^Γ*: For* $Q,P\in\mathcal{P}(S)$ *we have* $D_f^\Gamma(Q\|P)\ge 0$ *and* $D_f^\Gamma(P\|P)=0$. *If* f *and* Γ *are both strictly admissible then* $D_f^\Gamma(Q\|P)=0$ *if and only if* $Q=P$.

Proof

1. Define $H_1,H_2: C_b(S)\to(-\infty,\infty]$ by

$$H_1(g) = \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\} \tag{83}$$

and $H_2(g)=\infty 1_{\Gamma^c}(g)$ (note that $H_1>-\infty$ follows from the bound $f^*(y)\ge y$; see Lemma 42). We first show that H_1 and H_2 are convex and LSC. To see that H_1 is convex, note that convexity of f^* implies that the map $(g,\nu)\mapsto \nu + E_P[f^*(g-\nu)]$ is convex on $C_b(S)\times\mathbb{R}$. Therefore, taking the infimum over ν results in a convex function of g; see Theorem 2.2.6 in Bot et al. (2009). To show lower semicontinuity of H_1, first recall the variational formula (66)

$$\inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\} = \sup_{h\in\mathcal{M}_b(S):E_P[h]=1,E_P[f(h)]<\infty}\{E_P[gh]-E_P[f(h)]\}\,. \tag{84}$$

We can write $E_P[gh]=\int g dP_h$ where $dP_h\equiv h dP\in M(S)$. Recalling that $C_b(S)^* = \{\tau_\eta : \eta\in M(S)\}$, $\tau_\eta(g)\equiv\int g d\eta$ we see that $g\mapsto E_P[gh]$ is continuous on $C_b(S)$. Therefore Eq. (84) expresses H_1 as the supremum of a family of continuous functions, thus proving H_1 is LSC. H_2 is LSC and convex since Γ is closed and convex. We can write $D_f^\Gamma(\mu\|P)$ as the infinite-dimensional convex conjugate of H_1+H_2:

$$D_f^\Gamma(\mu\|P) = \sup_{g\in C_b(S)}\{\tau_\mu(g) - (H_1(g)+H_2(g))\} = (H_1+H_2)^*(\tau_\mu)\,.$$

We will now use the theory of infimal convolutions to compute the convex conjugate of H_1+H_2. Under appropriate assumptions, this theory allows one to show

$$(H_1+H_2)^*(\tau) = \inf\{H_1^*(\tau_1)+H_2^*(\tau_2) : \tau_1+\tau_2=\tau\} \equiv (H_1^*\square H_2^*)(\tau)\,,\ \ \tau\in C_b(S)^*\,, \tag{85}$$

where $H_1^*\square H_2^*$ is called the infimal convolution; see, e.g., Chapter 2 in Bot et al. (2009) for further information on infimal convolutions. Specifically, using Theorem 2.3.10 in Bot et al. (2009), if $\mathrm{dom}H_1 \cap \mathrm{dom}H_2 \neq \emptyset$ and $H_1^*\square H_2^*$ is LSC in the weak-* topology on $C_b(S)^*$ then $(H_1+H_2)^* = H_1^*\square H_2^*$. To show the first condition, note that f^* is not identically equal to ∞ and $0 \in \Gamma$; using these facts it is straightforward to show that $0 \in \mathrm{dom}H_1 \cap \mathrm{dom}H_2$. Therefore if we can prove lower semicontinuity of $H_1^*\square H_2^*$ then we can conclude (85). To accomplish this, first rewrite

$$H_1^*\square H_2^*(\tau_\mu) = \inf_{\eta \in M(S)} \{H_1^*(\tau_\eta) + W^\Gamma(\mu,\eta)\}\,. \tag{86}$$

Next we show that the infimum in (86) can be restricted to $\mathcal{P}(S)$. We do this in two steps:

(a) $H_1^*(\tau_\eta) = \infty$ when η is not positive: To show this, first note that

$$H_1^*(\tau_\eta) \geq \sup_{g \in C_b(S)} \left\{ \int g d\eta - E_P[f^*(g)] \right\}.$$

If there exists a measurable set $F \subset S$ with $\eta(F) < 0$ then by the extension of Lusin's theorem found in Appendix D of Dudley (2014), for any $\epsilon > 0$ there exists a closed set $E_\epsilon \subset S$ such that $|\eta|(E_\epsilon^c) < \epsilon$ and a $g_\epsilon \in C_b(S)$ such that $0 \leq g_\epsilon \leq 1$ and $g_\epsilon = 1_F$ on E_ϵ. For $n \in \mathbb{Z}^+$ define $g_{n,\epsilon} = -ng_\epsilon \in C_b(S)$. The assumption that f is admissible implies $\lim_{y\to-\infty} f^*(y) < \infty$. Therefore $f^*(g_{n,\epsilon})$ is bounded above independent of n,ϵ, and so there exists $D \in \mathbb{R}$ with

$$\begin{aligned} H_1^*(\tau_\eta) \geq& -n\int g_\epsilon d\eta - D = n|\eta(F)| + n\int 1_{E_\epsilon^c} 1_F d\eta - n\int_{E_\epsilon^c} g_\epsilon d\eta - D \\ \geq& n|\eta(F)| - 2\epsilon n - D\,. \end{aligned}$$

Letting $\epsilon < |\eta(F)|/2$ and sending $n \to \infty$ proves the claim.

(b) $H_1^*(\tau_\eta) = \infty$ if $\eta(S) \neq 1$: For $c \in \mathbb{R}$ we can use the fact that $(f^*)^* = f$ to compute

$$\begin{aligned} H_1^*(\tau_\eta) \geq& c\eta(S) - \inf_{\nu\in\mathbb{R}}\{\nu + f^*(c-\nu)\} = c(\eta(S)-1) + \sup_{\nu\in\mathbb{R}}\{c - \nu - f^*(c-\nu)\} \\ =& c(\eta(S)-1) + f(1) = c(\eta(S)-1)\,. \end{aligned} \tag{87}$$

Taking $c \to \pm\infty$ proves the claim.

Remark 75 *In light of the variational formula* (10)*, one might be motivated to define D_f^Γ using $E_P[f^*(g)]$ in place of $H_1(g)$. However, the property proven in Eq.* (87) *would fail in that case and we would be unable to proceed with our method of proof. If Γ is closed under the shift transformations $g \mapsto g - \nu$, $\nu \in \mathbb{R}$, then the choice of $H_1(g)$ versus $E_P[f^*(g)]$ does not impact the value of $D_f^\Gamma(\mu\|P)$ when $\mu = Q \in \mathcal{P}(S)$, but it can change the value if $\mu \in M(S)\setminus\mathcal{P}(S)$ and the choice can also impact the performance of numerical computations (see Ruderman et al., 2012 and Birrell et al., 2020 for discussion of this issue in the context of classical f-divergences).*

