Optimizing Variational Representations of Divergences and Accelerating their Statistical Estimation

Jeremiah Birrell, Markos A. Katsoulakis and Yannis Pantazis

Abstract

Variational representations of divergences and distances between high-dimensional probability distributions offer significant theoretical insights and practical advantages in numerous research areas. Recently, they have gained popularity in machine learning as a tractable and scalable approach for training probabilistic models and for statistically differentiating between data distributions. Their advantages include: 1) They can be estimated from data as statistical averages. 2) Such representations can leverage the ability of neural networks to efficiently approximate optimal solutions in function spaces. However, a systematic and practical approach to improving the tightness of such variational formulas, and accordingly accelerate statistical learning and estimation from data, is currently lacking. Here we develop such a methodology for building new, tighter variational representations of divergences. Our approach relies on improved objective functionals constructed via an auxiliary optimization problem. Furthermore, the calculation of the functional Hessian of objective functionals unveils the local curvature differences around the common optimal variational solution; this quantifies and orders the tightness gains between different variational representations. Finally, numerical simulations utilizing neural network optimization demonstrate that tighter representations can result in significantly faster learning and more accurate estimation of divergences in both synthetic and real datasets (of more than 1000 dimensions), often accelerated by nearly an order of magnitude.

J. Birrell and M.A. Katsoulakis are with the Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003, USA (email: birrell@math.umass.edu, markos@math.umass.edu).

Y. Pantazis is with the Institute of Applied and Computational Mathematics, Foundation for Research and Technology, Hellas, Heraklion, GR-70013, Greece (email: pantazis@iacm.forth.gr).

Index Terms

Divergences, Variational Representations, Statistical Estimation, Neural Networks, Hellinger-MINE

I. Introduction

Divergences and distances between multivariate probability distributions play a central role in many mathematical, engineering, and scientific fields ranging from statistical physics, large deviations theory, and uncertainty quantification to information theory, statistics, and machine learning. Variational representation formulas for divergences, also referred to as dual formulations, convert divergence calculation into an optimization problem over a function space and offer a valuable mathematical tool to build, train, and analyze probabilistic models and measure the similarity between data collections. Typical examples of variational representations are, among others, the Legendre transformation (LT) of an f-divergence [1], [2], the Donsker-Varadhan (DV) formula for the Kullback-Leibler (KL) divergence [3], [4] and the Rubinstein-Kantorovich duality formula for Wasserstein distance [5]. Variational representations have been used in statistical mechanics and interacting particles systems [6], large deviations [4], divergence estimation [7], [8], [9], determining variable independence through mutual information estimation [10], adversarial learning of generative models [11], [12], [13], uncertainty quantification (UQ) of stochastic processes [14], [15], bounding risk in probably approximately correct (PAC) learning [16], [17], [18], as well as in parameter estimation [19].

Two main mathematical ingredients are involved in the construction of a variational formula. First, the *function space* where the optimal solution will be searched for and, second, the representation expression, called here the '*objective functional*', whose optimization leads to the value of the divergence. Crucial practical advantages of variational formulas in statistics and machine learning include: a) they do not require an explicit form of the probability distributions (or their density ratio); related probabilistic quantities can be approximated by statistical estimators over available data; b) they can exploit the capacity of rich regression models such as neural networks to efficiently search the function space for optimal solutions; the optimal solution is typically related to the density ratio.

A single divergence can be derived from several different objective functionals. The key contribution of this paper is a systematic methodology that uses families of transformations (e.g., shifts, affine, and powers) to build new, tighter variational representations for divergences

by creating improved objective functionals, as described in our main Theorem 1. This idea is both simple and powerful; it provides a general framework that unifies many of the previous variational formulas in the literature, reveals new connections between them, drives the derivation of new variational formulas, and has practical implications in terms of accelerated statistical training, learning, and estimation from data.

Striking consequences of the proposed framework include: (i) the connection between LT-based KL, the DV representation formula, and a new, improved DV-type formula, (ii) a concrete representation of the abstract objective functional in [9], and (iii) a derivation of new representation formulas for α-divergence and connections with a recently derived, DV-type variational representation of Rényi divergences. Moreover, when the optimization over the transformation family is not analytically tractable, a second-order approximation is employed resulting in new variational representations.

The improved objective functionals constructed via our framework have the same optimal solution, but they are tighter in the sense that the same approximation of the optimum will provide a better approximation of the divergence, i.e., they are flatter around the optimal solution. We employ (functional) Hessians of the objective functionals to quantify and order relative tightness gains between different variational representations of divergences, in terms of the local curvature around the optimal solution. Although our goal here is not variance reduction, we do also study the asymptotic variance of the tighter variational representations. We obtain theoretical evidence that our optimization procedures do not increase the variance and provide some numerical evidence that the variance as well as the averaged error can be decreased.

Finally, we demonstrate that these tighter representation formulas can accelerate numerical optimization and estimation of divergences in a series of synthetic and real examples, such as the statistical estimation of f-divergences and mutual information, including cases with real and/or high-dimensional data (in excess of 1000 dimensions). Similarly to [10], we parameterize the function space using neural networks, hence the (parameter) optimization is efficiently performed with back-propagation algorithms. As an example of our method, we develop neural-based estimators of controlled sample complexity for Hellinger-based mutual information (Hellinger-MINE). Overall, we find that the improved, tighter representation formulas converge several times faster than the initial representation formula, often by nearly an order of magnitude in high-dimensional problems.

II. TIGHTENING THE VARIATIONAL REPRESENTATION OF f-DIVERGENCE

Background

Define $\mathcal{F}_1(a, b)$ to be the set of all convex functions $f : (a, b) \to \mathbb{R}$ with $f(1) = 0$. If a (resp. b) is finite, we extend f to a (resp. b) by continuity and set $f(x) = \infty$ for $x \notin [a, b]$. Such functions are appropriate for defining f-divergences, D_f, which have the variational characterization

$$D_f(Q\|P) \equiv E_P[f(dQ/dP)]$$
$$= \sup_{\phi \in \mathcal{M}_b(\Omega)} \{E_Q[\phi] - E_P[f^*(\phi)]\}, \tag{1}$$

where $\mathcal{M}_b(\Omega)$ denotes the set of all bounded measurable functions and f^* is the Legendre transform of f [20], [8]. Under appropriate assumptions (see Theorem 4.4 in [20]), the maximum is achieved at

$$\phi^* = f'(dQ/dP). \tag{2}$$

There are cases, such as α-divergence with $\alpha \in (0, 1)$, where $D_f(Q\|P)$ can be given a meaningful finite value even if $Q \not\ll P$ (see [21]). However, the right hand side of (1) is always $+\infty$ if $Q \not\ll P$ and so here we use the convention that $D_f(Q\|P) \equiv \infty$ when $Q \not\ll P$. The special case of Kullback-Leibler (KL) divergence (i.e., D_f with $f(x) = x \log(x)$) has a well-known alternative variational representation, the Donsker-Varadhan (DV) variational formula

$$D_{KL}(Q\|P) = \sup_{\phi \in \mathcal{M}_b(\Omega)} \{E_Q[\phi] - \log E_P[e^\phi]\}. \tag{3}$$

It is known that the objective functional in (3) is tighter than that of (1), in the sense that $E_Q[\phi] - \log E_P[e^\phi] \geq E_Q[\phi] - E_P[f^*(\phi)]$ for all $\phi \in \mathcal{M}_b(\Omega)$, [9]. In this paper, we present a general procedure for obtaining tighter variational representations of any f-divergence, for which the transition from (1) to (3) is just one special case.

Theoretical Results

Our method for deriving tighter variational representations for f-divergences is described in the following Theorem; a proof can be found in Section VI.

Theorem 1: Let $f \in \mathcal{F}_1(a, b)$ and suppose $D_f(Q\|P) < \infty$. With ϕ^* defined by Eq. (2) (when it exists), let Φ be a family of functions (the test functions) with

$$\mathcal{M}_b(\Omega) \subset \Phi \subset L^1(Q) \quad \text{or with} \quad \phi^* \in \Phi \subset L^1(Q). \tag{4}$$

Consider any family of transformations

$$\mathcal{T} \subset \{T = T(\phi)\,,\ \text{such that}\ T : \Phi \mapsto L^1(Q)\} \tag{5}$$

that includes the identity map. Then

$$D_f(Q\|P) = \sup_{\phi \in \Phi} H_{\mathcal{T}}[\phi]\,, \tag{6}$$

$$\text{where} \quad H_{\mathcal{T},f}[\phi] = \sup_{T \in \mathcal{T}} \{E_Q[T(\phi)] - E_P[f^*(T(\phi))]\}\,, \tag{7}$$

and the maximum in (6) is achieved at ϕ^*. Furthermore, the objective functional, $H_{\mathcal{T},f}$, in the variational representation (6) is tighter than the objective functional in (1), in the sense that

$$E_Q[\phi] - E_P[f^*(\phi)] \leq H_{\mathcal{T},f}[\phi] \leq D_f(Q\|P) \tag{8}$$

for all $\phi \in \Phi$.

Remark 1: When the choice of f and/or $\mathcal{T}$ is unimportant or clear from the context, we will omit the corresponding subscript on $H_{\mathcal{T},f}$.

(i) The main new insight and primary mathematical tool in this paper are formulas (6) and (7), which allow for the objective functional $H_{\mathcal{T}}[\phi]$ to be 'improved/tightened' using any appropriate family of transformations $\mathcal{T}$; examples of such families are discussed in the next two subsections. Eq. (6) is a simple but far-reaching idea that reveals connections between many known variational representations and also leads to the derivation of new ones. (ii) The extension of Eq. (1) from $\mathcal{M}_b(\Omega)$ to $L^1(\Omega)$ is useful because the exact optimizer, ϕ^*, is generally unbounded; various versions of this extension can be found in the literature [20], [8]. This extension is needed to justify the computation of variational derivatives around the optimum presented in Section III. It also implies that one does not need to impose boundedness condition via a cutoff function when employing neural-based statistical estimation. (iii) The generalization of Eq. (1) to a family, Φ, that contains the optimizer is the natural next step; again, see [20], [8]. It provides a great deal of flexibility in adapting the proposed variational representation (6) to different f-divergences and could guide the algorithmic implementation. We use this idea several times to restrict the optimization to, e.g., positive functions for the α-divergences and finite dimensional submanifolds for exponential families of distributions. (iv) If $\mathcal{T} \subset \{T : \Phi \to \Phi\}$ is a group under composition then the objective functional (7) is invariant under the family of transformations $\mathcal{T}$, i.e., $H_{\mathcal{T}}[T(\phi)] = H_{\mathcal{T}}[\phi]$ for all $T \in \mathcal{T}$. (v) If $(\Omega, \mathcal{M})$ is a metric space with the Borel σ-algebra then one can replace $\mathcal{M}_b(\Omega)$ with $C_b(\Omega)$ (bounded continuous functions) and $L^1(Q)$ with $L^1_c(Q)$

(continuous $L^1(Q)$ functions) in Theorem 1. This is a direct consequence of Lusin's theorem [22, Appendix D]. (vi) The auxiliary optimization problem in Eq. (7) is often computed analytically; alternatively and due to its low dimensionality, the corresponding optimization can be easily incorporated into the gradient descent framework without significant additional computational cost.

Families of Transformations

Next, we present several useful families of transformations that, in conjunction with Theorem 1, yield tighter variational formulas.

a. Identity: $\mathcal{T}^{id} = \{id\}$ leads to what we call the Legendre Transform (LT) f-divergence variational formula given by (1).

b. Shifts: $T_\nu^{shift}(\phi) = \phi - \nu$, $\nu \in \mathbb{R}$, lead to what we call the shift or ν-improved variational formula

$$D_f(Q\|P) = \sup_{\phi \in \Phi}\{\sup_{\nu \in \mathbb{R}}\{E_Q[\phi] - \nu - E_P[f^*(\phi - \nu)]\}\}. \tag{9}$$

This result was first obtained in [23], and in this sense Theorem 1 is a broad and systematic generalization of Theorem 4.4 in [23], where only shift transformations where considered.

c. Scaling Transformation: $T_\eta^{scale}(\phi) = \eta\phi$, $\eta \in \mathbb{R}$, which lead to the scaling or η-improved variational formula

$$D_f(Q\|P) = \sup_{\phi \in \Phi}\{\sup_{\eta \in \mathbb{R}}\{\eta E_Q[\phi] - E_P[f^*(\eta\phi)]\}\}. \tag{10}$$

d. Affine Transformation: The above two cases can be combined into a two parameter family $T_{\eta,\nu}^{affine}(\phi) = \eta\phi - \nu$.

e. Power Transformation: $T_\beta^{power}(\phi) = \phi^\beta$, $\beta \in \mathbb{R}$, which lead to the power or β-improved variational formula

$$D_f(Q\|P) = \sup_{\phi \in \Phi}\{\sup_{\beta \in \mathbb{R}}\{E_Q[\phi^\beta] - E_P[f^*(\phi^\beta)]\}\}. \tag{11}$$

As with the affine transformations, the power transformations can be combined with the shift and/or scaling transformations to form a multiparameter family. These are related to the well-known Box-Cox transformation [24], used in statistics to transform non-normal data sets to approximately normal.

New Derivations of Existing Variational Representations

In this and next subsections, we explore several specific cases of the above framework, focusing primarily on examples where the optimization over the shift and/or scaling parameter can be done analytically. Here, we uncover connections with previously known variational formulas.

1) Donsker-Varadhan formula (KL-divergence with shift transformations): KL-divergence is the f-divergence corresponding to $f(x) = x \log(x)$, which has Legendre transform $f^*(y) = e^{y-1}$. The maximum over the shift transformations $\mathcal{T}^{shift}$ in (6) occurs at $\nu^* = \log\left(E_P[e^\phi]\right) - 1$, and hence

$$\sup_{\nu \in \mathbb{R}} \{E_Q[\phi] - \nu - E_P[f^*(\phi - \nu)]\} \tag{12}$$

$$= E_Q[\phi] - \nu^* - E_P[e^\phi]e^{-\nu^*-1} = E_Q[\phi] - \log(E_P[e^\phi]).$$

The result is the objective functional in the well-known Donsker-Varadhan variational formula (3) and so this framework provides the connection between (3) and (1). We also note that in [25] a connection is derived between (3) and (1) by using a logarithmic change of variables in the function space, based on (2) for the KL case.

2) Connection with the results of [9]: In Theorem 1 of [9] the following improved variational formula was derived:

$$D_f(Q\|P) = \sup_{\phi \in \mathcal{M}_b(\Omega)} \{E_Q[\phi] - (I_{f,P}^R)^*[\phi]\}, \tag{13}$$

$$I_{f,P}^R(r) \equiv \begin{cases} E_P[f(r)], & r \in L^1(P),\, r \geq 0,\, E_P[r] = 1 \\ \infty, & \text{otherwise} \end{cases}.$$

This is another special case of our framework, as can be seen by first rewriting the Legendre-Fenchel transform and then using Theorem 4.2 in [23]:

$$(I_{f,P}^R)^*[\phi] = \sup_{Q \ll P} \{E_Q[\phi] - D_f(Q\|P)\} \tag{14}$$

$$= \inf_{\nu \in \mathbb{R}} \{\nu + E_P[f^*(\phi - \nu)]\}.$$

Hence the variational formula (13) is in fact the same as the ν-improved variational formula (9).

3) χ^2-Divergence: We use Theorem 1 to provide a variational perspective on the classical Hammersley-Chapman-Robbins bound for the χ^2-divergence. To our knowledge, the tightness

result of the bound is novel. The χ^2-divergence is a special case of the α-divergences, $\chi^2(Q\|P) = 2D_{f_2}(Q\|P)$. Here we optimize (6) over $\mathcal{T}^{affine}$ to obtain

$$\chi^2(Q\|P) = \sup_{\phi \in \Phi} \frac{(E_Q[\phi] - E_P[\phi])^2}{\mathrm{Var}_P[\phi]}, \tag{15}$$

where $\Phi = \{\phi \in L^1(Q) \cap L^2(P) : \phi \geq 0, \mathrm{Var}_P[\phi] > 0\}$ and the maximum is achieved at $\phi^* = dQ/dP \geq 0$. Eq. (15) implies the Hammersley-Chapman-Robbins bound for the χ^2 divergence (see, e.g., Eq. 4.13 in [26]) and shows tightness over the set Φ. The objective functional in (15) was proposed as loss function for χ^2-GANs, [27]; thus, (15) provides a complete and rigorous justification for this choice. Finally, if we instead optimize (6) over $\mathcal{T}^{shift}$, we obtain the objective functional for χ^2 derived in [25], which is thus less tight than (15); see Appendix A.3.

Deriving New Variational Representations and Further Connections

In this subsection, we produce new variational representations and reveal further connections between divergences. We do not claim these examples are exhaustive. Nevertheless, they cover many important cases and illustrate the power and flexibility of Theorem 1.

1) Improved Donsker-Varadhan (KL-Divergence with affine transformations): Introducing a scaling parameter into Eq. (12), i.e., optimizing over all affine transformations $\mathcal{T}^{affine}$ in (6), one finds the new KL variational representation

$$D_{KL}(Q\|P) = \sup_{\phi \in \Phi}\{\sup_{\eta \in \mathbb{R}}\{\eta E_Q[\phi] - \log E_P[e^{\eta\phi}]\}\}. \tag{16}$$

The inclusions $\mathcal{T}^{shift} \subset \mathcal{T}^{affine}$ implies that (16) is tighter than DV (3). Calculations and numerical results that quantify this improved tightness are found in Section III. The optimization over η in Eq. (16) cannot be evaluated analytically in general, but it can be done numerically (as discussed in Section V).

2) α-Divergences (scaling transformations): The α-divergences are the family of f-divergences corresponding to $f_\alpha(t) = \frac{t^\alpha - 1}{\alpha(\alpha - 1)}$, $\alpha \neq 0, 1$. See [21] for properties, related families, and further references. It includes the KL, Hellinger and χ^2 divergences as special cases [28], is closely related to the Tsallis entropies [29], and appears also in the context of information

geometry [30]. By optimizing (6) over the family of scaling transformations, $\mathcal{T}^{scale}$ (restricted to $\eta > 0$), we obtain a new variational representation of the α-divergences:

$$D_{f_\alpha}(Q\|P) \tag{17}$$

$$= \sup_{\phi \in \Phi_\alpha} \left\{ \frac{1}{\alpha(\alpha-1)} \left(E_Q[\phi]^\alpha E_P[\phi^{\alpha/(\alpha-1)}]^{-(\alpha-1)} - 1 \right) \right\},$$

where $\Phi_\alpha = \{\phi \in L^1(Q) : \phi \geq 0, 0 < E_P[\phi^{\alpha/(\alpha-1)}] < \infty\}$ if $\alpha > 1$ and $\Phi_\alpha = \{\phi : \phi > 0\}$ if $0 < \alpha < 1$. Eq. (17) has the exact optimizers $\phi_\alpha^* = (dQ/dP)^{\alpha-1}$. Theorem 1 guarantees that the objective functionals in the new variational representations (17) are tighter than that of the LT f-divergence objective functional from (1). See Appendix A for details on the calculations that lead to Eq. (17), as well as for connections to the KL divergence in the limits as $\alpha \to 0, 1$.

3) Variational representations of Rényi divergences: Eq. (17) leads to a variational characterization of Rényi divergences. Using the known connection between the α and Rényi divergences, along with Eq. (17) and the change of variables $\phi = e^{(\alpha-1)g}$ (see Appendix A for details) one obtains

$$R_\alpha(Q\|P) = \frac{1}{\alpha(\alpha-1)} \log(\alpha(\alpha-1)D_{f_\alpha}(Q\|P) + 1) \tag{18}$$

$$= \sup_g \left\{ \frac{1}{\alpha-1} \log(E_Q[e^{(\alpha-1)g}]) - \frac{1}{\alpha} \log(E_P[e^{\alpha g}]) \right\}.$$

This constitutes an independent derivation of the Renyi variational formula derived in [31], [25], while in the asymptotic limit $\alpha \to 1$ one recovers (3). The Renyi variational formula (18) was also used in [32] to construct cumulant-based, generative adversarial networks. Moreover, UQ bounds for risk-sensitive functionals in terms of Rényi divergences, which were obtained recently in [33], readily follow from (18) after appropriate manipulations and an optimization over $\alpha > 1$.

4) α-Divergences (scaling and power transformations): For $\alpha \in (0,1)$, Combining the scaling and power family of transformations yields

$$D_{f_\alpha}(Q\|P) = \frac{1}{\alpha(1-\alpha)} \sup_{\phi>0} \sup_{\beta \in \mathbb{R}} \left\{ 1 - E_Q[\phi^\beta]^\alpha E_P[\phi^{-\alpha\beta/(1-\alpha)}]^{1-\alpha} \right\}. \tag{19}$$

The optimization over scalings was evaluated as in Eq. (17) but the optimization over the power transformations, β, cannot be done analytically. In practice, β can be included as an additional parameter in a numerical optimization procedure; see Section V for further discussion.

5) Exponential Families: If P and Q are members of a parametric family then the set of test functions, Φ, in (6) can be reduced to a finite dimensional manifold. For instance, if $P = P_{\theta_p}$ and $Q = P_{\theta_q}$ are members of the same exponential family $dP_\theta = h(x)e^{\kappa(\theta) \cdot T(x) - \beta(\theta)}, \ \theta \in \Theta$ with $T : \Omega \to \mathbb{R}^n$ the vector of sufficient statistics then the explicit optimizer $\phi^* = f'(dQ/dP)$ lies on an $(n+1)$-dimensional manifold of functions, parameterized by the sufficient statistics and constants, $\phi_{(\beta,\kappa)} = f'(\exp(\kappa \cdot T + \beta))$, $(\beta, \kappa) \in \mathbb{R} \times \mathbb{R}^n$, and computation of the f-divergence reduces to the following finite-dimensional optimization problem:

$$D_f(Q\|P) = \sup_{(\beta,\kappa) \in \mathbb{R}^{n+1}} \left\{ E_Q[\phi_{(\beta,\kappa)}] - E_P[f^*(\phi_{(\beta,\kappa)})] \right\}. \tag{20}$$

This variational representation can be further combined with any appropriate family of transformations $\mathcal{T}$. We refer to Appendix E for further details.

6) Approximating the Improved Donsker-Varadhan: An analytic computation of the optimization over η in Eq. (6) is not possible in general. Nevertheless, an alternative variational characterization of the KL divergence can be derived by expanding around $\eta = 1$ and solving the quadratic approximation for $\Delta\eta^*$. Under appropriate assumptions given in Theorem 6 (see Appendix B), we derive a new variational formula for the KL divergence,

$$D_{KL}(Q\|P) = \sup_{\phi \in \Phi}\{(1 + \Delta\eta^*(\phi))E_Q[\phi] - \log E_P[e^{(1+\Delta\eta^*(\phi))\phi}]\}, \tag{21}$$

where the optimal $\Delta\eta^*$ is obtained by maximizing the second-order Taylor approximation and it is given by

$$\Delta\eta^*(\phi) = \frac{E_Q[\phi] - E_{P_\phi}[\phi]}{\mathrm{Var}_{P_\phi}[\phi]}, \tag{22}$$

with $dP_\phi = e^\phi dP/E_P[e^\phi]$ being the tilted measure. Since we quadratically approximate the improved Donsker-Varadhan, there is no guarantee that the objective functional in Eq. (21) is tighter than that of Donsker-Varadhan's representation. Despite no general guarantee, it is expected to be tighter when $\Delta\eta^*(\phi)$ is sufficiently small. This example demonstrates that even when there is no explicit formula for the transformation's optimization, one is still able to derive a closed-form formula for an approximate version of it and thereby obtain a new, rigorous variational formula. This same methodological approach can be used to analytically approximate the optimization over other families of transformations. As another demonstration, we refer to Theorem 7 in Appendix C where a new approximate variational formula is derived for the power-improved Rényi variational representation.

7) Connections with Uncertainty Quantification: The improved DV representation (16) provides an alternative and arguably more general derivation of model uncertainty bounds derived recently in [14], [15]. These results quantify the effects of model uncertainty by bounding expectations of an observable ϕ under an alternative model, Q, in terms of the behavior under a baseline model P and the model discrepancy, measured by $D_{KL}(Q\|P)$. Specifically, (16) implies after straightforward manipulation that $E_Q[\phi] \leq \inf_{\eta>0}\{\frac{1}{\eta}\log E_P[e^{\eta\phi}] + \frac{1}{\eta}D_{KL}(Q\|P)\}$ [14]. More generally, we can obtain similar UQ bounds when model discrepancy is measured by an f-divergence by using $T_{\eta,\nu}^{affine}(\phi) = \eta\phi - \nu$ in (6) and performing the analogous manipulations:

$$E_Q[\phi] \leq \inf_{\substack{\eta>0 \\ \nu\in\mathbb{R}}} \left\{ \frac{1}{\eta}\{E_P[f^*(\eta\phi - \nu)] + \nu\} + \frac{1}{\eta}D_f(Q\|P) \right\}. \tag{23}$$

The Hammersley-Chapman-Robbins bound can also be viewed as a special case of (23) in this UQ context.

III. VARIATIONAL DERIVATIVES AND TIGHTNESS GAINS

In Theorem 1, we established the general methodology for building tighter variational representations of f-divergences, by constructing suitable objective functionals $H_{\mathcal{T}}$. Here we will quantify relative tightness gains corresponding to different transformation families $\mathcal{T}$: for all such families the maximizer in (6) is always ϕ^* given by (2). Therefore, our approach relies on building quadratic variational approximations of each objective functional (7) around the common maximizer ϕ^*, and subsequently comparing the corresponding (variational) curvatures; see Figure 1 for a demonstration. Specifically, using that the maximum occurs at ϕ^*, an asymptotic expansion yields

$$H_{\mathcal{T}}[\phi^* + \psi] = D_f(Q\|P) + \frac{1}{2}\langle\nabla^2 H_{\mathcal{T}}[\phi^*]\psi, \psi\rangle + O(\|\psi\|^3), \tag{24}$$

where we formally define $\langle\nabla^2 H_{\mathcal{T}}[\phi]\psi, \psi\rangle \equiv \frac{d^2}{d\epsilon^2}\big|_{\epsilon=0} H_{\mathcal{T}}[\phi + \epsilon\psi]$ and ψ is any functional perturbation of the maximizer ϕ^*. Formally, the second order term $\nabla^2 H_{\mathcal{T}}[\phi^*]$, i.e., a variational Hessian, is necessarily non-positive and determines the behavior in a neighborhood of the maximizer. By comparing $\nabla^2 H_{\mathcal{T}}[\phi^*]$ for different families $\mathcal{T}$, we will quantify the 'tightness gains' provided by different transformation families. All these calculations can be made rigorous under appropriate assumptions as demonstrated next in Theorem 2 and in Theorem 8 in Appendix D.

Here we focus our analysis on affine transformations, $T_{\eta,\nu}^{affine}(\phi) = \eta\phi - \nu$, but a similar analysis can be performed for any family $\mathcal{T}$ with a smooth, finite-dimensional parameterization.

The LT f-divergence variational formula (1) corresponds to $\mathcal{T}$ containing only the identity, and we write the corresponding objective functional as $H_{id}[\phi]$. Specializing (7) to the affine case, we define the functional

$$
\begin{aligned}
H[\phi, \eta, \nu] =& E_Q[T_{\eta,\nu}^{affine}(\phi)] - E_P[f^*(T_{\eta,\nu}^{affine}(\phi))] \\
=& E_Q[\eta\phi - \nu] - E_P[f^*(\eta\phi - \nu)],
\end{aligned}
\tag{25}
$$

which leads to four different objective functionals and variational representations of the f-divergence

$$
D_f(Q\|P) = \sup_{\phi} \overbrace{H[\phi, 1, 0]}^{H_{id}[\phi]} = \sup_{\phi}\sup_{\nu} \overbrace{H[\phi, 1, \nu]}^{H_{shift}[\phi]}
\tag{26}
$$

$$
= \sup_{\phi}\underbrace{\sup_{\eta} H[\phi, \eta, 0]}_{H_{scale}[\phi]} = \sup_{\phi}\underbrace{\sup_{\eta,\nu} H[\phi, \eta, \nu]}_{H_{affine}[\phi]}
$$

and the corresponding Hessians $\nabla^2 H_{id}[\phi]$, $\nabla^2 H_{shift}[\phi]$, $\nabla^2 H_{scale}[\phi]$, and $\nabla^2 H_{affine}[\phi]$. Next we state a Theorem where we evaluate and compare these variational Hessians for the important case of KL divergence, where $f(x) = x\log(x)$ and $\phi^* = f'(dQ/dP) = \log(dQ/dP) + 1$. Detailed computations and rigorous analysis for general f-divergences can be found in Appendix D.
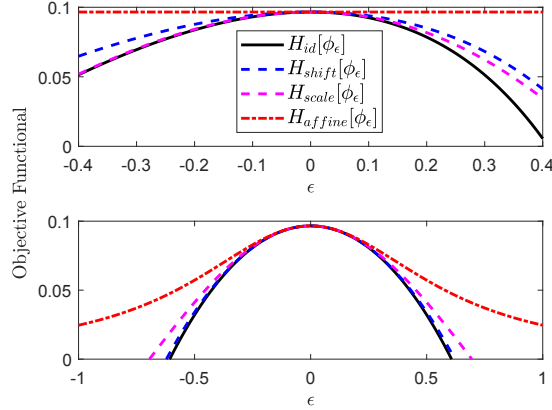

Fig. 1. Both plots demonstrate the improvement of the KL divergence objective functional in a neighborhood of the optimizer. Here, $\phi_\epsilon = \phi^* + \epsilon\psi$ where $\psi = x^2$ (top panel) and $\psi(x) = x$ (bottom panel). P and Q are 1-dimensional Gaussians. Black curves: LT-based f-divergence objective functional, Blue curves: shift-improved (i.e., Donkser-Varadhan), Magenta curves: η-improved, Red curves: shift-scaling-improved. Note that $\psi = x^2$ is related to ϕ^* by a shift and scaling, hence the shift-scaling-improved objective functional in the top plot has zero curvature in this direction, a manifestation of its shift and scale invariance.

Motivated by the formal calculation (24), we define

$$\mathcal{J}_{\mathcal{T}}(\epsilon) = H_{\mathcal{T}}[\phi^* + \epsilon\psi]\,, \quad \text{for any} \quad \psi \in \mathcal{M}_b(\Omega) \quad \text{and} \quad \epsilon \in \mathbb{R}\,. \tag{27}$$

The second derivative, when it exists, $\frac{d^2}{d\epsilon^2}\mathcal{J}_{\mathcal{T}}(0) = \frac{d^2}{d\epsilon^2}\big|_{\epsilon=0}H_{\mathcal{T}}[\phi^* + \epsilon\psi]$ is called a 2nd Gateaux derivative of $H_{\mathcal{T}}[\phi]$ in the direction ψ and describes the curvature of the objective functional $H_{\mathcal{T}}[\phi]$ along any given direction $\psi \in \mathcal{M}_b(\Omega)$ at the maximizer ϕ^*. Therefore we will define a corresponding variational Hessian in the direction ψ as

$$\langle \nabla^2 H_{\mathcal{T}}[\phi]\psi, \psi \rangle \equiv \frac{d^2}{d\epsilon^2}\mathcal{J}_{\mathcal{T}}(0) = \frac{d^2}{d\epsilon^2}\big|_{\epsilon=0}H_{\mathcal{T}}[\phi + \epsilon\psi]\,. \tag{28}$$

We remark that relation (28) is simply a notation that only intends to draw a parallel to finite dimensional optimization calculations using conventional Hessians at maxima or minima of finite dimensional functions. All rigorous results are stated in terms of the 2nd Gateaux derivative $\frac{d^2}{d\epsilon^2}\mathcal{J}_{\mathcal{T}}(0)$. Using the above notations we can formulate to following result for the KL divergence.