Having proven the above two properties we can now conclude that $H_1^*(\tau_\eta) = \infty$ if $\eta \notin \mathcal{P}(S)$, hence

$$H_1^* \square H_2^*(\tau_\mu) = \inf_{\eta \in \mathcal{P}(S)} \{D_f(\eta \| P) + W^\Gamma(\mu, \eta)\}\,, \tag{88}$$

where we used Corollary 51 and Remark 53 to evaluate $H_1^*(\tau_\eta)$. To prove lower semicontinuity of $H_1^* \square H_2^*$, let $a \in \mathbb{R}$ and take a net $\{\tau_{\mu_\alpha}\}_{\alpha \in A}$ in $\{H_1^* \square H_2^* \leq a\}$ with $\tau_{\mu_\alpha} \to \tau_\mu$. For any $\epsilon > 0$, Eq. (88) implies there exists $\eta_{\alpha,\epsilon} \in \mathcal{P}(S)$ with

$$\epsilon + a > D_f(\eta_{\alpha,\epsilon} \| P) + W^\Gamma(\mu_\alpha, \eta_{\alpha,\epsilon}) \geq D_f(\eta_{\alpha,\epsilon} \| P)\,,$$

where in the second inequality we used the fact that $W^\Gamma \geq 0$ (to see this, bound it below by taking $g = 0$ in Equation 76). $D_f(\cdot \| P)$ is lower semicontinuous (see Corollary 51) and has compact sublevel sets in the Prokhorov-metric topology (see Lemma 54), so there exists a subnet $\eta_{\alpha_\beta,\epsilon}$, $\beta \in B$ (where B is some directed set) with $\eta_{\alpha_\beta,\epsilon} \to \eta_\epsilon$ in the Prokhorov metric (i.e., weakly) and

$$D_f(\eta_\epsilon \| P) \leq \liminf_\beta D_f(\eta_{\alpha_\beta,\epsilon} \| P) \equiv \sup_{\tilde{\beta}} \inf_{\beta \geq \tilde{\beta}} D_f(\eta_{\alpha_\beta,\epsilon} \| P)\,.$$

The weak-* topology on $C_b(S)^*$ is generated by $\{\pi_g : g \in C_b(S)\}$, $\pi_g(\tau) \equiv \tau(g)$ and we have

$$W^\Gamma(\mu, \eta) = w^\Gamma(\tau_\mu, \tau_\eta)\,, \tag{89}$$

where $w^\Gamma : C_b(S)^* \times C_b(S)^* \to [0, \infty]$ is given by

$$w^\Gamma = \sup_{g \in \Gamma} \{\pi_g \circ \pi_1 - \pi_g \circ \pi_2\} \tag{90}$$

and is therefore LSC in the product topology (π_1, π_2 denote the projection maps onto the first and second components). By assumption, $\tau_{\mu_\alpha} \to \tau_\mu$ in the weak-* topology. The fact that $\eta_{\alpha_\beta,\epsilon} \to \eta_\epsilon$ weakly implies that $\tau_{\eta_{\alpha_\beta,\epsilon}} \to \tau_{\eta_\epsilon}$ in the weak-* topology as well. Therefore $(\tau_{\mu_{\alpha_\beta}}, \tau_{\eta_{\alpha_\beta,\epsilon}}) \to (\tau_\mu, \tau_{\eta_\epsilon})$ in the product topology on $C_b(S)^* \times C_b(S)^*$. Lower semicontinuity of w^Γ and the equality (89) then imply

$$W^\Gamma(\mu, \eta_\epsilon) \leq \liminf_\beta W^\Gamma(\mu_{\alpha_\beta}, \eta_{\alpha_\beta,\epsilon})\,.$$

Next we need the following simple to prove lemma about nets in $(-\infty, \infty]$: Let x_β, y_β, $\beta \in B$ be nets in $(-\infty, \infty]$. If x_β and y_β are nondecreasing then

$$\sup_\beta \{x_\beta + y_\beta\} = \sup_\beta x_\beta + \sup_\beta y_\beta\,.$$

Using this we have

$$\begin{aligned}
\epsilon + a \geq & \sup_{\tilde{\beta}\in B} \inf_{\beta\geq\tilde{\beta}} \left\{ D_f(\eta_{\alpha_\beta,\epsilon}\|P) + W^\Gamma(\mu_{\alpha_\beta}, \eta_{\alpha_\beta,\epsilon}) \right\} \\
\geq & \sup_{\tilde{\beta}\in B} \left\{ \inf_{\beta\geq\tilde{\beta}} D_f(\eta_{\alpha_\beta,\epsilon}\|P) + \inf_{\beta\geq\tilde{\beta}} W^\Gamma(\mu_{\alpha_\beta}, \eta_{\alpha_\beta,\epsilon}) \right\} \\
= & \sup_{\tilde{\beta}\in B} \inf_{\beta\geq\tilde{\beta}} D_f(\eta_{\alpha_\beta,\epsilon}\|P) + \sup_{\tilde{\beta}\in B} \inf_{\beta\geq\tilde{\beta}} W^\Gamma(\mu_{\alpha_\beta}, \eta_{\alpha_\beta,\epsilon}) \\
= & \liminf_\beta D_f(\eta_{\alpha_\beta,\epsilon}\|P) + \liminf_\beta W^\Gamma(\mu_{\alpha_\beta}, \eta_{\alpha_\beta,\epsilon}) \\
\geq & D_f(\eta_\epsilon\|P) + W^\Gamma(\mu, \eta_\epsilon) \\
\geq & \inf_{\eta\in\mathcal{P}(S)} \{ D_f(\eta\|P) + W^\Gamma(\mu,\eta) \} \\
= & H_1^* \square H_2^*(\tau_\mu) \,.
\end{aligned}$$

This holds for all $\epsilon > 0$ and so $\tau_\mu \in \{H_1^* \square H_2^* \leq a\}$. This proves that $\{H_1^* \square H_2^* \leq a\}$ is closed for all $a \in \mathbb{R}$ and hence we have proven lower semicontinuity of $H_1^* \square H_2^*$. Therefore we can conclude that

$$\begin{aligned}
D_f^\Gamma(\mu\|P) = & (H_1 + H_2)^*(\tau_\mu) = H_1^* \square H_2^*(\tau_\mu) \\
= & \inf_{\eta\in\mathcal{P}(S)} \{ D_f(\eta\|P) + W^\Gamma(\mu,\eta) \}
\end{aligned}$$

as claimed.

2. If $D_f^\Gamma(\mu\|P) < \infty$ then there exists $\eta_n \in \mathcal{P}(S)$ such that

$$D_f^\Gamma(\mu\|P) = \lim_n (D_f(\eta_n\|P) + W^\Gamma(\mu,\eta_n)) \,, \tag{91}$$

with $D_f(\eta_n\|P) + W^\Gamma(\mu,\eta_n)$ finite for all n. $W^\Gamma \geq 0$ and so (91) implies $D_f(\eta_n\|P)$ is a bounded sequence, i.e., there exists $M \in \mathbb{R}$ with $\eta_n \in \{Q : D_f(Q\|P) \leq M\}$. Lemma 54 implies $Q \mapsto D_f(Q\|P)$ has compact sublevel sets, hence there exists a convergence subsequence $\eta_{n_j} \to \eta_* \in \mathcal{P}(S)$. By Corollary 51, $(Q,P) \mapsto D_f(Q\|P)$ is LSC and so $D_f(\eta_*\|P) \leq \liminf_j D_f(\eta_{n_j}\|P)$. $\eta \mapsto W^\Gamma(\mu,\eta)$ is the supremum of a collection of continuous functions on $\mathcal{P}(S)$, and so is also LSC. Therefore $W^\Gamma(\mu,\eta_*) \leq \liminf_j W^\Gamma(\mu,\eta_{n_j})$ and we have

$$\begin{aligned}
& D_f^\Gamma(\mu\|P) \leq D_f(\eta_*\|P) + W^\Gamma(\mu,\eta_*) \leq \liminf_j D_f(\eta_{n_j}\|P) + \liminf_j W^\Gamma(\mu,\eta_{n_j}) \\
\leq & \liminf_j (D_f(\eta_{n_j}\|P) + W^\Gamma(\mu,\eta_{n_j})) = D_f^\Gamma(\mu\|P) \,.
\end{aligned}$$

This completes the proof of (82). If f is strictly convex then the map $\eta \mapsto D_f(\eta\|P)$ is strictly convex on the set $\{\eta : D_f(\eta\|P) < \infty\}$ (see Lemma 55). It is straightforward to see that $\eta \mapsto W^\Gamma(\mu,\eta)$ is convex. Therefore the objective functional in (81) is strictly convex and hence has at most one minimizer.