Theorem 2 (Tightness gains for KL divergence): Consider the KL divergence, i.e., $f(x) = x\log(x)$. In addition to the assumptions of Theorem 1, we assume that the maximizer $\phi^* = \log(dQ/dP) + 1$ in (2) is bounded (i.e., $\phi^* \in \mathcal{M}_b(\Omega)$) and select the function space $\Phi = \mathcal{M}_b(\Omega)$ in Theorem 1. Then for $\mathcal{T} = \mathcal{T}^{id}, \mathcal{T}^{shift}, \mathcal{T}^{affine}$ the function $\mathcal{J}_{\mathcal{T}}$ in (27) is twice differentiable at $\epsilon = 0$. Furthermore, using the notation (28), the corresponding 2nd Gateaux derivatives $\frac{d^2}{d\epsilon^2}\mathcal{J}_{\mathcal{T}}(0) = \frac{d^2}{d\epsilon^2}\big|_{\epsilon=0}H_{\mathcal{T}}[\phi^*+\epsilon\psi]$ for all $\mathcal{T} = \mathcal{T}^{id}, \mathcal{T}^{shift}, \mathcal{T}^{affine}$ satisfy the following:

$$\frac{d^2}{d\epsilon^2}\mathcal{J}_{\mathcal{T}^{id}}(0) = \langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle = -\operatorname{Var}_Q[\psi] - E_Q[\psi]^2\,, \tag{29}$$

$$\frac{d^2}{d\epsilon^2}\mathcal{J}_{\mathcal{T}^{shift}}(0) = \langle \nabla^2 H_{shift}[\phi^*]\psi, \psi \rangle = -\operatorname{Var}_Q[\psi]\,, \tag{30}$$

$$\frac{d^2}{d\epsilon^2}\mathcal{J}_{\mathcal{T}^{affine}}(0) = \langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle = -\operatorname{Var}_Q[\psi] + \frac{\operatorname{Cov}_Q(\phi^*, \psi)^2}{\operatorname{Var}_Q[\phi^*]}\,, \tag{31}$$

corresponding to (1), (3), and (16), respectively.

The complete proof of Theorem 2 is presented in Appendix D in the form of Theorem 8 that describes the more general case of f-divergences.

Remark 2: The gains inherent in the inclusions $\mathcal{T}^{id} \subset \mathcal{T}^{shift} \subset \mathcal{T}^{affine}$ in Theorem 1 are quantified in Theorem 2 by comparing the variational curvatures Eq. (29), Eq. (30), and Eq. (31) as computed by these 2nd Gateaux derivatives; note that they are progressively smaller in magnitude. These curvature computations demonstrate how one can rigorously and precisely quantify heuristics such as those presented in Figure 1 from [9]. Furthermore, our Hessian

computations in the form of Theorem 8 also quantify and extend to f-divergences the accuracy gains observed in the neural estimation of mutual information in [10].

Remark 3: The boundedness assumptions of Theorems 2 and 8 may appear restrictive compared to the generality assumed for f-divergence definition, but they provide the simplest conditions under which Theorems can be rigorously stated. Under appropriate technical assumptions these results can be easily generalized without assuming the boundedness of ϕ^* and for more general function spaces Φ. Any necessary assumptions need to ensure the applicability of the implicit function theorem and the dominated convergence theorem [34] (Theorem 2.27), following their use in the proof of Theorem 8 presented in Appendix D.

Figure 1 is a simple demonstration of Theorem 2, using 1D Gaussians: $P = N(0, 1)$, $Q = N(0, 1/2)$, with perturbations in the directions $\psi = x^2$ (top) and $\psi = x$ (bottom). Optimizing over all affine transformations (red curves) provides noticeable curvature gains when compared to optimization over only shifts (blue curves), i.e., the improved DV proposed in (16) compared to the classical DV objective functional (3) and even more so compared to the Legendre transform case, (1) (black curves).

IV. f-DIVERGENCE ESTIMATOR BIAS AND VARIANCE

The variational formula Eq. (6) suggests the following natural f-divergence estimator

$$\widehat{D}_f^n(Q\|P) = \sup_{\phi \in \widetilde{\Phi}} \sup_{T \in \mathcal{T}} \{E_{Q_n}[T(\phi)] - E_{P_n}[f^*(T(\phi))]\}, \tag{32}$$

where Q_n and P_n are the n-sample empirical measures and $\widetilde{\Phi} \subset \Phi$ is a function space that can be optimized over numerically (e.g., a family of neural networks). A natural question is therefore the bias and variance of this estimator. In practice, the optimizations are performed via some stochastic gradient descent (SGD) algorithm, and one is actually interested in the bias and variance after a finite number of training steps. Addressing this complicated problem, which depends heavily on the choice of space $\widetilde{\Phi}$ and the SGD algorithm, is outside the scope of the present work. However, in this section we will follow the prior work in [35] for KL divergences and discuss the bias and variance of the objective functional in (32). Finally, we emphasize that our goal of this paper is to develop tighter objective functionals and study the impact of improved curvature on the speed of convergence of numerical estimators. This is a separate question from that of variance reduction, which we do not pursue here. However, we will show that, for α-divergences, optimizing over scalings does not worsen the variance at the optimizer.

Objective Functional Bias

The estimator (32) can be viewed either as a single-stage optimization problem over both ϕ and T, in which case the objective functional is unbiased, or as a two-stage optimization,

$$\widehat{D}_f^n(Q\|P) = \sup_{\phi \in \widetilde{\Phi}} \widehat{H}_{\mathcal{T},f}[\phi; Q_n, P_n], \tag{33}$$

$$\widehat{H}_{\mathcal{T},f}[\phi; Q_n, P_n] = \sup_{T \in \mathcal{T}}\{E_{Q_n}[T(\phi)] - E_{P_n}[f^*(T(\phi))]\},$$

with a biased objective functional. If the optimization over T must be performed numerically then we take the former view, but if part of the optimization can be performed analytically (such as in (12) and (17)) then we take the latter view. In any case, the divergence estimator as a whole, Eq. (32), is not unbiased. This can be seen in the simple case of discrete measures on a finite sample space, where one can numerically optimize over all of $\mathcal{M}_b(\Omega)$,

$$\widehat{D}_f^n(Q\|P) = \sup_{\phi \in \mathcal{M}_b(\Omega)} \{E_{Q_n}[\phi] - E_{P_n}[f^*(\phi)]\} = D_f(Q_n\|P_n)$$

$$= E_{P_n}[f(dQ_n/dP_n)], \tag{34}$$

which is biased. This fact renders the biased/unbiased objective functional question less relevant in practice, as one's goal is generally to estimate the divergence and not just the objective functional at a fixed ϕ.

Objective Functional Variance

For general f, the objective functional estimator for the LT variational formula is

$$\widehat{H}_f[\phi; Q_n, P_n] = E_{Q_n}[\phi] - E_{P_n}[f^*(\phi)]. \tag{35}$$

The two terms are independent, therefore

$$\mathrm{Var}[\widehat{H}_f[\phi; Q_n, P_n]] = \mathrm{Var}[E_{Q_n}[\phi]] + \mathrm{Var}[E_{P_n}[f^*(\phi)]] \tag{36}$$

$$= \frac{1}{n}\mathrm{Var}_Q[\phi] + \frac{1}{n}\mathrm{Var}_P[f^*(\phi)].$$

In particular, for α divergences (36) reduces to

$$n\,\mathrm{Var}[\widehat{H}_{f_\alpha}[\phi; Q_n, P_n]] = \mathrm{Var}_Q[\phi] + \alpha^{-2}|\alpha - 1|^{2\alpha/(\alpha-1)}\mathrm{Var}_P[\phi^{\alpha/(\alpha-1)}], \tag{37}$$

where for $\alpha \in (0, 1)$ we made a change of variables $\phi \to -\phi$.

To compute the asymptotic variance of an optimized objective functional, we rely on the delta method. This method can be applied to any objective functional that can be expressed as a

function of expectations. We provide the details in the case of scaling-optimized α-divergences Eq. (17).

Theorem 3: Let $\alpha > 0$, $\alpha \neq 1$ and suppose $\phi \in \Phi$ satisfies $0 < c \leq \phi \leq d$ for some $c, d \in \mathbb{R}$. Then we have:

(a) The scaling-optimized α-divergence objective functional (17),

$$\widehat{H}_{scale,f_\alpha}[\phi; Q_n, P_n] = \frac{1}{\alpha(\alpha-1)} \left(E_{Q_n}[\phi]^\alpha E_{P_n}[\phi^{\alpha/(\alpha-1)}]^{-(\alpha-1)} - 1 \right), \tag{38}$$

has asymptotic variance

$$\lim_{n\to\infty} n \operatorname{Var}[\widehat{H}_{scale,f_\alpha}[\phi; Q_n, P_n]] \tag{39}$$

$$= (\alpha-1)^{-2} (E_Q[\phi]/E_P[\psi_\alpha])^{2(\alpha-1)} \operatorname{Var}_Q[\phi] + \alpha^{-2} (E_Q[\phi]/E_P[\psi_\alpha])^{2\alpha} \operatorname{Var}_P[\psi_\alpha]),$$

where $\psi_\alpha \equiv \phi^{\alpha/(\alpha-1)}$.

(b) At the exact optimizer, $\phi_\alpha^* = (dQ/dP)^{\alpha-1}$, Eq. (39) reduces to

$$\lim_{n\to\infty} n \operatorname{Var}[H_{scale,f_\alpha}[\phi^*; Q_n, P_n]] \tag{40}$$

$$= |\alpha-1|^{-2} \operatorname{Var}_Q[(dQ/dP)^{\alpha-1}] + \alpha^{-2} \operatorname{Var}_P[(dQ/dP)^\alpha].$$

See Appendix F for the proof.

Theorem 3 implies that we need $(dQ/dP)^{2\alpha} \in L^1(P)$ to achieve a finite variance; this fact can motivate an appropriate choice of α; we also refer to the Hellinger-MINE discussion and Corollary 1 in Section V. The asymptotic variance (40) agrees with the variance of the LT objective functional, Eq. (37), at its optimizer $\phi_\alpha^* = (dQ/dP)^{\alpha-1}/|\alpha-1|$. Hence, from the perspective of the variance at the optimizer, neither method has an advantage. Away from the optimizer there is no consistent relationship between the two variances. In practice, both methods will take different paths to the optimizer and this further complicates any variance comparison away from the optimizer. However, empirically we found that the estimators constructed via the improved variational formulas have smaller variance at the estimated optimal which is different in general from the theoretical optimizer ϕ^* (i.e., after a finite number of SGD steps); see Figure 3.

V. NUMERICAL EXAMPLES: FASTER STATISTICAL ESTIMATION AND LEARNING

Next we discuss practical implications of using tighter variational representations developed in Theorem 1, focusing on accelerating neural-based statistical learning and estimation. In recent

works, variational representations such as (1) or (3) were used to estimate f-divergences and likelihood ratios based solely on available data [8]. This variational perspective proved also to be a crucial mathematical step in training generative adversarial networks (GAN) [11], [12], [13] and towards developing neural-based estimators for mutual information, [10], taking advantage of the ability of neural networks to search efficiently through function spaces.

Improved variational formulas for statistical estimation and learning were previously studied in: i) [9], using Eq. (13) and assuming a Hilbert space (RKHS) function space, ii) [10], where the DV and LT formulas for the KL divergence were used to estimate mutual information with improved accuracy. Both of these implicitly rely on the shift-improved variational formula (see Eq. (12) and Eq. (14)). Our Theorem 1 provides a broad generalization of these ideas to other transformation families, allows for practical implementation of the method in [9] to more general functions space parametrizations (e.g., neural networks), and generalizes the ideas in [10] to other f-divergences beyond KL, where it can provide improved mutual information estimators based on (6).

In the following, we employ the outcomes from Sections II & III and build several variational neural network estimators, in the general spirit of [8], [10]. We demonstrate the performance improvements that result from representations such as Eq. (6). We start with the heuristic observation, illustrated in Figure 1, that tighter representations can improve the accuracy of statistical estimators for f-divergences, in the sense that the same approximation of the optimal ϕ^* will provide a better approximation of the divergence. Moreover, tighter variational formulas can lead to faster convergence of the search algorithm, as we now motivate: suppose one minimizes a convex function $f(x)$ by the simple gradient descent algorithm $x_{n+1} = x_n - \gamma \nabla f(x_n)$. If ∇f is L-Lipschitz (i.e., the Hessian is bounded by L) then this algorithm converges if $0 < \gamma < 2/L$, and the analysis suggests the optimal learning rate of $\gamma = 1/L$ and leads to the error bound $|f(x_n) - f(x^*)| \leq \|x_0 - x^*\|^2 / (2\gamma n)$ (see, e.g., Theorem 3.3 in [36]). If $\widetilde{f}$ has the same optimizer and optimal value, but has a smaller Hessian bound, $\widetilde{L}$, then the optimal learning rate, $\widetilde{\gamma} = 1/\widetilde{L}$ is larger, and the error bound after an equal number of steps is smaller, i.e., the use of $\widetilde{f}$ in place of f can lead to faster convergence.

The above argument is only heuristic; the constant learning rate algorithm is far from optimal in most cases and the above analysis does not capture the complexity of the current setting. Nonetheless, it does provide important insight into the numerical results, presented below, which demonstrate that, in practice, the improved variational formulas do generally lead to faster

convergence of the estimators, letting all other factors be equal.

The examples below use the new variational formulas (17) and (19) for α-divergence, largely focusing on the well-known Hellinger divergence ($\alpha = 1/2$), defined as

$$D_{f_{1/2}}(Q\|P) = 4\left(1 - \int p^{1/2}q^{1/2}d\mu\right),\tag{41}$$

where $dQ = qd\mu$, $dP = pd\mu$. Computations were done in TensorFlow using the AdamOptimizer [37], an adaptive learning-rate SGD optimizer, with all methods given the same initial learning rate. When working with neural-network based estimators of α-divergence, we enforce positivity of the test functions (see (17) and (19)) via the parameterization $\phi_\theta = \exp(g_\theta)$ where g_θ, $\theta \in \Theta$, is a neural network family with ReLU activation functions. We compare the LT method, (1) with the improved estimators based on our formulas (17) and (19). For instance, in the case of Hellinger divergences we compare the following estimators.

LT Hellinger Estimator:

$$\sup_{\theta \in \Theta}\{E_{Q_N}[-\exp(g_\theta)] - E_{P_N}[f^*_{1/2}(-\exp(g_\theta))]\},\tag{42}$$

$$f^*_{1/2}(y) = \infty 1_{y \geq 0} + 4(|y|^{-1} - 1)1_{y<0},\tag{43}$$

Scaling-Improved Hellinger Estimator:

$$4\sup_{\theta \in \Theta}\left\{1 - E_{Q_N}[\exp(g_\theta)]^{1/2}E_{P_N}[\exp(-g_\theta)]^{1/2}\right\},\tag{44}$$

Scaling-Power-Improved Hellinger Estimator:

$$4\sup_{\theta \in \Theta, \beta \in \mathbb{R}}\left\{1 - E_Q[\exp(\beta g_\theta)]^{1/2}E_P[\exp(-\beta g_\theta)]^{1/2}\right\}.\tag{45}$$

In the above, Q_N and P_N denote the expectation under the empirical distributions using N iid samples from Q and P respectively.

If the optimization over a parameterized family of transformations, T_δ, cannot be performed analytically then we solve the minimization problem (6) - (7) by performing stochastic gradient descent (SGD) on the full collection of parameters, (θ, δ). In such cases, our two-step formulation can be thought of as parametric enhancement of the neural network architecture. The nested nature of the minimization over ϕ and T also allows for more sophisticated methods (not explored here), e.g., for each ϕ one can perform several SGD steps for T, thus solving the (generally low dimensional) problem (7) to high accuracy, before performing another SGD step for ϕ in (6); this is reminiscent of multiscale numerical methods [38]. The parameterization, δ, of the families of

transformations considered here is at most two dimensional. For example in the case of T^{affine} we have $\delta = (\nu, \eta)$, where $\nu, \eta \in \mathbb{R}$; see also the remaining examples in Section II. Including this small number of additional parameters in the stochastic gradient descent iterations is expected to add a negligible additional computational cost, as compared to the (generally) much larger number of neural-network parameters, θ. In practice, we do find the additional computational cost to be negligible.

Hellinger-MINE

Here we consider the problem of computing Hellinger mutual information (Hellinger-MI), $D_{f_{1/2}}(P_{(X,Y)}\|P_X \times P_Y)$. Typically the divergence in mutual information $D_f(P_{(X,Y)}\|P_X \times P_Y)$ is chosen to be KL. However, one can consider a whole array of different f-divergences for this purpose, see for instance [39]. A motivation for choosing Hellinger over KL is rigorously based on the variance calculations in Section IV. In particular, as a direct consequence of Theorem 3 we obtain the following:

Corollary 1: Under the assumptions of Theorem 3, the relative variance for the Hellinger α-divergence where $\alpha = 1/2$, at the optimizer ϕ^* is

$$
\frac{n \operatorname{Var}[\widehat{H}_{f_{1/2}}^n[\phi^*; Q, P]]}{D_{f_{1/2}}(Q\|P)^2} = \frac{\lim_{n\to\infty} n \operatorname{Var}[\widehat{H}_{scale,f_{1/2}}[\phi^*; Q_n, P_n]]}{D_{f_{1/2}}(Q\|P)^2}
\tag{46}
$$
$$
= \frac{8 - D_{f_{1/2}}(Q\|P)}{2D_{f_{1/2}}(Q\|P)} \, .
$$

Therefore, the sample complexity of the estimator $\widehat{H}_{scale,f_{1/2}}[\phi^*; Q_n, P_n]$ at the optimizer ϕ^* is $n = \mathcal{O}(1)$, when $Q \neq P$.

Comparing Corollary 1 with the corresponding result for the KL divergence from Theorem 2 of Ref. [35],

$$
\frac{\lim_{n\to\infty} n \operatorname{Var}[\widehat{H}_{DV}[\phi^*; Q_n, P_n]]}{D_{KL}(Q\|P)^2} \geq \frac{e^{D_{KL}(Q\|P)} - 1}{D_{KL}(Q\|P)^2} \, ,
\tag{47}
$$

we see that in practice the KL divergence requires $n = \mathcal{O}(e^{D_{KL}(Q,P)})$ samples, while the Hellinger divergence requires $n = \mathcal{O}(1)$ samples due to (46).

In Figure 2, we present the computation of Hellinger mutual information (Hellinger-MI), $D_{f_{1/2}}(P_{(X,Y)}\|P_X \times P_Y)$, via neural network optimization, where X and Y are correlated 20-dimensional Gaussians with component-wise correlation ρ. The results demonstrate that, for a given computational budget (i.e., fixed number of SGD iterations) the improved variational
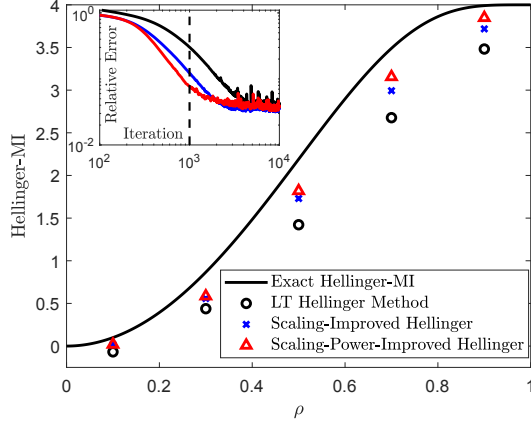
Fig. 2. Estimation of Hellinger-based mutual information between 20-dimensional correlated Gaussians with component-wise correlation ρ. We use a fully-connected neural network with one hidden layer of 64 nodes while training is performed with a minibatch size of 100. We show the Hellinger MI as a function of ρ after 1000 steps of SGD and averaged over 50 runs. The inset shows the relative error for $\rho = 0.7$, as a function of the number of SGD iterations.

formulas (red and blue) yield more accurate results, i.e., they converge faster than the LT f-divergence method (1) (black). Moreover, optimizing over both scalings and powers (19) (red) provides a non-trivial improvement over the scaling-improved method (17) (blue). This is a generalization of the findings in [10], which compared the DV variational formula (3) with (1) for the KL divergence. We emphasize that despite the lack of an analytical formula for the optimization over β, the inclusion of this single additional parameter in the variational formula (a negligible addition to the computational cost) leads to a clear performance gain.

As a followup, we demonstrate the effectiveness of our method in estimating Hellinger-MI for high-dimensional problems with low-dimensional structure. Specifically, in Figure 3 we compare 20-dimensional Gaussians embedded in high dimensional space via a nonlinear map. The left panels demonstrate the performance gain when using the curvature-improved objective functionals; we find that the optimized methods significantly outperform the LT method, especially in higher dimensions. The right panels are a (partial) demonstration that the variances are comparable, if not improved, when using the optimized objective functionals. Specifically, we find that the optimized methods approach the exact asymptotic variance faster while the LT method, which has not yet converged, has a larger variance.
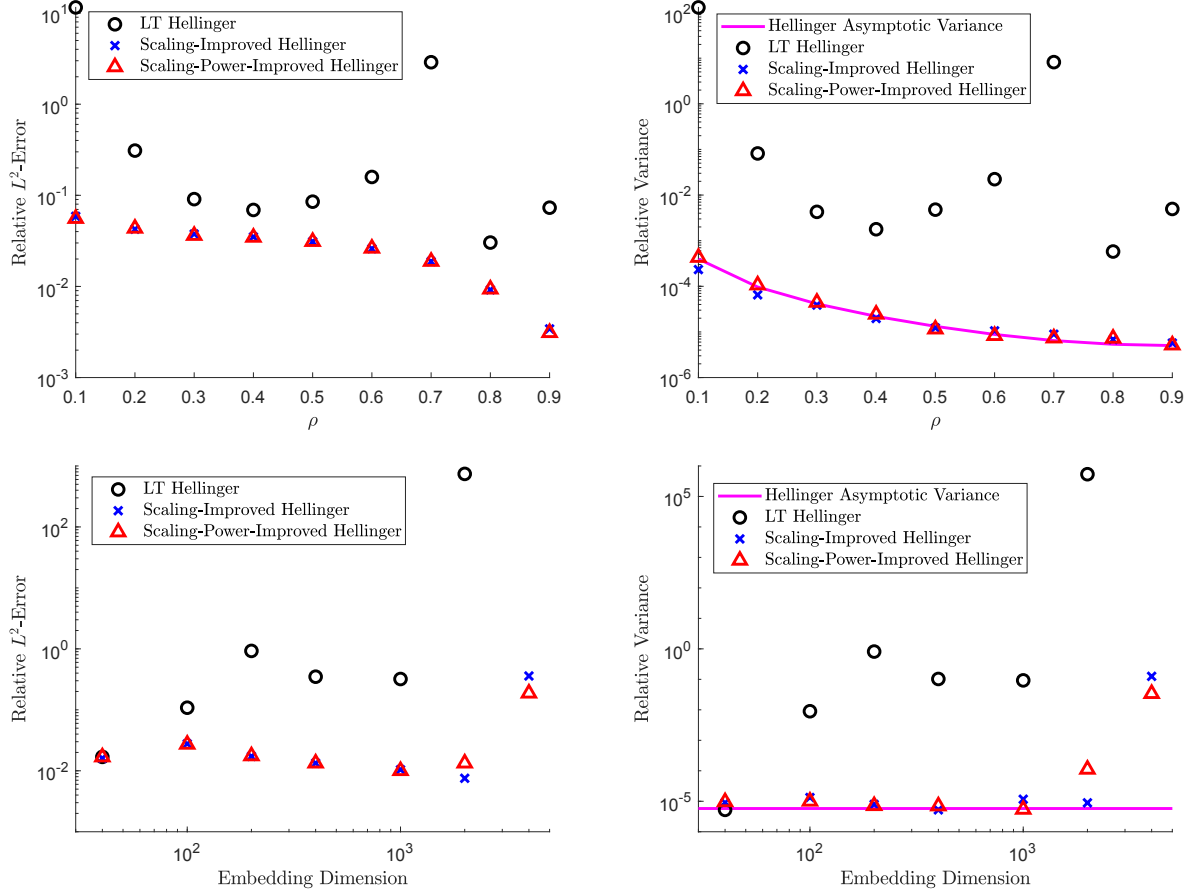
Fig. 3. Hellinger-MINE between correlated Gaussians embedded in high-dimensional space by a map of the following form: $h_i(x) = x_i$ for $i = 1, ..., 20$ and for $i > 20$ we define $h_i(x) = A_i(x) + \cos(x_{j_{1,i}}) \sin(x_{j_{2,i}}) + x_{j_{3,i}} x_{j_{4,i}}$, where A is an affine function and $j_{k,i} \in \{1, ..., 20\}$; the parameters of A were randomly selected. We employ a fully-connected neural network and performed training with a minibatch size of 1000 from a dataset of 100000 samples. We show results after 50000 steps of SGD and the relative L^2-error was computed using data from 75 runs. The left figures show the relative error while the right figures show the relative variance; the solid lines show the exact relative asymptotic variance (46). The top row show the result as a function of ρ for embedding-dimension 200 and used a single hidden layer of 64 nodes. The bottom row shows the result for $\rho = 0.75$ as a function of problem dimension (i.e., embedding dimension for X plus embedding dimension for Y) and used two hidden layers of 64 and 8 nodes respectively, irrespective of the embedding dimension; the majority of LT method runs diverged when the dimension equaled 4000.

Submanifold Parameterization for Exponential Families

Our method allows for a great deal of flexibility in the choice of function space parameterization. In a 'small-data' setting, the assumption of an exponential family structure can serve as an effective regularization. We illustrate this with Figure 4, which shows the estimation of the α-divergence with $\alpha = 0.25$ between 10-dimensional Gaussians using a data set of 5000 samples

from each distribution for SGD (minibatch size of 100) and using another 5000 samples for Monte Carlo estimation of the value of the objective functional. Using the submanifold estimation formula (20) and its scaling-improved variant (see Appendix E) we obtain the magenta and red curves, respectively. In blue, we show the result from the scaling-improved variational formula (17) and in black we show the result using the LT f-divergence objective functional (1); both use
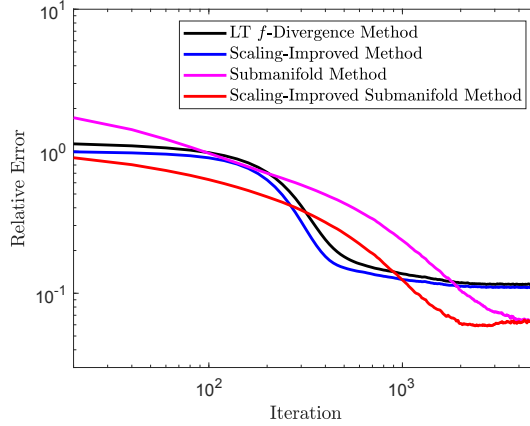

Fig. 4. Estimation of the α-divergence with $\alpha = 0.25$ between two 10-dimensional Gaussians with randomly generated variances and one of the means randomly perturbed from zero. We compare the convergence performance between the neural network and submanifold parameterizations. The relative error was averaged over 50 runs.

neural network families with one fully connected hidden layer (5 nodes). The number of nodes was chosen so that all methods use approximately the same number of parameters. The neural network parameterization converges faster, but ends up with a larger bias than the submanifold parameterization. The scaling-improved variational formulas lead to faster convergence than the LT variational formula in both cases as expected by our theory.

MNIST Dataset Examples

Next we illustrate the accelerated speed of convergence on high-dimensional ($28 \times 28 = 784$ dimensional) realistic data by estimating the Hellinger divergence between two distributions obtained by (iid) randomly translating the MNIST handwritten digits image dataset [40]. This provides an effective test case wherein we know the exact answer ($D_{f_{1/2}}(Q\|P) = 0$). Figure 5 shows the error, as a function of the number of SGD iterations, and once again demonstrates that the improved variational formulas lead to faster convergence; in this case, nearly one order of magnitude fewer SGD iterations are required to reach an accuracy of 10^{-2} when using the
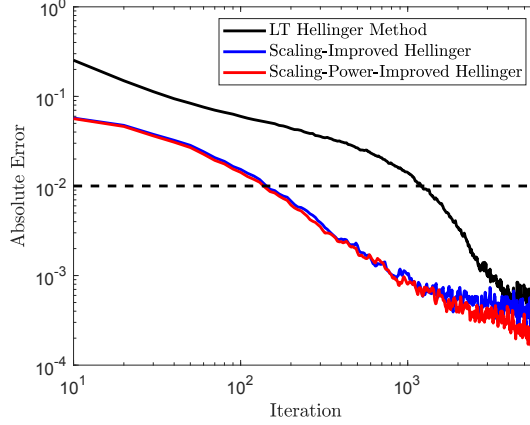
Fig. 5. Estimation of the Hellinger divergence between two distributions obtained by (iid) randomly translating the MNIST handwritten digits dataset [40]: each sample is a random translation of an MNIST image in the x and y directions (iid $N(0, 3^2)$ shifts, rounded to the nearest pixel and with periodic boundary conditions). Each step of SGD uses two independent minibatches of 100 such samples (one minibatch for P and one for Q). Monte Carlo estimation of the value of the Hellinger divergence was done using the corresponding objective functional and with samples coming from two separate datasets of 10000 randomly shifted MNIST images (one collection of images for P and an independent collection for Q). The function space was parameterized via fully-connected neural networks with one hidden layer of 128 nodes. The results were averaged over 50 runs.

tighter objective functionals. In practice, this means that one can more quickly detect whether or not the two data streams are in fact coming from the same distribution.

To further illustrate that our estimators are behaving appropriately, we perform a pair of consistency checks using the MNIST dataset, similar to the tests in [35]. While not a perfect substitute for computing the relative error, tests such as these are very useful in situations where the exact value of the divergence is nonzero and unknown. In Figure 6(a) we test the data processing inequality for f-divergences:

$$D_f(Q \otimes \kappa \| P \otimes \kappa) = D_f(Q \| P), \tag{48}$$

where κ is a probability kernel. Here we let Q be the MNIST dataset, P be the MNIST dataset of digits 0 through N_P where N_P ranges from 0 to 8, and we let κ_x be the distribution of random translations of the image x (specifically, $N(0, 1)$ translations, with components rounded to the nearest integer). The plot shows the ratio of the estimators for $D_{f_{1/2}}(Q \otimes \kappa \| P \otimes \kappa)$ and $D_{f_{1/2}}(Q \| P)$, using various objective functionals. In Figure 6(b) we test the product measure

identity for α-divergences:

$$D_{f_\alpha}(Q_1 \times ... \times Q_k \| P_1 \times ... \times P_k) = \frac{\prod_{i=1}^{k}(\alpha(\alpha-1)D_{f_\alpha}(Q_i\|P_i)+1)-1}{\alpha(\alpha-1)} . \qquad (49)$$

Here we let Q_i be copies of the MNIST dataset and P_i be copies of the MNIST dataset of digits 0 through N_P, where N_P again ranges from 0 to 8. The plot shows the ratio of the right-hand-side of Eq. (49) to the left-hand-side. The black horizontal lines in both panels show the exact value of 1. We find that all methods perform well on these consistency checks, though the best performing method is different for the two tests. This is unsurprising, as our methods are designed to accelerate the convergence to the optimum value, which is a different goal than preserving these the above two properties. Note that as N_P increases the distributions Q and P become more similar and so both the numerator and denominator approach zero, making the task of estimating the ratio more difficult; this is reflected in Figure 6.
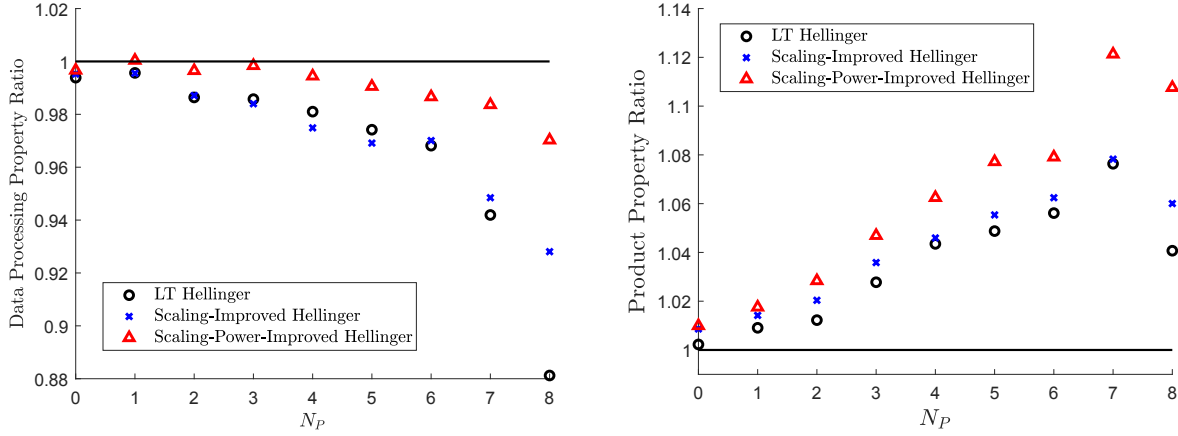

Fig. 6.   Hellinger divergence consistency tests (Left: testing the data processing inequality (48). Right: testing the product property (49)) using MNIST dataset. We employ a fully connected neural network with one hidden layer of 128 nodes. Training was performed with a minibatch size of 100. We show results after 5000 steps of SGD and results were averaged over 100 runs.