3. We have already noted that $W^\Gamma \geq 0$. The property $W^\Gamma(\mu,\mu) = 0$ is trivial from the definition. If Γ is strictly admissible and $Q, P \in \mathcal{P}(S)$ with $W^\Gamma(Q,P) = 0$ then for $\psi \in \Psi$ we can find $c \in \mathbb{R}$, $\epsilon > 0$ such that $c \pm \epsilon\psi \in \Gamma$. Therefore

$$0 = W^\Gamma(Q,P) \geq E_Q[c \pm \epsilon\psi] - E_P[c \pm \epsilon\psi] = \pm\epsilon(E_Q[\psi] - E_P[\psi])\,.$$

From this we can conclude that $E_Q[\psi] = E_P[\psi]$ for all $\psi \in \Psi$ and hence $Q = P$.

4. Both D_f and W^Γ are non-negative, hence the infimal convolution formula (81) implies $D_f^\Gamma \geq 0$. By taking $\eta = P$ in (81) it is easy to see that $D_f^\Gamma(P\|P) = 0$. Now suppose f and Γ are strictly admissible. Strict convexity of f at 1 implies that D_f has the divergence property (Liese and Vajda, 2006). If $Q, P \in \mathcal{P}(S)$ with $D_f^\Gamma(Q\|P) = 0$ then Eq. (82) implies there exists $\eta_* \in \mathcal{P}(S)$ with

$$0 = D_f^\Gamma(Q\|P) = D_f(\eta_*\|P) + W^\Gamma(Q,\eta_*)\,.$$

Therefore $D_f(\eta_*\|P) = 0 = W^\Gamma(Q,\eta_*)$. The first equality implies $\eta_* = P$ and so we have $W^\Gamma(Q,P) = 0$. We have assumed Γ is strictly admissible, hence Part 3 implies $Q = P$.

■

Under slightly stronger assumptions we find that $D_f^\Gamma(\mu\|P)$ is infinite when $\mu \notin M_1(S) \equiv \{\mu \in M(S) : \mu(S) = 1\}$.

Corollary 76 *Suppose f and Γ are admissible and Γ contains the constant functions. Then for $P \in \mathcal{P}(S)$, $\mu \in M(S) \setminus M_1(S)$ we have $D_f^\Gamma(\mu\|P) = \infty$.*

Proof We have assumed that Γ contains the constant functions. Therefore, for any $\eta \in \mathcal{P}(S)$ we have

$$W^\Gamma(\mu,\eta) \geq c(\mu(S) - \eta(S)) = c(\mu(S) - 1)$$

for all $c \in \mathbb{R}$. If $\mu(S) \neq 1$ then taking $c \to \pm\infty$ implies $W^\Gamma(\mu,\eta) = \infty$. This holds for all $\eta \in \mathcal{P}(S)$ and so Eq. (81) implies $D_f^\Gamma(\mu\|P) = \infty$. ■

Here we prove that, under appropriate assumptions, the unit ball in a RKHS is an admissible set and hence falls under the purview of Theorem 15. See Chapter 4 in Steinwart and Christmann (2008) for a detailed treatment of the properties of an RKHS, several of which are used below.

Lemma 77 *Let $X \subset C_b(S)$ be a separable RKHS with reproducing-kernel $k : S \times S \to \mathbb{R}$. Let $\Gamma = \{g \in X : \|g\|_X \leq 1\}$ be the unit ball in X. Then Γ is admissible.*

Proof We clearly have $0 \in \Gamma$ and Γ is convex. Therefore we just need to show that Γ is a closed subset of $C_b(S)$ under the weak topology generated by $M(S)$: First note that Γ is compact in the weak topology induced by X^*; this follows from Alaoglu's theorem (see, e.g., Theorem 5.18 in Folland, 2013) together with the fact that X is reflexive.

Next we show that the topology on Γ induced by $M(S)$ is the same as the topology induced by X^*. To do this, first recall that the assumption $X \subset C_b(S)$ implies k is bounded, separately continuous, and jointly measurable. This allows us to define the linear map $\mu_X : M(S) \to X$ by $\mu_X(\nu) = \int k(\cdot, x)\nu(dx)$ that satisfies

$$\tau_\nu(g) \equiv \int g d\nu = \langle g, \mu_X(\nu)\rangle_X$$

for all $g \in X$. This shows that $\tau_\nu \in X^*$ for all $\nu \in M(S)$. Therefore the topology on Γ induced by $M(S)$ is weaker than the topology induced by X^*. The former is Hausdorff, since $M(S)$ separates points, and, as shown above, the latter is compact. Therefore the two topologies are in fact equal (see, e.g., Proposition 4.28 in Folland, 2013).

Combining the above two properties we conclude that the weak topology on Γ induced by $M(S)$ is compact. The topology induced by $M(S)$ on $C_b(S)$ is Hausdorff (again, because $M(S)$ separates points) and Γ is a compact subset of this space, hence we have proven that Γ is closed in $C_b(S)$. ■

Remark 78 *By imposing various additional conditions on the kernel (e.g., if it is characteristic or universal) one can ensure that the unit ball in X is measure determining and hence is strictly admissible; see Sriperumbudur et al. (2011) and references therein.*

Now we prove the limiting properties from Theorem 17, which are repeated below.

Theorem 79 *Let $Q, P \in \mathcal{P}(S)$ and Γ, f both be admissible. Then for all $c > 0$ the set $\Gamma_c \equiv \{cg : g \in \Gamma\}$ is admissible and we have the following two limiting formulas.*

1. *If Γ is strictly admissible then the sets Γ_L are strictly admissible for all $L > 0$ and*

$$\lim_{L\to\infty} D_f^{\Gamma_L}(Q\|P) = D_f(Q\|P)\,.$$

2. *If f is strictly admissible then*

$$\lim_{\delta \searrow 0} \frac{1}{\delta} D_f^{\Gamma_\delta}(Q\|P) = W^\Gamma(Q,P)\,.$$