To conclude these examples, we note that although the proposed optimization framework was applied on statistical learning and estimation, it can be of broader interest, among others, in epistemic uncertainty quantification [14], in coarse-graining and model reduction [41], [42], [43], as well as in PAC learning [18] and adversarial learning [11], [13]. In particular, we intend to explore the use of variational formulas derived via the quadratic approximation method (e.g., Eq. (21) and Eq. (100)) for uncertainty quantification, along with the UQ bound (23).

VI. Proof of Theorem 1

In this section we provide a detailed proof of Theorem 1. For the convenience of the reader, we will recall the relevant definitions and notation below.

Let P, Q be probability measures on a measurable space $(\Omega, \mathcal{M})$ and, for any $-\infty \le a < 1 < b \le \infty$ define $\mathcal{F}_1(a, b)$ to be the set of convex functions $f : (a, b) \to \mathbb{R}$ with $f(1) = 0$. If a (resp. b) is finite, we extend f to a (resp. b) by continuity and set $f(x) = \infty$ for $x \notin [a, b]$. The result is a convex, lower semicontinuous function, $f : \mathbb{R} \to (-\infty, \infty]$. The f-divergence of Q with respect to P is defined by

$$
D_f(Q\|P) = \begin{cases} E_P[f(dQ/dP)], & Q \ll P \\ \infty, & Q \not\ll P \end{cases}.
\tag{50}
$$

Our starting point is the the following variational characterization [20], [8]:

$$
D_f(Q\|P) = \sup_{\phi \in \mathcal{M}_b(\Omega)} \{E_Q[\phi] - E_P[f^*(\phi)]\},
\tag{51}
$$

where $\mathcal{M}_b(\Omega)$ denotes the set of all bounded and measurable functions and

$$
f^*(y) = \sup_{x \in \mathbb{R}} \{yx - f(x)\} = \sup_{x \in (a,b)} \{yx - f(x)\}
\tag{52}
$$

is the Legendre transform of f.

Remark 4: Note that $f^*(y) \ge y$. This implies $f^*(\phi) \ge \phi$ is bounded below for $\phi \in \mathcal{M}_b(\Omega)$, and hence $E_P[f^*(\phi)] \in (-\infty, \infty]$ is always well-defined.

The technical aspects of the proof of Theorem 1 revolve around ensuring that all of the required expectations and operations are well-defined (without requiring any arbitrary convention regarding the definition of $\infty - \infty$). Modulo those details, the derivation of Eq. (6) is quite simple. As a first step, we show that Eq. (51) can be extended to certain unbounded ϕ. This is similar to results in [20], [8] but we will prove explicit conditions for which the expectations exist. To do this, we will need the following lemmas:

Lemma 1: Let $f \in \mathcal{F}_1(a, b)$. Then one of the following holds

1) f^* is bounded below.

2) The set $\operatorname{dom}(f^*) \equiv \{y : f^*(y) < \infty\}$ is of the form $\operatorname{dom}(f^*) = (-\infty, d)$ or $\operatorname{dom}(f^*) = (-\infty, d]$ for some $d \in (-\infty, \infty]$ and f^* is non-decreasing.

Proof: Suppose f^* is not bounded below. Take $y_n \in I$ with $f^*(y_n) \to -\infty$. We know $f^*(y) \ge y$ and so $y_n \le f^*(y_n) \to -\infty$ and hence $y_n \to -\infty$. I is convex so it we let $d = \sup I$ then this implies $(-\infty, d) \subset I \subset (-\infty, d]$.

To show f^* is non-decreasing, suppose that we have $x_1 < x_2$ with $f^*(x_1) > f^*(x_2)$. Taking y_n as above, find an n such that $y_n < x_1$ and $f^*(y_n) < f^*(x_2)$. f^* is convex and so, letting $t = (x_2 - x_1)/(x_2 - y_n) \in (0,1)$, we have

$$f^*(x_1) = f^*(ty_n + (1-t)x_2) \leq tf^*(y_n) + (1-t)f^*(x_2)$$

$$< tf^*(x_2) + (1-t)f^*(x_2) = f^*(x_2) < f^*(x_1). \tag{53}$$

This is a contradiction, hence f^* is non-decreasing. ∎

Lemma 2: Let $f \in \mathcal{F}_1(a,b)$ and suppose f^* is bounded below or $D_f(Q\|P) < \infty$. Then $E_P[f^*(\phi)^-] < \infty$ for all $\phi \in L^1(Q)$.

Remark 5: We use the notation $g^+ \equiv g1_{g \geq 0}$ and $g^- \equiv -g1_{g \leq 0}$ for positive and negative parts of a function $g = g^+ - g^-$.

Proof: Fix $\phi \in L^1(Q)$. If f^* is bounded below then $(f^*)^-$ is bounded above and the result is trivial, so suppose not. Lemma 1 then implies that f^* is non-decreasing. General properties of Legendre transforms on the real line imply that f^* is convex, lower semicontinuous, and f^* is continuous on $\overline{\mathrm{dom}(f^*)}$. Hence there exists $b \in \mathbb{R}$ such that $f^* \leq 0$ on $(-\infty, b]$ and $f^* \geq 0$ on (b, ∞) (note that $f^*(0) \geq 0$). Define $\phi_b = \phi1_{\phi \leq b} + b1_{\phi > b}$, so that $\phi_b \leq b$, $\phi_b \in L^1(Q)$, and

$$E_P[f^*(\phi)^-] = E_P[1_{\phi \leq b}f^*(\phi)^-] = E_P[1_{\phi \leq b}f^*(\phi_b)^-] \leq E_P[f^*(\phi_b)^-] = E_P[-f^*(\phi_b)]. \tag{54}$$

Hence $E_P[f^*(\phi_b)] \leq -E_P[f^*(\phi)^-]$.

Now define $\phi_{b,n} = -n1_{\phi_b < -n} + \phi_b1_{\phi_b \geq -n}$. ϕ_b is bounded above and so $\phi_{b,n} \in \mathcal{M}_b(\Omega)$ and we can use Eq. (51) to find

$$E_Q[\phi_{b,n}] \leq D_f(Q\|P) + E_P[f^*(\phi_{b,n})]. \tag{55}$$

We have $\phi_{b,n} \to \phi_b$ pointwise, $|\phi_{b,n}| \leq |\phi_b|$, and $\phi_b \in L^1(Q)$, so we can use the dominated convergence theorem to obtain

$$E_Q[\phi_b] \leq D_f(Q\|P) + \liminf_n E_P[f^*(\phi_{b,n})] \tag{56}$$

(here it was important that we are in the case where $D_f(Q\|P) < \infty$). We also have $\phi_{b,n+1} \leq \phi_{b,n}$, hence $f^*(\phi_{b,n+1}) \leq f^*(\phi_{b,n})$ (recall we are in the case where f^* is nondecreasing) and for N large enough we have $\phi_{b,n} \leq b$ for all $n \geq N$. f^* is continuous on $(-\infty, b]$, hence so

$$0 \leq -f^*(\phi_{b,n}) \nearrow -f^*(\phi_b). \tag{57}$$

Therefore the monotone convergence theorem implies $\lim_n E_P[f^*(\phi_{b,n})] = E_P[f^*(\phi_b)]$, and so

$$-\infty < E_Q[\phi_b] \leq D_f(Q\|P) + E_P[f^*(\phi_b)] \leq D_f(Q\|P) - E_P[f^*(\phi)^-]. \tag{58}$$

We therefore conclude that $E_P[f^*(\phi)^-] < \infty$. ∎

We can now prove that Eq. (51) can be extended to $\phi \in L^1(Q)$.

Theorem 4: Let $f \in \mathcal{F}_1(a, b)$ and suppose either f^* is bounded below or $D_f(Q\|P) < \infty$. Then

$$D_f(Q\|P) = \sup_{\phi \in L^1(Q)} \{E_Q[\phi] - E_P[f^*(\phi)]\}, \tag{59}$$

where the objective functional is valued in $[-\infty, \infty)$.

Proof: Lemma 2 implies $E_P[f^*(\phi)^-] < \infty$ for all $\phi \in L^1(Q)$ and so the objective functional in Eq. (59) is valued in $[-\infty, \infty)$. If we can show $E_Q[\phi] - E_P[f^*(\phi)] \leq D_f(Q\|P)$ for all $\phi \in L^1(Q)$ then the claimed result will follow by using Eq. (51).

Fix $\phi \in L^1(Q)$. If $D_f(Q\|P) = \infty$ or $E_P[f^*(\phi)] = \infty$ then the required bound is trivial, so suppose not. Then $f^*(\phi) < \infty$ P-a.s. We are in the case where $D_f(Q\|P) < \infty$, and so $Q \ll P$ and $f^*(\phi) < \infty$ Q-a.s. as well.

In summary, it suffices to show $E_Q[\phi] - E_P[f^*(\phi)] \leq D_f(Q\|P)$ in the case where $\phi \in L^1(Q)$, $f^*(\phi) \in L^1(P)$, $D_f(Q\|P) < \infty$, $\mathrm{range}(\phi) \subset \mathrm{dom}(f^*)$. To do this, fix $y_0 \in I$ and define $\phi_n = y_0 1_{\phi < -n} + \phi 1_{-n \leq \phi \leq n} + y_0 1_{\phi > n}$. $\phi_n \in \mathcal{M}_b(\Omega)$ and so Eq. (51) gives

$$D_f(Q\|P) \geq E_Q[\phi_n] - E_P[f^*(\phi_n)]. \tag{60}$$

We have $\phi_n \to \phi$ pointwise and $|\phi_n| \leq |\phi| + |y_0| \in L^1(Q)$, therefore the dominated convergence theorem $E_Q[\phi_n] \to E_Q[\phi]$.

We have $\mathrm{range}(\phi_n), \mathrm{range}(\phi) \subset \mathrm{dom}(f^*)$ and f^* is continuous on $\mathrm{dom}(f^*)$, therefore $f^*(\phi_n) \to f^*(\phi)$ pointwise. We also have

$$|f^*(\phi_n)| = |f^*(\phi_n)| 1_{\phi < -n} + |f^*(\phi_n)| 1_{-n \leq \phi \leq n} + |f^*(\phi_n)| 1_{\phi > n} \tag{61}$$

$$\leq |f^*(y_0)| + |f^*(\phi)| \in L^1(P),$$

hence the dominated convergence theorem implies $E_P[f^*(\phi_n)] \to E_P[f^*(\phi)]$. Combining these gives

$$E_Q[\phi_n] - E_P[f^*(\phi_n)] \to E_Q[\phi] - E_P[f^*(\phi)] \tag{62}$$

and so

$$D_f(Q\|P) \geq E_Q[\phi] - E_P[f^*(\phi)].$$ (63)

This proves the claim. ∎

We now prove Theorem 1 from the main text, which we restate below. For completeness, we also provide a derivation of the formula for the optimizer, ϕ^*, which was obtained in [20].

Theorem 5: Let $f \in \mathcal{F}_1(a, b)$ and suppose either f^* is bounded below or $D_f(Q\|P) < \infty$. Then:

1) Suppose $Q \ll P$, f is C^1, f' is strictly increasing, and one of the following holds:
   a) $a < dQ/dP < b$
   b) $a \leq dQ/dP \leq b$ and if the value a (resp. b) is achieved then $f'(a) \equiv \lim_{t \searrow a} f'(t)$ (resp. $f'(b) \equiv \lim_{t \nearrow b} f'(t)$) exists and is finite.

   Define $\phi^* = f'(dQ/dP)$. If $\phi^* \in L^1(Q)$ then the supremum in Eq. (59) is achieved at ϕ^*.

2) Let Φ be a family of functions with $\mathcal{M}_b(\Omega) \subset \Phi \subset L^1(Q)$ or with $\phi^* \in \Phi \subset L^1(Q)$ (in the latter case, we also assume that the conditions from part (1) hold, so that ϕ^* is the optimizer). Consider any family of transformations $\mathcal{T} \subset \{T = T(\phi), \text{ such that } T : \Phi \mapsto L^1(Q)\}$ that includes the identity map. Then

$$D_f(Q\|P) = \sup_{\phi \in \Phi}\{\sup_{T \in \mathcal{T}}\{E_Q[T(\phi)] - E_P[f^*(T(\phi))]\}\}.$$ (64)

3) If $Q \ll P$ and $(\Omega, \mathcal{M})$ is a metric space with the Borel σ-algebra then one can replace $\mathcal{M}_b(\Omega)$ with $C_b(\Omega)$ (bounded continuous functions) and $L^1(Q)$ with $L^1_c(Q)$ (continuous $L^1(Q)$ functions) in the above.

Remark 6: Recall that $-\infty \leq a < 1 < b \leq \infty$. In particular, one could have $b = \infty$ and so the assumptions of part (1) do not necessary imply dQ/dP is bounded.

Proof: Suppose that f satisfies the additional conditions from item (1). Then the Legendre transform can be computed:

$$f^*(y) = y(f')^{-1}(y) - f((f')^{-1}(y))$$ (65)

for all $y \in \text{range}(f')$. This implies that for $x \in (a, b)$ we have $f(x) = xf'(x) - f^*(f'(x))$. By taking limits and using the assumptions on dQ/dP, we find $f(dQ/dP) = dQ/dPf'(dQ/dP) - $

$f^*(f'(dQ/dP))$. Therefore, assuming $\phi^* \equiv f'(dQ/dP) \in L^1(Q)$, we have

$$E_Q[\phi^*] - E_P[f^*(\phi^*)] \tag{66}$$

$$= E_P[f'(dQ/dP)dQ/dP - f^*(f'(dQ/dP))]$$

$$= E_P[f(dQ/dP)] = D_f(Q\|P),$$

which proves the claim.

Now, let Φ and $\mathcal{T}$ be as in item (2). Since every $T \in \mathcal{T}$ maps Φ into $L^1(Q)$ and $\mathcal{T}$ contains the identity we can use Theorem 4 to obtain

$$E_Q[\phi] - E_P[f^*(\phi)] \leq \sup_{T \in \mathcal{T}} \{E_Q[T(\phi)] - E_P[f^*(T(\phi))]\} \leq D_f(Q\|P) \tag{67}$$

for all $\phi \in \Phi$. Maximizing over Φ and using the fact that either $\phi^* \in L^1(Q)$ or $\mathcal{M}_b \subset \Phi \subset L^1(Q)$ along with Eq. (59) gives

$$D_f(Q\|P) \leq \sup_{\phi \in \Phi}\{E_Q[\phi] - E_P[f^*(\phi)]\} \tag{68}$$

$$\leq \sup_{\phi \in \Phi} \sup_{T \in \mathcal{T}} \{E_Q[T(\phi)] - E_P[f^*(T(\phi))]\} \leq D_f(Q\|P),$$

which proves the claim.

Finally, on a metric space, one can approximate measurable functions with continuous functions via Lusin's theorem (see, e.g., Appendix D in [22]). Using this fact it is straightforward to show that the same results are obtained if one replaces $\mathcal{M}_b(\Omega)$ with $C_b(\Omega)$ and $L^1(Q)$ with $L^1_c(Q)$ (see the Proof of Theorem 1 in [31] for details on the use of this technique in a similar context). This proves item (3).