Proof

1. From the definition, it is straightforward to see that Γ_c is strictly admissible for all $c > 0$ and $W^{\Gamma_c}(\mu,\kappa) = cW^\Gamma(\mu,\kappa)$. To prove that $\lim_{L\to\infty} D_f^{\Gamma_L}(Q\|P) = D_f(Q\|P)$, first suppose that $D_f(Q\|P) < \infty$: From Eq. (81) we see that $D_f^{\Gamma_L}(Q\|P) \le D_f(Q\|P) < \infty$ for all $L > 0$ and (82) implies that there exists $\eta_{*,L} \in \mathcal{P}(S)$ such that $D_f^{\Gamma_L}(Q\|P) = D_f(\eta_{*,L}\|P) + W^{\Gamma_L}(Q,\eta_{*,L})$. Take a sequence $L_n \nearrow \infty$. We have

$$D_f(\eta_{*,L_n}\|P) \le D_f^{\Gamma_L}(Q\|P) \le D_f(Q\|P) \equiv M < \infty\,,$$

and so for all n we have $\eta_{*,L_n} \in \{D_f(\cdot\|P) \leq M\}$, a compact set (see Lemma 54). Hence there exists a weakly convergence subsequence $\eta_{*,L_{n_j}} \to \eta_*$. We can compute

$$\begin{aligned} W^\Gamma(Q,\eta_*) \leq& \liminf_j W^\Gamma(Q,\eta_{*,L_{n_j}}) = \liminf_j \frac{1}{L_{n_j}} W^{\Gamma_{L_{n_j}}}(Q,\eta_{*,L_{n_j}}) \\ \leq& \liminf_j \frac{1}{L_{n_j}} D_f^{\Gamma_{L_{n_j}}}(Q\|P) \leq \liminf_j \frac{1}{L_{n_j}} D_f(Q\|P) = 0\,. \end{aligned}$$

Therefore $W^\Gamma(Q,\eta_*) = 0$. Γ is strictly admissible, hence W^Γ has the divergence property (see Part 3 of Theorem 15) and so $\eta_* = Q$. Therefore we can use lower semicontinuity of D_f to compute

$$\begin{aligned} \liminf_j D_f^{\Gamma_{L_{n_j}}}(Q\|P) =& \liminf_j (D_f(\eta_{*,L_{n_j}}\|P) + W^{\Gamma_L}(Q,\eta_{*,L_{n_j}})) \\ \geq& \liminf_j D_f(\eta_{*,L_{n_j}}\|P) \geq D_f(Q\|P)\,. \end{aligned}$$

Combining this with the fact that $D_f^{\Gamma_{L_n}}(Q\|P) \leq D_f(Q\|P)$ we find $\lim_j D_f^{\Gamma_{L_{n_j}}}(Q\|P) = D_f(Q\|P)$. Therefore we have shown that every sequence $L_n \nearrow \infty$ has a subsequence with $D_f^{\Gamma_{L_{n_j}}}(Q\|P) \to D_f(Q\|P)$. This implies $D_f^{\Gamma_{L_n}}(Q\|P) \to D_f(Q\|P)$ and so we have proven the result in the case where $D_f(Q\|P) < \infty$.

Now suppose $D_f(Q\|P) = \infty$: If $\lim_{L\to\infty} D_f^{\Gamma_L}(Q\|P) \neq \infty$ then there exists $R \in \mathbb{R}$ and $L_n \to \infty$ with $D_f^{\Gamma_{L_n}}(Q\|P) \leq R$ for all n. Eq. (82) implies there exists $\eta_{*,n} \in \mathcal{P}(S)$ such that

$$R \geq D_f^{\Gamma_{L_n}}(Q\|P) = D_f(\eta_{*,n}\|P) + W^{\Gamma_{L_n}}(Q,\eta_{*,n}) \geq D_f(\eta_{*,n}\|P)\,. \tag{92}$$

Using compactness of sublevel sets we again see that there is a convergent subsequence $\eta_{*,n_j} \to \eta_*$. Similarly to (92), we can compute

$$R \geq D_f^{\Gam_{L_n}}(Q\|P) = D_f(\eta_{*,n}\|P) + W^{\Gamma_{L_n}}(Q,\eta_{*,n}) \geq W^{\Gamma_{L_n}}(Q,\eta_{*,n}) = L_n W^\Gamma(Q,\eta_{*,n})\,.$$

This implies

$$R/L_n \geq W^\Gamma(Q,\eta_{*,n}) \geq 0\,,$$

and so $W^\Gamma(Q,\eta_{*,n}) \to 0$. Γ is strictly admissible and $\eta_{*,n_j} \to \eta_*$ weakly, hence a similar argument to that of the previous case implies $Q = \eta_*$ and

$$\begin{aligned} D_f(Q\|P) \leq& \liminf_j D_f(\eta_{*,n_j}\|P) \leq \liminf_j (D_f(\eta_{*,n_j}\|P) + W^{\Gamma_{L_{n_j}}}(Q,\eta_{*,n_j})) \\ =& \liminf_j D_f^{\Gamma_{L_{n_j}}}(Q\|P) \leq R\,. \end{aligned}$$

Therefore $D_f(Q\|P) \leq R < \infty$, a contradiction. This completes the proof.

2. It is again straightforward to see that Γ_δ is admissible and $W^{\Gamma_\delta} = \delta W^\Gamma$. Fix $Q, P \in \mathcal{P}(S)$ and take a sequence $\delta_n \searrow 0$. Using the infimal convolution formula (81) we see that

$$\frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) = \inf_{\eta\in\mathcal{P}(S)} \{\delta_n^{-1} D_f(\eta\|P) + W^\Gamma(Q,\eta)\} \leq W^\Gamma(Q,P)$$

and the left-hand-side is nondecreasing in n. Therefore

$$\lim_{n\to\infty} \frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) = \sup_n \frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) \leq W^\Gamma(Q,P)\,.$$

Suppose $\sup_n \frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) < W^\Gamma(Q,P)$: This implies $D_f^{\Gamma_{\delta_n}}(Q\|P) < \infty$ for all n, hence Eq. (82) implies there exists $\eta_{*,n} \in \mathcal{P}(S)$ such that $D_f^{\Gamma_{\delta_n}}(Q\|P) = D_f(\eta_{*,n}\|P) + W^{\Gamma_{\delta_n}}(Q,\eta_{*,n})$. Therefore

$$\infty > \sup_n \frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) \geq \sup_n \frac{1}{\delta_n} D_f(\eta_{*,n}\|P)\,.$$

$\delta_n \searrow 0$ and so this implies $D_f(\eta_{*,n}\|P)$ is uniformly bounded. $\eta \mapsto D_f(\eta\|P)$ has compact sublevel sets (see Lemma 54), hence there exists a weakly convergent subsequence $\eta_{*,n_j} \to \eta_*$. The fact that $\sup_n \frac{1}{\delta_n} D_f(\eta_{*,n}\|P) < \infty$ together with $\delta_n \searrow 0$ implies $D_f(\eta_{*,n}\|P) \to 0$. D_f is LSC and $\eta_{*,n_j} \to \eta_*$ so this implies

$$0 \leq D_f(\eta_*\|P) \leq \liminf_j D_f(\eta_{*,n_j}\|P) = 0\,,$$

i.e., $D_f(\eta_*\|P) = 0$. f is strictly convex at 1, hence D_f has the divergence property and so $\eta_* = P$. Therefore $\eta_{*,n_j} \to P$ weakly and we can compute

$$\begin{aligned} W^\Gamma(Q,P) &> \sup_n \frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) \geq \liminf_j \frac{1}{\delta_{n_j}} D_f^{\Gamma_{\delta_{n_j}}}(Q\|P) \\ &\geq \liminf_j \frac{1}{\delta_{n_j}} W^{\Gamma_{\delta_{n_j}}}(Q,\eta_{*,n_j}) = \liminf_j W^\Gamma(Q,\eta_{*,n_j}) \geq W^\Gamma(Q,P)\,. \end{aligned}$$

This is a contradiction, hence we can conclude that

$$\lim_{n\to\infty} \frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) = \sup_n \frac{1}{\delta_n} D_f^{\Gamma_{\delta_n}}(Q\|P) = W^\Gamma(Q,P)\,.$$

$\delta_n \searrow 0$ was arbitrary, therefore

$$\lim_{\delta\searrow 0} \frac{1}{\delta} D_f^{\Gamma_\delta}(Q\|P) = W^\Gamma(Q,P)\,.$$

■

Next we prove the convergence and continuity results from Theorem 18.

Theorem 80 *Let $f \in \mathcal{F}_1(a,b)$ and $\Gamma \subset \mathcal{M}_b(\Omega)$. Then:*

1. *If there exists* $c_0 \in \Gamma \cap \mathbb{R}$ *then* $W^\Gamma(Q_n, P) \to 0 \implies D_f^\Gamma(Q_n\|P) \to 0$ *and* $D_f(Q_n\|P) \to 0 \implies D_f^\Gamma(Q_n\|P) \to 0$*, and similarly if one exchanges* Q_n *and* P*.*

2. *Suppose* f *and* Γ *also satisfy the following:*

   (a) *There exist a nonempty set* $\Psi \subset \Gamma$ *with the following properties:*

      i. Ψ *is* $\mathcal{P}(\Omega)$*-determining.*

      ii. *For all* $\psi \in \Psi$ *there exists* $c_0 \in \mathbb{R}$*,* $\epsilon_0 > 0$ *such that* $c_0 + \epsilon\psi \in \Gamma$ *for all* $|\epsilon| < \epsilon_0$*.*

   (b) f *is strictly convex on a neighborhood of* 1*.*

   (c) f^* *is finite and* C^1 *on a neighborhood of* $\nu_0 \equiv f'_+(1)$*.*

   Let $P, Q_n \in \mathcal{P}(\Omega)$*,* $n \in \mathbb{Z}_+$*. If* $D_f^\Gamma(Q_n\|P) \to 0$ *or* $D_f^\Gamma(P\|Q_n) \to 0$ *then* $E_{Q_n}[\psi] \to E_P[\psi]$ *for all* $\psi \in \Psi$*.*

3. *On a metric space* S*, if* f *is admissible then the map* $(Q, P) \in \mathcal{P}(S) \times \mathcal{P}(S) \mapsto D_f^\Gamma(Q\|P)$ *is lower semicontinuous.*

Proof Part 1 follows from the upper bound (21) and the lower bound from Part 3 of Theorem 8. Now work under the assumptions of Part 2 and suppose $D_f^\Gamma(Q_n\|P) \to 0$. Fix $\delta > 0$ and take N_δ such that for all $n \geq N_\delta$ we have $D_f^\Gamma(Q_n\|P) \leq \delta$. Fix $\psi \in \Psi$ and, per Assumption 2.a.ii, take $c_0 \in \mathbb{R}$ and $\epsilon_0 > 0$ such that $c_0 + \epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$. Using (17) we obtain

$$E_{Q_n}[\nu_0 + \epsilon\psi] - E_P[f^*(\nu_0 + \epsilon\psi)] \leq D_f^\Gamma(Q_n\|P) \leq \delta$$

for all $n \geq N_\delta$, $|\epsilon| < \epsilon_0$, where ν_0 is as in (80). Taylor expanding f^* then gives

$$E_{Q_n}[\nu_0 + \epsilon\psi] - E_P[f^*(\nu_0) + (f^*)'(\nu_0)\epsilon\psi + R(\epsilon\psi)\epsilon\psi] \leq \delta$$

for all $n \geq N_\delta$, $|\epsilon| < \epsilon_0$ (using a possibly smaller ϵ_0), where the remainder, R, is continuous, bounded, and satisfies $R(0) = 0$. The identities (80) then imply

$$\epsilon(E_{Q_n}[\psi] - E_P[\psi]) \leq \delta + \epsilon E_P[R(\epsilon\psi)\psi]$$

for all $n \geq N_\delta$, $|\epsilon| < \epsilon_0$. By appropriately choosing the sign of ϵ, we therefore find

$$\sup_{n \geq N_\delta} |E_{Q_n}[\psi] - E_P[\psi]| \leq \delta/\epsilon + \|\psi\|_\infty \sup_{[-\epsilon\|\psi\|_\infty, \epsilon\|\psi\|_\infty]} |R|$$

for all $0 < \epsilon < \epsilon_0$. For δ sufficiently small we can let $\epsilon = \delta^{1/2}$ and therefore find

$$\sup_{n \geq N_\delta} |E_{Q_n}[\psi] - E_P[\psi]| \leq \delta^{1/2} + \|\psi\|_\infty \sup_{[-\delta^{1/2}\|\psi\|_\infty, \delta^{1/2}\|\psi\|_\infty]} |R| \to 0$$

as $\delta \to 0$. Hence $E_{Q_n}[\psi] \to E_P[\psi]$ as claimed. The proof in the case where $D_f^\Gamma(P\|Q_n) \to 0$ is similar. Finally, Part 3 follows from (17), which show that $D_f^\Gamma(Q\|P)$ is the supremum of functions that are continuous in (Q, P), where the topology on $\mathcal{P}(S)$ is induced by the Prokhorov metric; here we used the fact that Lemma 41 implies f^* is finite and continuous on $\mathbb{R}$ and hence $f^*(g - \nu) \in C_b(S)$. ■

Now we derive the data processing inequality from Theorem 21.

Theorem 81 (Data Processing Inequality) *Let $f \in \mathcal{F}_1(a,b)$, $Q,P \in \mathcal{P}(\Omega)$, and K be a probability kernel from $(\Omega,\mathcal{M})$ to $(N,\mathcal{N})$.*

1. *Let $\Gamma \subset \mathcal{M}_b(N)$ be nonempty. Then*

$$D_f^\Gamma\left(K[Q]\|K[P]\right) \le D_f^{K[\Gamma]}(Q\|P)\,. \tag{93}$$

2. *Let $\Gamma \subset \mathcal{M}_b(\Omega \times N)$ be nonempty. Then*

$$D_f^\Gamma\left(Q\otimes K\|P\otimes K\right) \le D_f^{K[\Gamma]}(Q\|P)\,. \tag{94}$$

Proof From Eq. (17) we have

$$D_f^\Gamma(K[Q]\|K[P]) = \sup_{g\in\Gamma,\nu\in\mathbb{R}}\left\{\int\int (g(y)-\nu)K_x(dy)Q(dx) - \int\int f^*(g(y)-\nu)K_x(dy)P(dx)\right\}.$$

Using convexity of f^* we can apply Jensen's inequality to find

$$\int f^*(g(y)-\nu)K_x(dy) \ge f^*\left(\int (g(y)-\nu)K_x(dy)\right)$$

for all $x\in\Omega$. Hence

$$D_f^\Gamma(K[Q]\|K[P]) \le \sup_{g\in\Gamma,\nu\in\mathbb{R}}\left\{E_Q[K[g]-\nu]-E_P[f^*(K[g]-\nu)]\right\} = D_f^{K[\Gamma]}(Q\|P)\,.$$

This proves Eq. (93). The proof of Eq. (94) is very similar and so we omit it. ■

Next we prove (a generalization of) Theorem 24, which gives existence and uniqueness results regarding the dual optimization problem (12) for the classical f-divergences.