∎

APPENDIX A
DERIVATION OF VARIATIONAL FORMULAS FOR α AND RÉNYI DIVERGENCES

Here we provide additional details regarding the derivation of the new variational formulas for α-divergences, as well as their connection to the Rényi divergences.

1. Recall that the α-divergences are the family of f-divergences corresponding to

$$f_\alpha(t) = \frac{t^\alpha - 1}{\alpha(\alpha - 1)}, \quad \alpha \neq 0, 1. \tag{69}$$

First consider the case $\alpha > 1$: The Legendre transform of f_α is

$$f_\alpha^*(y) = y^{\alpha/(\alpha-1)}\alpha^{-1}(\alpha-1)^{\alpha/(\alpha-1)} + \frac{1}{\alpha(\alpha-1)}, \quad y \geq 0 \tag{70}$$

and the exact optimizer (2) is non-negative. With this in mind, fix $\phi \geq 0$, $\phi \in L^1(Q)$, and $\phi^{\alpha/(\alpha-1)} \in L^1(P)$, with ϕ not P-a.s. zero. For $\eta > 0$ we can write

$$E_Q[\eta\phi] - E_P[f_\alpha^*(\eta\phi)] \tag{71}$$

$$= \eta E_Q[\phi] - \eta^{\alpha/(\alpha-1)}\alpha^{-1}(\alpha-1)^{\alpha/(\alpha-1)}E_P[\phi^{\alpha/(\alpha-1)}] - \frac{1}{\alpha(\alpha-1)}.$$

For $a, c \geq 0$, $b > 0$ we have the following general solution to the optimization problem

$$\sup_{\eta > 0}\{a\eta - b\eta^{\alpha/(\alpha-1)} - c\} = \frac{a^\alpha}{\alpha b^{\alpha-1}}\left(\frac{\alpha-1}{\alpha}\right)^{\alpha-1} - c, \tag{72}$$

where the maximum occurs at $\eta^* = (a(\alpha-1)/(b\alpha))^{\alpha-1}$. Hence we can use this formula to obtain

$$\sup_{\eta > 0}\{E_Q[\eta\phi] - E_P[f_\alpha^*(\eta\phi)]\} = \frac{1}{\alpha(\alpha-1)}E_Q[\phi]^\alpha E_P[\phi^{\alpha/(\alpha-1)}]^{-(\alpha-1)} - \frac{1}{\alpha(\alpha-1)}. \tag{73}$$

Using this, for $Q \ll P$ with $(dQ/dP)^\alpha \in L^1(P)$ we find

$$D_{f_\alpha}(Q\|P = \frac{1}{\alpha(\alpha-1)}\sup_{\phi \in L^1(Q):\phi \geq 0}\{E_Q[\phi]^\alpha E_P[\phi^{\alpha/(\alpha-1)}]^{1-\alpha} - 1\}, \tag{74}$$

and the maximum is achieved at

$$\phi_\alpha^* = (dQ/dP)^{\alpha-1}. \tag{75}$$

In Eq. (74) one should interpret $0/0 \equiv 1$ and $c/\infty \equiv 0$, or else restrict to the subset of functions with $0 < E_P[\phi^{\alpha/(\alpha-1)}] < \infty$ for which these cases do not occur. Also note that the objective functional (74) is now invariant under scaling and so we were able to drop the factor of $1/(\alpha-1)$ in Eq. (75).

If $\alpha \in (0,1)$ and $y < 0$ then

$$f_\alpha^*(y) = |y|^{-\alpha/(1-\alpha)}\alpha^{-1}(1-\alpha)^{-\alpha/(1-\alpha)} - \frac{1}{\alpha(1-\alpha)}, \tag{76}$$

and, if $Q \ll P$, a similar computation to the above yields

$$D_{f_\alpha}(Q\|P) = \frac{1}{\alpha(1-\alpha)}\sup_{\phi > 0}\{1 - E_Q[\phi]^\alpha E_P[\phi^{-\alpha/(1-\alpha)}]^{1-\alpha}\}. \tag{77}$$

If $dQ/dP > 0$ then the exact optimizer is $\phi_\alpha^* = (dQ/dP)^{\alpha-1}$. Note that we reparameterized $\phi \to -\phi$ in Eq. (77) so that we optimize over strictly positive functions, rather than strictly

negative functions. Also, note that it is not necessary to restrict the optimization in (77) to $\phi \in L^1(Q)$; if $E_Q[\phi] = \infty$ then the objective functional equals $-\infty$ and hence including such ϕ's does not change the value of the supremum. Theorem 1 guarantees that the objective functionals in the variational representations Eq. (74), and Eq. (77) are tighter than that of the LT f-divergence objective functional from (51).

2. In particular, for the Hellinger distance ($\alpha = 1/2$) we find the scaling-improved formula

$$D_{f_{1/2}}(Q\|P) = 4\sup_{\phi>0}\left\{1 - E_Q[\phi]^{\frac{1}{2}}E_P[\phi^{-1}]^{\frac{1}{2}}\right\}, \tag{78}$$

as compared to the LT f-divergence variational formula (after changing variables $\phi \to -\phi$)

$$D_{f_{1/2}}(Q\|P) = \sup_{\phi>0}\{4 - E_Q[\phi] - 4E_P[\phi^{-1}]\}. \tag{79}$$

The improved tightness, $4 - E_Q[\phi] - 4E_P[\phi^{-1}] \leq 4(1 - E_Q[\phi]^{\frac{1}{2}}E_P[\phi^{-1}]^{\frac{1}{2}})$, is guaranteed by the general result of Theorem 5, but it can also be seen to be a consequence of the inequality $ab \leq a^2 + b^2/4$ applied to $a = E_P[\phi^{-1}]^{\frac{1}{2}}$, $b = E_Q[\phi]^{\frac{1}{2}}$.

3. For the χ^2-divergence (equal to $2D_{f_2}$) one can evaluate the optimization over all affine tranformations. We have $f_2(x) = (x^2 - 1)/2$ and $f_2^*(y) = (y^2 + 1)/2$. Suppose $D_{f_2}(Q\|P) < \infty$. Optimizing first over shifts we find, for $\phi \in L^1(Q) \cap L^2(P)$,

$$\sup_{\nu\in\mathbb{R}}\{E_Q[\phi - \nu] - \frac{1}{2}E_P[(\phi - \nu)^2 + 1]\} \tag{80}$$

$$= E_Q[\phi] + \frac{1}{2}(E_P[\phi] - 1)^2 - \frac{1}{2}(1 + E_P[\phi^2])$$

$$= E_Q[\phi] - E_P[\phi] - \frac{1}{2}\operatorname{Var}_P[\phi],$$

with the maximum occurring at $\nu^* = E_P[\phi] - 1$. Therefore we obtain the variational representation

$$\chi^2(Q\|P) = \sup_{\phi\in L^1(Q)\cap L^2(P)}\{E_Q[2\phi] - E_P[2\phi] - \operatorname{Var}_P[\phi]\}$$

$$= \sup_{\phi\in L^1(Q)\cap L^2(P)}\{E_Q[\phi] - E_P[\phi] - \frac{1}{4}\operatorname{Var}_P[\phi]\}. \tag{81}$$

Note that the objective functional in the last line is the same as that obtained in Eq. (48) of [25].

Further optimizing the objective functional over the scaling parameter $\eta \in \mathbb{R}$ we find

$$\sup_{\eta\in\mathbb{R}}\{E_Q[\eta\phi] + \frac{1}{2}(E_P[\eta\phi] - 1)^2 - \frac{1}{2}(1 + E_P[(\eta\phi)^2])\}$$

$$= \frac{1}{2}\frac{(E_Q[\phi] - E_P[\phi])^2}{\operatorname{Var}_P[\phi]} \tag{82}$$

(if $\mathrm{Var}_P[\phi] > 0$), with the maximum occurring at $\eta^* = (E_Q[\phi] - E_P[\phi])/\mathrm{Var}_P[\phi]$ (if $\mathrm{Var}_P[\phi] = 0$ then the supremum equals zero, since $Q \ll P$). Therefore

$$\chi^2(Q\|P) = 2 \sup_{\phi \in L^1(Q)} \sup_{\eta,\nu \in \mathbb{R}} \{ E_Q[\eta\phi - \nu] - E_P[f_2^*(\eta\phi - \nu)] \} \tag{83}$$

$$= \sup_{\phi \in L^1(Q) \cap L^2(P): \mathrm{Var}_P[\phi] > 0} \frac{(E_Q[\phi] - E_P[\phi])^2}{\mathrm{Var}_P[\phi]} .$$

Equality is achieved at $\phi^* = dQ/dP$ and so one can further restrict the optimization to $\phi \geq 0$. This provides a rigorous justification of the loss function for χ^2-GANs proposed in, [27]. We emphasize that the objective functional in (83) is tighter than the one in (81).

4. The α-divergences for $\alpha = 0, 1$ are the KL divergences: $D_{f_0}(Q\|P) = D_{KL}(P\|Q)$ and $D_{f_1}(Q\|P) = D_{KL}(Q\|P)$. Reparameterizing the objective functional in Eq. (77) via

$$\phi \to e^{\pm(\alpha-1)g} \tag{84}$$

(with the optimization running over all measurable g) provides connections to the Donsker-Varadhan formula for the KL divergence in the limits $\alpha \to 0, 1$. Under the reparameterization $\phi_\alpha = e^{(\alpha-1)g}$ one has

$$E_Q[\phi_\alpha]^\alpha E_P[\phi_\alpha^{-\alpha/(1-\alpha)}]^{1-\alpha} = 1 + (E_Q[g] - \log E_P[e^g])(\alpha - 1) + O((\alpha - 1)^2) \tag{85}$$

and so

$$\lim_{\alpha \to 1} \frac{1}{\alpha(1-\alpha)} (1 - E_Q[\phi_\alpha]^\alpha E_P[\phi_\alpha^{-\alpha/(1-\alpha)}]^{1-\alpha}) = E_Q[g] - \log E_P[e^g] , \tag{86}$$

which is the the DV objective functional for $D_{KL}(Q\|P)$.

Similarly, reparameterizing via $\phi_\alpha = e^{-(\alpha-1)g}$ gives

$$E_Q[\phi_\alpha]^\alpha E_P[\phi_\alpha^{\alpha/(\alpha-1)}]^{-(\alpha-1)} = 1 - (E_P[g] - \log E_Q[e^g])\alpha + O(\alpha^2) \tag{87}$$

and so

$$\lim_{\alpha \to 0} \frac{1}{\alpha(1-\alpha)} (1 - E_Q[\phi_\alpha]^\alpha E_P[\phi_\alpha^{-\alpha/(1-\alpha)}]^{1-\alpha}) = E_P[g] - \log E_Q[e^g] , \tag{88}$$

which is the objective functional for the Donsker-Varadhan representation of $D_{KL}(P\|Q)$.

5. Using the same reparametrization as in Eq. (84) we can also derive a connection with the Rényi family of divergences: Fix $\alpha \in (0, 1)$, reparametrize $\phi = e^{(\alpha-1)g}$ (g is any measurable function), and rewrite Eq. (77) in terms of an optimization over g:

$$D_{f_\alpha}(Q\|P) = \frac{1}{\alpha(1-\alpha)} \sup_g \left\{ 1 - E_Q[e^{(\alpha-1)g}]^\alpha E_P[e^{\alpha g}]^{1-\alpha} \right\} . \tag{89}$$

Using the connection between α-divergences and Rényi divergences (see [21], but note that our definition of Rényi divergence differs from theirs by a factor of $1/\alpha$), we obtain

$$
\begin{aligned}
R_\alpha(Q\|P) =& \frac{1}{\alpha(\alpha-1)}\log(\alpha(\alpha-1)D_{f_\alpha}+1) && (90)\\
=& \frac{1}{\alpha(\alpha-1)}\log(\inf_g E_Q[e^{(\alpha-1)g}]^\alpha E_P[e^{\alpha g}]^{1-\alpha})\\
=& \sup_g\left\{\frac{1}{\alpha-1}\log(E_Q[e^{(\alpha-1)g}])-\frac{1}{\alpha}\log(E_P[e^{\alpha g}])\right\}.
\end{aligned}
$$

A similar calculation applies when $\alpha > 1$; in either case, we obtain an independent derivation of the Rényi divergence variational formula from [31].

APPENDIX B
IMPROVED DONSKER-VARADHAN VARIATIONAL FORMULA

Here we collect some additional properties of the improved Donsker-Varadhan (DV) variational formula (16):

$$
D_{KL}(Q\|P) = \sup_{\phi\in\Phi}\{\sup_{\eta\in\mathbb{R}}\{\eta E_Q[\phi]-\log E_P[e^{\eta\phi}]\}\}. \tag{91}
$$

One obviously has $\mathcal{T}^{id}\subset\mathcal{T}^{shift}\subset\mathcal{T}^{affine}$. As a consequence, the objective functional in (91) is tighter than DV, which in turn is tighter than the LT f-Divergence objective functional from Eq. (51):

$$
\overbrace{\sup_{\eta\in\mathbb{R}}\{\eta E_Q[\phi]-\log E_P[e^{\eta\phi}]\}}^{\text{Improved DV Eq.(91)}} \geq \overbrace{E_Q[\phi]-\log E_P[e^{\phi}]}^{\text{DV Eq.(3)}} \geq \overbrace{E_Q[\phi]-E_P[e^{\phi-1}]}^{\text{LT } f\text{-Divergence Eq.(51)}}.
$$

Although the supremum over η in Eq. (91) cannot in general be evaluated analytically, one can obtain an explicit approximation as follows: Define $G_\phi(\eta)=\eta E_Q[\phi]-\log E_P[e^{\eta\phi}]$ and Taylor expand around $\eta=1$ to obtain

$$
G_\phi(1+\Delta\eta) = E_Q[\phi]-\log E_P[e^\phi]+(E_Q[\phi]-E_{P_\phi}[\phi])\Delta\eta-\frac{\mathrm{Var}_{P_\phi}[\phi]}{2}\Delta\eta^2+O(\Delta\eta^3). \tag{92}
$$

We approximate the optimal $\Delta\eta$ by maximizing the quadratic approximation (92) to find

$$
\Delta\eta^*(\phi) = \frac{E_Q[\phi]-E_{P_\phi}[\phi]}{\mathrm{Var}_{P_\phi}[\phi]}, \tag{93}
$$

where $dP_\phi = e^\phi dP/E_P[e^\phi]$ is the tilted measure and $\Delta\eta^*(\phi)$ is defined to be 0 if $\mathrm{Var}_{P_\phi}[\phi]=0$. Using this, we obtain a new variational representation:

Theorem 6: Define $\Phi = \{\phi \in L^1(Q) : e^\phi, \phi e^\phi \in L^1(P)\}$ and for $\phi \in \Phi$ define $\Delta\eta^*(\phi)$ as in Eq. (93). Suppose $\log(dQ/dP) \in L^1(Q)$. Then

$$D_{KL}(Q\|P) = \sup_{\phi \in \Phi}\{(1 + \Delta\eta^*(\phi))E_Q[\phi] - \log E_P[e^{(1+\Delta\eta^*(\phi))\phi}]\}. \qquad (94)$$

Proof: It is easy to see that the objective functional in Eq. (21) is well defined for all $\phi \in \Phi$. To see that Eq. (21) is an equality, first recall that $D_{KL}(Q\|P) \geq G_\phi(\eta)$ for all η and all $\phi \in L^1(Q)$ (see Eq. (91)), and so $D_{KL}(Q\|P) \geq \sup_{\phi \in \Phi} G_\phi(1 + \Delta\eta^*(\phi))$. Computing the value at $\phi^* = \log(dQ/dP) \in \Phi$ we see that $P_{\phi^*} = Q$ and so $\Delta\eta^*(\phi^*) = 0$ and

$$\sup_{\phi \in \Phi}\{(1 + \Delta\eta^*(\phi))E_Q[\phi] - \log E_P[e^{(1+\Delta\eta^*(\phi))\phi}]\} \qquad (95)$$

$$\geq (1 + \Delta\eta^*(\phi^*))E_Q[\phi^*] - \log E_P[e^{(1+\Delta\eta^*(\phi^*))\phi^*}]$$

$$= E_Q[\phi^*] - \log E_P[e^{\phi^*}] = D_{KL}(Q\|P).$$

This proves (21). ∎

Eq. (21) is a new variational representation of the KL-divergence, however we make no claim that the objective functional in (21) is tighter than Donsker-Varadhan for every ϕ; Eq. (21) is only guaranteed to be tighter than DV when $\Delta\eta^*(\phi)$ is sufficiently small. This is because we only maximized the quadratic approximation in η, and hence only obtained an approximation to the scaling-improved objective functional from Eq. (91). One can of course circumvent this by using the maximum of the two:

$$D_{KL}(Q\|P) = \sup_{\phi \in \Phi}\left\{\max\left\{(1 + \Delta\eta^*(\phi))E_Q[\phi] - \log E_P[e^{(1+\Delta\eta^*(\phi))\phi}], E_Q[\phi] - \log E_P[e^\phi]\right\}\right\}.$$
$$(96)$$

Although Eq. (96) is less than aesthetically appealing, note that its objective functional is certainly no worse than DV and will be tighter when $\Delta\eta^*(\phi)$ is sufficiently small. Also note that no additional expectations need to be computed to evaluate the objective functional in Eq. (96) as compared to that of Eq. (21).

APPENDIX C
APPROXIMATING THE POWER-IMPROVED RÉNYI VARIATIONAL REPRESENTATION

The same method used to derive the second-order approximation of the improved Donsker-Varadhan variational formula (21) can be applied to the Rényi divergence with power transfor-

mations. Recall that the power-improved Rényi variational is given by

$$R_\alpha(Q\|P) = \sup_g \sup_\beta \left\{ \frac{1}{\alpha-1} \log(E_Q[e^{(\alpha-1)\beta g}]) - \frac{1}{\alpha} \log(E_P[e^{\alpha\beta g}]) \right\}. \tag{97}$$

Define $G_{g,\alpha}(\beta) = \frac{1}{\alpha-1} \log(E_Q[e^{(\alpha-1)\beta g}]) - \frac{1}{\alpha} \log(E_P[e^{\alpha\beta g}])$ to be the objective functional as a function of β. A Taylor expansion around $\beta = 1$ gives

$$G_{g,\alpha}(1 + \Delta\beta) = G_{g,\alpha}(1) + (E_{Q_{\alpha-1}}[g] - E_{P_\alpha}[g])\Delta\beta \tag{98}$$
$$- \frac{1}{2}((1-\alpha)\operatorname{Var}_{Q_{\alpha-1}}[g] + \alpha\operatorname{Var}_{P_\alpha}[g])\Delta\beta^2 + O(\Delta\beta^3),$$

with $dQ_\alpha = e^{\alpha g} dQ / E_Q[e^{\alpha g}]$ being the α-tilted measure and similarly for P. The quadratically-optimal $\Delta\beta^*$ is obtained by maximizing the second-order Taylor approximation and is given by

$$\Delta\beta^*(g) = \frac{E_{Q_{\alpha-1}}[g] - E_{P_\alpha}[g]}{(1-\alpha)\operatorname{Var}_{Q_{\alpha-1}}[g] + \alpha\operatorname{Var}_{P_\alpha}[g]}, \tag{99}$$

where we assumed $0 < \alpha < 1$. If $(1-\alpha)\operatorname{Var}_{Q_{\alpha-1}}[g] + \alpha\operatorname{Var}_{P_\alpha}[g] = 0$ then we set $\Delta\beta^*(g) = 0$. A new Rényi representation formula is then obtained as the following theorem asserts.

Theorem 7: Define $\Phi = \{g : g^k e^{(\alpha-1)g} \in L^1(Q), g^k e^{\alpha g} \in L^1(P) \text{ with } k = 0, 1\}$ and for $g \in \Phi$ define $\Delta\beta^*(g)$ as in Eq. (99). Suppose $\log(dQ/dP) \in \Phi$. Then for $\alpha \in (0,1)$ we have

$$R_\alpha(Q\|P) = \sup_{g \in \Phi} \left\{ \frac{1}{\alpha-1} \log(E_Q[e^{(\alpha-1)(1+\Delta\beta^*(g))g}]) - \frac{1}{\alpha} \log(E_P[e^{\alpha(1+\Delta\beta^*(g))g}]) \right\}. \tag{100}$$

Proof: The proof is similar to the proof of Theorem 6. First, the integrability assumptions ensure that $\Delta\beta^*(g)$, and hence the objective functional, are well-defined (the latter possibly equaling $-\infty$). Second, $R_\alpha(Q\|P) \geq G_{g,\alpha}(\beta)$ for all β and $g \in \Phi$ thus $R_\alpha(Q\|P) \geq G_{g,\alpha}(1 + \Delta\beta^*(g))$. It remains to show that the there is a g such that the supremum is attained. Taking $g^* = \log(dQ/dP) \in \Phi$, it is sufficient to show that $\Delta\beta^*(g^*) = 0$. We compute the two terms of the numerator of $\Delta\beta^*(g^*)$:

$$E_{Q_{\alpha-1}}[g^*] = \frac{E_Q[e^{(\alpha-1)\log(dQ/dP)}\log(dQ/dP)]}{E_Q[e^{(\alpha-1)\log(dQ/dP)}]}$$
$$= \frac{E_Q[(dQ/dP)^{(\alpha-1)}\log(dQ/dP)]}{E_Q[(dQ/dP)^{(\alpha-1)}]}$$
$$= \frac{E_P[(dQ/dP)^{\alpha}\log(dQ/dP)]}{E_P[(dQ/dP)^{\alpha}]},$$

and

$$E_{P_\alpha}[g^*] = \frac{E_P[e^{\alpha \log(dQ/dP)} \log(dQ/dP)]}{E_P[e^{\alpha \log(dQ/dP)}]} \tag{101}$$

$$= \frac{E_P[(dQ/dP)^\alpha \log(dQ/dP)]}{E_P[(dQ/dP)^\alpha]} .$$

Thus $\Delta\beta^*(g^*) = 0$, which completes the proof. ∎

APPENDIX D

TIGHTNESS GAINS AND VARIATIONAL DERIVATIVES

Here we prove an extension of Theorem 2 for the more general case of f-divergences.

Theorem 8 (Tightness gains for f-divergences): In addition to the assumptions of Theorem 1 suppose f' is strictly increasing and $f^* \in C^3((c,d))$ with $(f^*)'' > 0$. We select the function space $\Phi = \mathcal{M}_b(\Omega)$ and assume that the maximizer ϕ^* in (2) is valued in a compact subset of (c,d). Then, for $\mathcal{T} = \mathcal{T}^{id}, \mathcal{T}^{shift}, \mathcal{T}^{scale}, \mathcal{T}^{affine}$ we have that $\mathcal{J}_\mathcal{T}$ in (27) is twice differentiable at $\epsilon = 0$ for any $\psi \in \mathcal{M}_b(\Omega)$. Furthermore, using the notation (28), the corresponding 2nd Gateaux derivatives $\frac{d^2}{d\epsilon^2}\mathcal{J}_\mathcal{T}(0) = \frac{d^2}{d\epsilon^2}\big|_{\epsilon=0} H_\mathcal{T}[\phi^* + \epsilon\psi]$ for all $\mathcal{T} = \mathcal{T}^{id}, \mathcal{T}^{shift}, \mathcal{T}^{affine}$ satisfy the following:

$$\langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle = - E_P[(f^*)''(\phi^*)]E_{P^*}[(\psi)^2]$$

$$= - E_P[(f^*)''(\phi^*)] \left(\mathrm{Var}_{P^*}[\psi] + E_{P^*}[\psi]^2 \right) , \tag{102}$$

$$\langle \nabla^2 H_{shift}[\phi^*]\psi, \psi \rangle = - E_P[(f^*)''(\phi^*)] \mathrm{Var}_{P^*}[\psi], \tag{103}$$

$$\langle \nabla^2 H_{scale}[\phi^*]\psi, \psi \rangle = - E_P[(f^*)''(\phi^*)] \left(\mathrm{Var}_{P^*}[\psi] + E_{P^*}[\psi]^2 - \frac{(E_{P^*}[\phi^*\psi])^2}{E_{P^*}[(\phi^*)^2]} \right)$$

$$= - E_P[(f^*)''(\phi^*)] \left(E_{P^*}[(\psi)^2] - \frac{(E_{P^*}[\phi^*\psi])^2}{E_{P^*}[(\phi^*)^2]} \right) , \tag{104}$$

$$\langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle = - E_P[(f^*)''(\phi^*)] \left[\mathrm{Var}_{P^*}[\psi] - \frac{\mathrm{Cov}_{P^*}(\phi^*, \psi)^2}{\mathrm{Var}_{P^*}[\phi^*]} \right]$$

$$= - E_P[(f^*)''(\phi^*)] \mathrm{Var}_{P^*}[\psi] \left[1 - \rho_{P^*}^2(\phi^*, \psi) \right] , \tag{105}$$

corresponding to (1), (9), (10) and the combination of the last two respectively. The tilted probability measure P^* is defined as

$$dP^* = (f^*)''(\phi^*)dP/E_P[(f^*)''(\phi^*)], \tag{106}$$

while $\rho_{P^*}(\phi^*, \psi)$ denotes the correlation between ϕ^* and ψ.

Remark 7: The relations (102), (103), (104) and (105) imply the following comparisons in variational curvatures around the optimizer ϕ^*, extending the discussion in Section III to f-divergences:

$$\langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle \geq \langle \nabla^2 H_{scale}[\phi^*]\psi, \psi \rangle \geq \langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle \tag{107}$$

and

$$\langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle \geq \langle \nabla^2 H_{shift}[\phi^*]\psi, \psi \rangle \geq \langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle . \tag{108}$$

Furthermore, relations (102), (103), (104) and (105) quantify precisely the gains in each inequality; compare also to the demonstration in Figure 1. We note that the inequality $\langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle \geq \langle \nabla^2 H_{scale}[\phi^*]\psi, \psi \rangle$ readily follows from (104) and (105) when all pertinent terms are rewritten using the correlation $\rho_{P^*}(\phi^*, \psi)$.

Remark 8: In the KL case we readily have $P^* = Q$ and $E_P[(f^*)''(\phi^*)] = 1$, thus Theorem 8 implies the results of Theorem 2.

Proof: For the convenience of the reader, we start by repeating several of the equations from the main text. First define

$$H[\phi, \eta, \nu] = E_Q[\eta\phi - \nu] - E_P[f^*(\eta\phi - \nu)] . \tag{109}$$

Optimization over the affine family leads to four different variational representations of the f-divergence

$$D_f(Q\|P) = \sup_\phi \overbrace{H[\phi, 1, 0]}^{H_{id}[\phi]} = \sup_\phi \sup_\nu \overbrace{H[\phi, 1, \nu]}^{H_{shift}[\phi]} \tag{110}$$

$$= \sup_\phi \sup_\eta \overbrace{H[\phi, \eta, 0]}^{H_{scale}[\phi]} = \sup_\phi \sup_{\eta,\nu} \overbrace{H[\phi, \eta, \nu]}^{H_{affine}[\phi]} .$$

Then, using the notation (28) for the 2nd Gateaux derivatives, we have the corresponding variational Hessians

$$\langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle = \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} H[\phi^* + \epsilon\psi, 1, 0] , \tag{111}$$

$$\langle \nabla^2 H_{shift}[\phi^*]\psi, \psi \rangle = \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \sup_\nu H[\phi^* + \epsilon\psi, 1, \nu] ,$$

$$\langle \nabla^2 H_{scale}[\phi^*]\psi, \psi \rangle = \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \sup_\eta H[\phi^* + \epsilon\psi, \eta, 0] ,$$

$$\langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle = \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \sup_{\eta,\nu} H[\phi^* + \epsilon\psi, \eta, \nu] .$$

The following general computation will facilitate the computation of the variational Hessians that appear in Eq. (111): The maxima in Eq. (110) are achieved at $\eta = 1$, $\nu = 0$, $\phi = \phi^* \equiv f'(dQ/dP)$. Given C^2 families $\phi_\epsilon = \phi^* + \epsilon\psi$, η_ϵ with $\eta_0 = 1$, and ν_ϵ with $\nu_0 = 0$, we can compute the derivatives

$$\frac{d}{d\epsilon}H[\phi_\epsilon, \eta_\epsilon, \nu_\epsilon] = E_Q[\eta_\epsilon'\phi_\epsilon + \eta_\epsilon\psi - \nu_\epsilon'] \tag{112}$$
$$- E_P[(f^*)'(\eta_\epsilon\phi_\epsilon - \nu_\epsilon)(\eta_\epsilon'\phi_\epsilon + \eta_\epsilon\psi - \nu_\epsilon')]\,,$$

$$\frac{d^2}{d\epsilon^2}H[\phi_\epsilon, \eta_\epsilon, \nu_\epsilon] = E_Q[\eta_\epsilon''\phi_\epsilon + 2\eta_\epsilon'\psi - \nu_\epsilon'']$$
$$- E_P[(f^*)''(\eta_\epsilon\phi_\epsilon - \nu_\epsilon)(\eta_\epsilon'\phi_\epsilon + \eta_\epsilon\psi - \nu_\epsilon')^2]$$
$$- E_P[(f^*)'(\eta_\epsilon\phi_\epsilon - \nu_\epsilon)(\eta_\epsilon''\phi_\epsilon + 2\eta_\epsilon'\psi - \nu_\epsilon'')]\,.$$

Here we used the boundedness assumptions on f^*, the differentiablity of f^* and the dominated convergence theorem [34] (Theorem 2.27); the latter allowed us to rigorously exchange derivatives and expectations when we calculated the 1st and the 2nd Gateaux derivatives of the objective functionals above. The same argument is also used in the evaluation of all Gateaux derivatives below. Furthermore, for any choice of ψ, ν_ϵ, η_ϵ that have the above properties, the maximum of $\epsilon \to H[\phi_\epsilon, \eta_\epsilon, \nu_\epsilon]$ is achieved at $\epsilon = 0$. Therefore the first derivative vanishes at $\epsilon = 0$. In particular, by considering the case $\eta_\epsilon \equiv 1$, $\nu_\epsilon \equiv 0$ we see that

$$E_P[(f^*)'(\phi^*)\psi] = E_Q[\psi] \tag{113}$$

for all ψ. The second derivative at the maximum is given by

$$\frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0}H[\phi_\epsilon, \eta_\epsilon, \nu_\epsilon] \tag{114}$$
$$=E_Q[\eta_0''\phi^* + 2\eta_0'\psi - \nu_0''] - E_P[(f^*)''(\phi^*)(\eta_0'\phi^* + \psi - \nu_0')^2]$$
$$- E_P[(f^*)'(\phi^*)(\eta_0''\phi^* + 2\eta_0'\psi - \nu_0'')]$$
$$= - E_P[(f^*)''(\phi^*)(\eta_0'\phi^* + \psi - \nu_0')^2]\,,$$

where we used Eq. (113) to cancel several terms. We will specialize this to compute all four Hessians from Eq. (111).

A. LT f - Divergence objective functional

Here we fix $\eta_\epsilon \equiv 1$, $\nu_\epsilon \equiv 0$:

$$H[\phi, 1, 0] = E_Q[\phi] - E_P[f^*(\phi)] \tag{115}$$

and hence

$$
\begin{aligned}
\langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle &= \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} H[\phi_\epsilon, 1, 0] \\
&= \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0}(E_Q[\phi_\epsilon] - E_P[f^*(\phi_\epsilon)]) \\
&= -E_P[(f^*)''(\phi^*)\psi^2].
\end{aligned}
\tag{116}
$$

We have assumed $(f^*)'' > 0$ and so we can define the probability measure

$$
dP^* = (f^*)''(\phi^*)dP/E_P[(f^*)''(\phi^*)]
\tag{117}
$$

and thereby write

$$
\langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle = -E_P[(f^*)''(\phi^*)]E_{P^*}[\psi^2]
\tag{118}
$$

(compare with the KL-case Eq. (29)).