Theorem 82 *Let $f\in\mathcal{F}_1(a,b)$, $a\ge 0$, $P\in\mathcal{P}(\Omega)$, and $g\in\mathcal{M}_b(\Omega)$.*

1. *If f is strictly convex then the optimization problem*

$$\sup_{Q\in\mathcal{P}(\Omega)}\left\{E_Q[g]-D_f(Q\|P)\right\} \tag{95}$$

has at most one optimizer.

2. *Suppose there exists $\nu_*\in\mathbb{R}$ such that Range$(g-\nu_*)\subset\{f^*<\infty\}^o$ and*

$$E_P[(f^*)'_+(g-\nu_*)] = 1\,. \tag{96}$$

Then

$$dQ_* \equiv (f^*)'_+(g-\nu_*)dP \tag{97}$$

is a probability measure and

$$\sup_{Q\in\mathcal{P}(\Omega)}\left\{E_Q[g]-D_f(Q\|P)\right\} = E_{Q_*}[g]-D_f(Q_*\|P)$$
$$=\nu_* + E_P[f^*(g-\nu_*)] = \inf_{\nu\in\mathbb{R}}\left\{\nu+E_P[f^*(g-\nu)]\right\}.$$

3. If f is strictly convex on (a,b) and $\{f^* < \infty\} = \mathbb{R}$ then there exists $\nu_* \in \mathbb{R}$ such that

$$E_P[(f^*)'(g-\nu_*)] = 1\,.$$

Proof

1. We obviously have

$$\sup_{Q\in\mathcal{P}(\Omega)} \{E_Q[g] - D_f(Q\|P)\} = \sup_{Q\in\mathcal{P}(\Omega):D_f(Q\|P)<\infty} \{E_Q[g] - D_f(Q\|P)\}\,, \tag{98}$$

and optimizers (if they exist) must be in $\{Q : D_f(Q\|P) < \infty\}$. If f is strictly convex then Lemma 55 implies that the map $Q \mapsto D_f(Q\|P)$ is strictly convex on the set $\{Q : D_f(Q\|P) < \infty\}$. The objective functional $Q \mapsto E_Q[g] - D_f(Q\|P)$ is therefore strictly concave and hence has at most one maximizer.

2. Lemma 46 implies f^* is nondecreasing, and so $(f^*)'_+ \geq 0$. Together with the assumption (96), this implies Q_* is a probability measure. From Lemma 48 we have

$$f((f^*)'_+(g-\nu_*)) = (g-\nu_*)(f^*)'_+(g-\nu_*) - f^*(g-\nu_*) \tag{99}$$

and so we can compute

$$\begin{aligned}
&\sup_{Q\in\mathcal{P}(\Omega)} \{E_Q[g] - D_f(Q\|P)\} \geq E_{Q_*}[g] - D_f(Q_*\|P)\\
=&\nu_* + E_P[(g-\nu_*)(f^*)'_+(g-\nu_*) - f((f^*)'_+(g-\nu_*))]\\
=&\nu_* + E_P[f^*(g-\nu_*)] \geq \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}\\
=&\sup_{Q:D_f(Q\|P)<\infty} \{E_Q[g] - D_f(Q\|P)\}\,,
\end{aligned}$$

where we used Eq. (99) to go from the second to the third line and we used Eq. (67) to obtain the last line. The equality (98) then completes the proof.

3. Strict convexity of f implies f^* is C^1 (see Theorem 26.3 in Rockafellar, 1970). g is bounded and so the dominated convergence theorem implies that the map $h : \mathbb{R} \to \mathbb{R}$, $h(\nu) = E_P[(f^*)'(g-\nu)]$ is continuous. From Lemma 49 we see that $\nu_0 \equiv f'_+(1)$ satisfies $(f^*)'(\nu_0) = 1$. Convexity of f^* implies that $(f^*)'$ is nondecreasing, therefore

$$h(\|g\|_\infty - \nu_0) \leq (f^*)'(\nu_0) = 1$$

and

$$h(-\|g\|_\infty - \nu_0) \geq (f^*)'(\nu_0) = 1\,.$$

Continuity therefore implies there exists $\nu_* \in [-\|g\|_\infty - \nu_0, \|g\|_\infty - \nu_0]$ with $h(\nu_*) = 1$ as claimed.

■

Finally, we prove the characterization from Theorem 25 in the more general case of $D_f^\Gamma(\mu\|P)$ where $\mu \in M(S)$.

Theorem 83 *Let $\Gamma \subset C_b(S)$ be admissible and $f \in \mathcal{F}_1(a,b)$ be admissible, where $a \geq 0$. Fix $P \in \mathcal{P}(S)$, $\mu \in M(S)$. Suppose we have $g_* \in \Gamma$ and $\nu_* \in \mathbb{R}$ that satisfy the following:*

1. *$f((f^*)'_+(g_* - \nu_*)) \in L^1(P)$,*
2. *$E_P[(f^*)'_+(g_* - \nu_*)] = 1$,*
3. *$W^\Gamma(\mu, \eta_*) = \int g_* d\mu - \int g_* d\eta_*$, where $d\eta_* \equiv (f^*)'_+(g_* - \nu_*)dP$.*

Then $\eta_* \in \mathcal{P}(S)$ solves the infimal convolution problem (81) (Equation 22 for the case of $\mu = Q \in \mathcal{P}(S)$) and

$$D_f^\Gamma(\mu\|P) = \int g_* d\mu - (\nu_* + E_P[f^*(g_* - \nu_*)])\,. \tag{100}$$

If f is strictly convex then η_* is the unique solution to the infimal convolution problem.

Proof Admissibility of f implies $\{f^* < \infty\} = \mathbb{R}$. Convexity of f^* then implies that the right derivative $(f^*)'_+$ exists everywhere and so $\text{Range}(g_* - \nu_*) \subset \{(f^*)'_+ < \infty\}$. Therefore we can use Theorem 82 to conclude that

$$d\eta_* = (f^*)'_+(g_* - \nu_*)dP$$

is a probability measure, $D_f(\eta_*\|P) < \infty$, and

$$\begin{aligned}\sup_{Q: D_f(Q\|P)<\infty} \{E_Q[g_*] - D_f(Q\|P)\} =& E_{\eta_*}[g_*] - D_f(\eta_*\|P) = \nu_* + E_P[f^*(g_* - \nu_*)] \\ &= \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g_* - \nu)]\}\,.\end{aligned}$$

In particular,

$$D_f(\eta_*\|P) - E_{\eta_*}[g_*] = -(\nu_* + E_P[f^*(g_* - \nu_*)])\,.$$

Using Part 1 of Theorem 74 we can compute

$$\begin{aligned}D_f^\Gamma(\mu\|P) \leq& D_f(\eta_*\|P) + W^\Gamma(\mu,\eta_*) \\ =& D_f(\eta_*\|P) + \int g_* d\mu - \int g_* d\eta_* \\ =& \int g_* d\mu - (\nu_* + E_P[f^*(g_* - \nu_*)]) \\ =& \int g_* d\mu - \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g_* - \nu)]\} \\ \leq& \sup_{g\in\Gamma} \left\{ \int g d\mu - \inf_{\nu\in\mathbb{R}} \{\nu + E_P[f^*(g - \nu)]\} \right\} = D_f^\Gamma(\mu\|P)\,.\end{aligned}$$

Therefore η_* solves the infimal convolution problem and (100) holds. If f is strictly convex then, as shown in Theorem 74, the solution to the infimal convolution problem is unique. ■

Appendix D. Strict Concavity of the (f, Γ)-Divergence Objective Functional

Here we will (formally) compute the Taylor expansion of the objective functional in (15) for the (f, Γ)-divergences. The computation is very similar to the classical f-divergence case considered in Appendix C of Birrell et al. (2020). First define

$$H_f[g; Q, P] = E_Q[g] - \inf_{\nu \in \mathbb{R}} \{\nu + E_P[f^*(g - \nu)]\}$$

and note that this map is concave in g, due to the convexity of f^*. In fact, under weak assumptions it is strictly concave, as we now show. Take a line segment $g_\epsilon = g_0 + \epsilon\psi \in \Gamma$, $\epsilon \in (-\delta, \delta)$. We will compute $\frac{d^2}{d\epsilon^2}|_{\epsilon=0} H_f[g_\epsilon; Q, P]$.

The optimization problem $\inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g_\epsilon - \nu)]\}$ is solved by ν_ϵ that satisfies

$$0 = \partial_\nu|_{\nu=\nu_\epsilon}\{\nu + E_P[f^*(g_\epsilon - \nu)]\}\,,$$

i.e.,

$$E_P[(f^*)'(g_\epsilon - \nu_\epsilon)] = 1 \tag{101}$$

for all ϵ. Differentiating this at $\epsilon = 0$ we find

$$\nu_0' = E_{P_0}[\psi]\,, \quad dP_0 \equiv \frac{(f^*)''(g_0 - \nu_0)}{E_P[(f^*)''(g_0 - \nu_0)]} dP\,. \tag{102}$$

Note that convexity of f^* implies $E_P[(f^*)''(g_0 - \nu_0)] \geq 0$. We assume this inequality is strict. We can compute

$$\begin{aligned}
\frac{d}{d\epsilon}|_{\epsilon=0} H[g_\epsilon; Q, P] =& E_Q[\psi] - \nu_0' - E_P[(f^*)'(g_0 - \nu_0)\psi] + E_P[(f^*)'(g_0 - \nu_0)]\nu_0' \\
=& E_Q[\psi] - E_P[(f^*)'(g_0 - \nu_0)\psi]\,, \qquad (103) \\
\frac{d^2}{d\epsilon^2}|_{\epsilon=0} H_f[g_\epsilon; Q, P] =& -\nu_0'' - E_P[(f^*)''(g_0 - \nu_0)(\psi - \nu_0')^2] + E_P[(f^*)'(g_0 - \nu_0)]\nu_0'' \\
=& -E_P[(f^*)''(g_0 - \nu_0)]\,\mathrm{Var}_{P_0}[\psi]\,, \qquad (104)
\end{aligned}$$

where we used Eq. (101) and Eq. (102) to simplify and

$$\nu_0 = \mathrm{argmin}_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g_0 - \nu)]\}\,.$$

In particular, the second derivative is strictly negative when $\mathrm{Var}_{P_0}[\psi] \neq 0$, i.e., $H_f[g; Q, P]$ is strictly concave at g_0 in all directions, ψ, of nonzero variance under P_0. This can be made more explicit in the KL case. First recall the objective functional from Eq. (13),

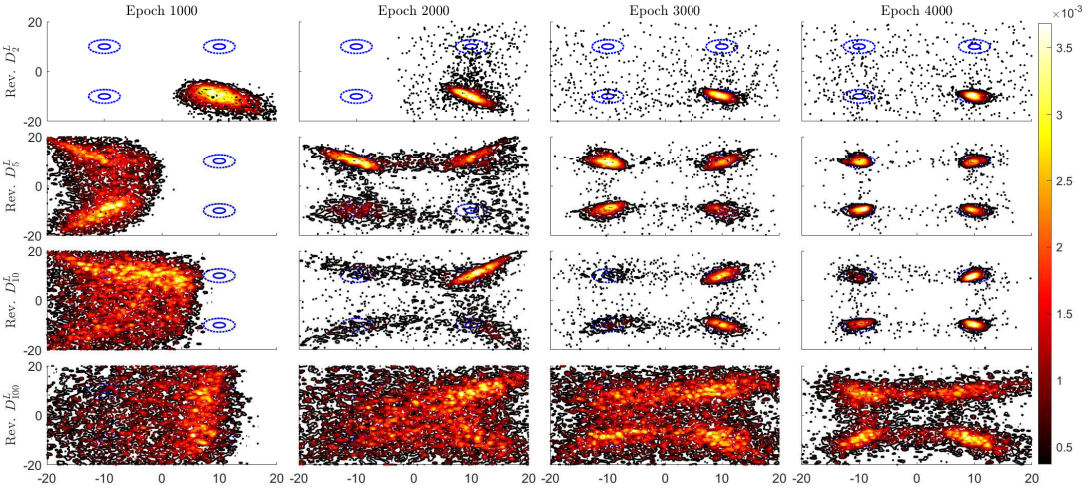
$$H_{KL}[g; Q, P] \equiv E_Q[g] - \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f_{KL}^*(g - \nu)]\} = E_Q[g] - \log E_P[e^g]\,. \tag{105}$$

Fixing $g_0 \in \Gamma$ and perturbing in a direction ψ we can compute

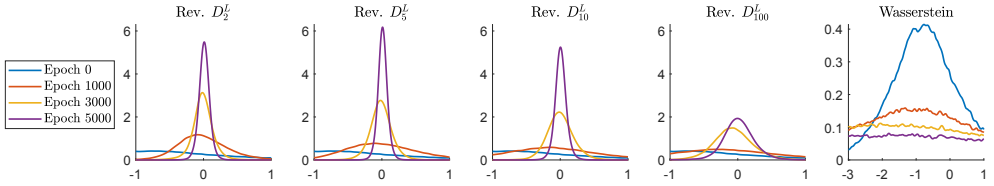
$$\begin{aligned}
\frac{d^2}{d\epsilon^2}|_{\epsilon=0} H_{KL}[g_0 + \epsilon\psi; Q, P] =& -\left(E_P[\psi^2 e^{g_0}]E_P[e^{g_0}]^{-1} - E_P[\psi e^{g_0}]^2 E_P[e^{g_0}]^{-2}\right) \qquad (106) \\
=& -\mathrm{Var}_{P_0}[\psi], \quad dP_0 \equiv e^{g_0} dP / E_P[e^{g_0}]\,.
\end{aligned}$$

Thus we again have strict convexity in all directions ψ of nonzero variance under P_0.

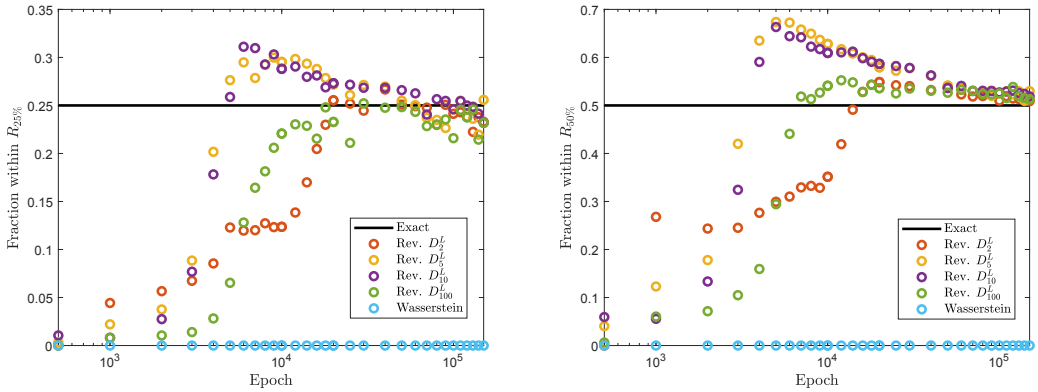
Appendix E. Additional Figures



(a)



(b)



(c)

Figure 5: Here we present generator samples and their statistical behavior from Wasserstein and reverse Lipschitz α-GAN methods using the same setup as in Figure 5, except that training was done with a much larger set of samples (100000 samples). We obtain similar results, with the primary difference being that the training converges faster.

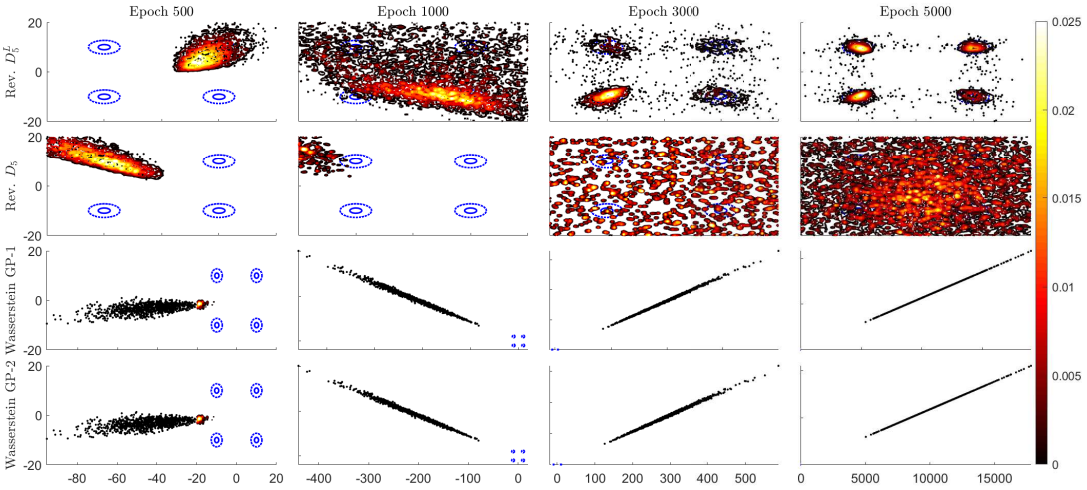


Figure 6: Here we present generator samples from WGAN-GP, reverse classical f-GAN (denoted D_α), and reverse Lipschitz α-GAN using the setup described in Section 6.2 and Figure 5, except that we do not embed in higher-dimensional space; the results are similar to what was described in Section 6.2. In the case of classical f-GAN this is intriguing since, unlike in Figure 5, here we have $D_f(P_\theta \| Q) < \infty$ yet the classical f-GAN still fails to converge. This suggests that the Lipschitz constraint aids in the stability of the training even in such cases where the classical f-divergence is finite. We show the result only for $\alpha = 5$ but the behavior for other values of α is similar.

References

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 (Methodological)*, 28(1): 131–142, 1966.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

R. Atar, K. Chowdhary, and P. Dupuis. Robust bounds on risk-sensitive functionals via Rényi divergence. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):18–33, 2015.

M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

A. Ben-Tal and M. Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.

P. Billingsley. *Probability and Measure*. Wiley, 2012.

P. Billingsley. *Convergence of Probability Measures*. Wiley, 2013.

J. Birrell, M. A. Katsoulakis, and Y. Pantazis. Optimizing variational representations of divergences and accelerating their statistical estimation. *arXiv preprint*, 2006.08781, 2020.

J. Birrell, P. Dupuis, M. A. Katsoulakis, L. Rey-Bellet, and J. Wang. Variational representations and neural network estimation for Rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116, 2021.

R. Bot, S. Grad, and G. Wanka. *Duality in Vector Optimization*. Springer Berlin Heidelberg, 2009.

T. Breuer and I. Csiszár. Measuring distribution model risk. *Mathematical Finance*, 26(2): 395–411, 2016.

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

K. Chowdhary and P. Dupuis. Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification. *ESAIM: Mathematical Modelling and Numerical Analysis*, 47(3):635–662, 2013.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:303–314, 1989.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training GANs with optimism. In *Proceedings of the International Conference on Learning Representations*, 2018.

R. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 2014.

P. Dupuis and R. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley, 1997.

P. Dupuis and Y. Mao. Formulation and properties of a divergence used to compare probability measures without absolute continuity. *arXiv preprint*, 1911.07422, 2019.

P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and P. Plechac. Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):80–111, 2016.

P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet. Sensitivity analysis for rare events based on Rényi divergence. *Annals of Applied Probability*, 30(4):1507–1533, 08 2020.

L. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

F. Farnia and D. Tse. A convex duality framework for GANs. In *Advances in Neural Information Processing Systems 31*, 2018.

G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley, 2013.

P. Glaser, M. Arbel, and A. Gretton. KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. *arXiv preprint*, 2106.08929, 2021.

P. Glasserman and X. Xu. Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58, 2014.

I. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint*, 1701.00160, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, 2014.

K. Gourgoulias, M. A. Katsoulakis, L. Rey-Bellet, and J. Wang. How biased is your model? Concentration inequalities, information and model bias. *IEEE Transactions on Information Theory*, 66(5):3079–3097, 2020.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30*, 2017.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems 30*, 2017.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

H. Husain, R. Nock, and R. C. Williamson. A primal-dual link between GANs and autoencoders. In *Advances in Neural Information Processing Systems 32*, 2019.

K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, et al. Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *Advances in Neural Information Processing Systems 28*, 2015.

P. Kidger and T. Lyons. Universal approximation with deep narrow networks. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint*, 1412.6980, 2014.

A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. Wasserman. Nonparametric estimation of Rényi divergence and friends. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

H. Lam. Robust sensitivity analysis for stochastic systems. *Mathematics of Operations Research*, 41(4):1248–1275, 2016.

F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

S. Liu and K. Chaudhuri. The inductive bias of restricted f-GANs. *arXiv preprint*, 1809.04542, 2018.

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems 30*, 2017.

D. Luenberger. *Optimization by Vector Space Methods*. Wiley, 1997.

T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

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

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems 29*, 2016.

Y. Pantazis, D. Paul, M. Fasoulakis, Y. Stylianou, and M. A. Katsoulakis. Cumulant GAN. *arXiv*, 2006.06625, 2020.

A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999.

A. Roberts and D. Varberg. *Convex Functions*. Elsevier Science, 1974.

R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

A. Ruderman, M. D. Reid, D. García-García, and J. Petterson. Tighter variational representations of f-divergences via restriction to probability measures. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, 2012.

W. Rudin. *Functional Analysis*. McGraw-Hill, 2006.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems 29*, 2016.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Springer International Publishing, 2015.

J. Song and S. Ermon. Bridging the gap between f-GANs and Wasserstein GANs. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, and G. R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. *arXiv preprint*, 0901.2698.

B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12 (70):2389–2410, 2011.

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.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer New York, 2008.

S. S. Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

C. Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

Q. Wang, S. R. Kulkarni, and S. Verdu. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9): 3064–3074, 2005.

Q. Wang, S. R. Kulkarni, and S. Verdu. A nearest-neighbor approach to estimating divergence between continuous random vectors. In *2006 IEEE International Symposium on Information Theory*, 2006.