B. Optimization over shifts

Here $\eta_\epsilon \equiv 1$ and ν_ϵ is chosen so that

$$
\sup_\nu \{-\nu - E_P[f^*(\phi_\epsilon - \nu)]\} = -\nu_\epsilon - E_P[f^*(\phi_\epsilon - \nu_\epsilon)].
\tag{119}
$$

By using the convexity of the objective functional in ν along with implicit function theorem, one can see that the assumptions on f are sufficient to ensure that such a smooth ν_ϵ exists for ϵ in a neighborhood of 0. We can simplify

$$
\begin{aligned}
\langle \nabla^2 H_{shift}[\phi^*]\psi, \psi \rangle &= \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \sup_\nu H[\phi_\epsilon, 1, \nu] \\
&= \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} H[\phi_\epsilon, \eta_\epsilon, \nu_\epsilon] = -E_P[(f^*)''(\phi^*)(\psi - \nu_0')^2].
\end{aligned}
\tag{120}
$$

The derivative of ν_ϵ can be computed as follows: Eq. (119) implies

$$
0 = \partial_\nu|_{\nu=\nu_\epsilon}(-\nu - E_P[f^*(\phi_\epsilon - \nu)]) = -1 + E_P[(f^*)'(\phi_\epsilon - \nu_\epsilon)]
\tag{121}
$$

for all ϵ. Differentiating with respect to ϵ gives

$$
0 = E_P[(f^*)''(\phi_\epsilon - \nu_\epsilon)(\psi - \nu_\epsilon')],
\tag{122}
$$

hence

$$
\nu_0' = \frac{E_P[(f^*)''(\phi^*)\psi]}{E_P[(f^*)''(\phi^*)]} = E_{P^*}[\psi]
\tag{123}
$$

where P^* is the probability measure defined in Eq. (117). Therefore

$$\frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \sup_\nu H[\phi_\epsilon, 1, \nu] = -E_P[(f^*)''(\phi^*)]E_{P^*}[(\psi - \nu_0')^2]$$

$$= -E_P[(f^*)''(\phi^*)]\operatorname{Var}_{P^*}[\psi] \tag{124}$$

(compare with the KL-case Eq. (30)).

Recalling the result (118) we arrive at

$$\langle \nabla^2 H_{shift}[\phi^*]\psi, \psi\rangle = \langle \nabla^2 H_{id}[\phi^*]\psi, \psi\rangle + E_P[(f^*)''(\phi^*)\psi]^2/E_P[(f^*)''(\phi^*)]. \tag{125}$$

This shows that the magnitude of the second derivative (which is guaranteed to be non-positive) has been reduced by the amount $E_P[(f^*)''(\phi^*)\psi]^2/E_P[(f^*)''(\phi^*)]$.

C. Optimization over scaling transformations

Here $\nu_\epsilon \equiv 0$ and η_ϵ is chosen so that

$$\sup_\eta\{E_Q[\eta\phi_\epsilon] - E_P[f^*(\eta\phi_\epsilon)]\} = E_Q[\eta_\epsilon\phi_\epsilon] - E_P[f^*(\eta_\epsilon\phi_\epsilon)]. \tag{126}$$

Again, the existence and smoothness of η_ϵ is guaranteed by convexity and the implicit function theorem. Next note that

$$0 = \partial_\eta|_{\eta=\eta_\epsilon}(E_Q[\eta\phi_\epsilon] - E_P[f^*(\eta\phi_\epsilon)]) = E_Q[\phi_\epsilon] - E_P[(f^*)'(\eta_\epsilon\phi_\epsilon)\phi_\epsilon] \tag{127}$$

for all ϵ. Taking the derivative with respect to ϵ gives

$$0 = E_Q[\psi] - E_P[(f^*)''(\phi^*)(\eta_0'\phi^* + \psi)\phi^* + (f^*)'(\phi^*)\psi]. \tag{128}$$

Solving for η_0' and using Eq. (113) to simplify yields

$$\eta_0' = -\frac{E_P[(f^*)''(\phi^*)\psi\phi^*]}{E_P[(f^*)''(\phi^*)(\phi^*)^2]}. \tag{129}$$

Therefore, the Hessian is

$$\langle \nabla^2 H_{scale}[\phi^*]\psi, \psi\rangle = \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \sup_\eta H[\phi_\epsilon, \eta, 0] \tag{130}$$

$$= -E_P[(f^*)''(\phi^*)(\eta_0'\phi^* + \psi)^2]$$

$$= \langle \nabla^2 H_{id}[\phi^*]\psi, \psi\rangle - (\eta_0')^2 E_P[(f^*)''(\phi^*)(\phi^*)^2] - 2\eta_0' E_P[(f^*)''(\phi^*)\phi^*\psi]$$

$$= \langle \nabla^2 H_{id}[\phi^*]\psi, \psi\rangle + \frac{(E_P[(f^*)''(\phi^*)\psi\phi^*])^2}{E_P[(f^*)''(\phi^*)(\phi^*)^2]}.$$

Once again, the magnitude of the second derivative has been reduced.

D. *Optimization over affine transformations*

Finally, let η_ϵ and ν_ϵ be defined by

$$\sup_{\eta,\nu}\{E_Q[\eta\phi_\epsilon - \nu] - E_P[f^*(\eta\phi_\epsilon - \nu)] = E_Q[\eta_\epsilon\phi_\epsilon - \nu_\epsilon] - E_P[f^*(\eta_\epsilon\phi_\epsilon - \nu_\epsilon)]\,. \tag{131}$$

The existence and smoothness of η_ϵ and ν_ϵ are again guaranteed by convexity and the implicit function theorem. They satisfy

$$\partial_\nu|_{\nu=\nu_\epsilon}(E_Q[\eta_\epsilon\phi_\epsilon - \nu] - E_P[f^*(\eta_\epsilon\phi_\epsilon - \nu)]) = 0\,, \tag{132}$$

$$\partial_\eta|_{\eta=\eta_\epsilon}(E_Q[\eta\phi_\epsilon - \nu_\epsilon] - E_P[f^*(\eta\phi_\epsilon - \nu_\epsilon)]) = 0$$

for all ϵ. Simplifying this and differentiating with respect to ϵ we obtain the two equations

$$0 = E_P[(f^*)''(\phi^*)\phi^*]\eta_0' - E_P[(f^*)''(\phi^*)]\nu_0' + E_P[(f^*)''(\phi^*)\psi]\,, \tag{133}$$

$$0 = -E_P[(f^*)''(\phi^*)(\phi^*)^2]\eta_0' + E_P[(f^*)''(\phi^*)\phi^*]\nu_0' - E_P[(f^*)''(\phi^*)\phi^*\psi]\,.$$

If we define

$$a = E_P[(f^*)''(\phi^*)\phi^*]\,, \quad b = E_P[(f^*)''(\phi^*)]\,, \tag{134}$$

$$c = E_P[(f^*)''(\phi^*)(\phi^*)^2]\,, \quad g = E_P[(f^*)''(\phi^*)\psi]\,,$$

$$h = E_P[(f^*)''(\phi^*)\phi^*\psi]\,,$$

then these have the solution

$$\begin{bmatrix} \eta_0' \\ \nu_0' \end{bmatrix} = \frac{1}{bc - a^2} \begin{bmatrix} a & -b \\ c & -a \end{bmatrix} \begin{bmatrix} g \\ h \end{bmatrix} \tag{135}$$

(note that $a^2 < bc$ follows from the Cauchy Schwarz inequality together with the assumptions that f' is strictly increasing and $Q \neq P$). We can compute

$$\langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle$$

$$= \frac{d^2}{d\epsilon^2}\Big|_{\epsilon=0} \sup_{\eta,\nu\in\mathbb{R}} H[\phi_\epsilon, \eta, \nu] = -E_P[(f^*)''(\phi^*)(\eta_0'\phi^* + \psi - \nu_0')^2]$$

$$= \langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle - (\eta_0')^2 c - 2\eta_0' h + 2\eta_0'\nu_0' a + 2\nu_0' g - (\nu_0')^2 b$$

$$= \langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle + \frac{1}{bc - a^2} \begin{bmatrix} h & g \end{bmatrix} \begin{bmatrix} b & -a \\ -a & c \end{bmatrix} \begin{bmatrix} h \\ g \end{bmatrix}\,. \tag{136}$$

We have $b + c \geq 0$ and $bc - a^2 > 0$ hence the matrix in Eq. (136) is positive semi-definite. Therefore the gain term is non-negative and the second derivative is reduced in magnitude, as

compared to the unoptimized objective functional (118). Finally, we can rewrite the terms a, c, d in (136) in terms of $b = E_P[(f^*)''(\phi^*)]$ and the tilted measure (117):

$$a = E_P[(f^*)''(\phi^*)\phi^*] = b \cdot E_{P^*}[\phi^*]\,, \tag{137}$$

$$c = E_P[(f^*)''(\phi^*)(\phi^*)^2] = b \cdot E_{P^*}[(\phi^*)^2]\,,$$

$$g = E_P[(f^*)''(\phi^*)\psi] = b \cdot E_{P^*}[\psi]\,,$$

$$h = E_P[(f^*)''(\phi^*)\phi^*\psi] = b \cdot E_{P^*}[\phi^*\psi]\,.$$

Then (136) becomes

$$\langle \nabla^2 H_{affine}[\phi^*]\psi, \psi \rangle = \langle \nabla^2 H_{id}[\phi^*]\psi, \psi \rangle + b \cdot \left[\frac{\mathrm{Cov}_{P^*}(\phi^*, \psi)^2}{\mathrm{Var}_{P^*}[\phi^*]} + (E_{P^*}[\psi])^2 \right]. \tag{138}$$

■

APPENDIX E

EXPONENTIAL FAMILIES AND THE MANIFOLD OF SUFFICIENT STATISTICS

Suppose $P = P_{\theta_p}$ and $Q = P_{\theta_q}$ are members of the same exponential family $dP_\theta = h(x)e^{\kappa(\theta) \cdot T(x) - \beta(\theta)}$, $\theta \in \Theta$ with $T : \Omega \to \mathbb{R}^n$ the vector of sufficient statistics. Then (under appropriate assumptions) we have

$$D_f(Q\|P) = \sup_{\phi \in L^1(Q)} \{ E_Q[\phi] - E_P[f^*(\phi)] \} = E_Q[\phi^*] - E_P[f^*(\phi^*)]\,, \tag{139}$$

and the explicit optimizer $\phi^* = f'(dQ/dP)$ lies on a (generally nonlinear) $(n+1)$-dimensional manifold of functions, parameterized by the sufficient statistics and constants:

$$\phi_{(\beta,\kappa)} = f'(\exp(\kappa \cdot T + \beta)), \quad (\beta, \kappa) \in \mathbb{R} \times \mathbb{R}^n\,. \tag{140}$$

Therefore, computing the f-divergence reduces to the following finite-dimensional optimization problem:

$$D_f(Q\|P) = \sup_{(\beta,\kappa) \in \mathbb{R}^{n+1}} \{ E_Q[\phi_{(\beta,\kappa)}] - E_P[f^*(\phi_{(\beta,\kappa)})] \}\,. \tag{141}$$

In some cases one can further reduce the dimension by optimizing over an appropriate family of transformations.

1) For KL divergence, we have

$$\phi_{(\beta,\kappa)} = \log(\exp(\kappa \cdot T + \beta)) + 1 = \kappa \cdot T + (\beta + 1)\,, \tag{142}$$

and so the optimizer lies on a linear manifold. If one optimizes over the family of shifts (i.e., one uses the Donsker-Varadhan formula) then the $\beta + 1$ factor is eliminated and one finds that

$$D_{KL}(Q\|P) = \sup_{\kappa \in \mathbb{R}^n} \{E_Q[\kappa \cdot T] - \log E_P[e^{\kappa \cdot T}]\} \tag{143}$$

for any Q, P that are members of the exponential family, i.e., the optimization is over the n-dimensional subspace spanned by the sufficient statistics: $\phi_\kappa = \kappa \cdot T, \ \ \kappa \in \mathbb{R}^n$. Also, note that $\kappa \to E_Q[\kappa \cdot T] - \log E_P[e^{\kappa \cdot T}]$ is concave.

2) For α-divergences, the maximizer of the scaling-improved variational formula (17) lies on the manifold $\phi_\kappa = \exp((\alpha - 1)\kappa \cdot T)$, hence

$$D_{f_\alpha}(Q\|P) = \sup_{\kappa \in \mathbb{R}^n} \left\{ \frac{1}{\alpha(\alpha - 1)} E_Q[\phi_\kappa]^\alpha E_P[\phi_\kappa^{\alpha/(\alpha-1)}]^{-(\alpha-1)} \right\} - \frac{1}{\alpha(\alpha - 1)}$$

for any members Q, P, of the exponential family.

Appendix F
Asymptotic Variance of the Shift-Optimized α-Divergence Objective Functional

In this section we provide details regarding the computation of asymptotic variance of the objective functional estimators for α-divergences (see Theorem 3), generalizing some of the work on KL divergences from [35]. The main tool is the following lemma, which is based on the delta method.

Lemma 3: Let $h : \mathbb{R}^d \to \mathbb{R}$ be C^1 and X_n be iid $\mathbb{R}^d$-valued random variables with mean μ, covariance Σ, and $\sup_n \|X_n\| < \infty$. Define

$$E_n[X] = \frac{1}{n} \sum_{i=1}^n X_i. \tag{144}$$

Then

$$E[\sqrt{n}(h(E_n[X]) - h(\mu))] \to 0 \tag{145}$$

and

$$\mathrm{Var}[\sqrt{n}(h(E_n[X]) - h(\mu))] \to \nabla h(\mu) \cdot \Sigma \cdot \nabla h(\mu). \tag{146}$$

Proof: The central limit theorem implies

$$\sqrt{n}(E_n[X] - \mu) \xrightarrow{D} N(0, \Sigma). \tag{147}$$

The delta method (see, e.g., Theorem 5.15 in [44]) then implies

$$\sqrt{n}\left(h(E_n[X]) - h(\mu)\right) \xrightarrow{D} N(0, \nabla h(\mu) \cdot \Sigma \cdot \nabla h(\mu)). \tag{148}$$

Convergence of the mean and variance will follow from (148) if we can show uniform integrability of the random variables $|\sqrt{n}(h(E_n[X]) - h(\mu))|^2$; see, e.g., Theorem 5.9 in [45]. The quantities $E_n[X]$ and μ valued in $B_M(0)$, where M is a bound on $\sup_n \|X_n\|$. The C^1 function h is therefore Lipschitz on this ball; let L denote the Lipschitz constant. We can now compute

$$\sup_n \int 1_{|\sqrt{n}(h(E_n[X])-h(\mu))|^2 > c} |\sqrt{n}(h(E_n[X]) - h(\mu))|^2 dP \tag{149}$$

$$\leq L^2 \sup_n n \int 1_{nL^2 \|E_n[X]-\mu\|^2 > c} \|E_n[X] - \mu\|^2 dP$$

$$\leq L^2 \sup_n n E[\|E_n[X] - \mu\|^4]^{1/2} P(nL^2 \|E_n[X] - \mu\|^2 > c)^{1/2}$$

$$\leq L^2 \sup_n n E[\|E_n[X] - \mu\|^4]^{1/2}$$

$$\times \left(\sum_{k=1}^d P\left(|E_n[X^k] - \mu^k| > \frac{c^{1/2}}{n^{1/2}d^{1/2}L} \right) \right)^{1/2}.$$

Hoeffding's inequality (see, e.g., Theorem 2.8 in [46]) implies

$$P\left(|E_n[X^k] - \mu^k| > \frac{c^{1/2}}{n^{1/2}d^{1/2}L} \right) \leq 2 \exp\left(-\frac{n}{2M^2} \left(\frac{c^{1/2}}{n^{1/2}d^{1/2}L} \right)^2 \right) \tag{150}$$

$$= 2 \exp\left(-\frac{c}{2dM^2L^2} \right).$$

Therefore

$$\sup_n \int 1_{|\sqrt{n}(h(E_n[X])-h(\mu))|^2 > c} |\sqrt{n}(h(E_n[X]) - h(\mu))|^2 dP \tag{151}$$

$$\leq 2^{1/2} d^{1/2} L^2 \exp\left(-\frac{c}{4dM^2L^2} \right) \sup_n E[n^2 \|E_n[X] - \mu\|^4]^{1/2}.$$

If $\sup_n E[n^2 \|E_n[X] - \mu\|^4] < \infty$ then the limit of (151) as $c \to \infty$ equals zero and we are done. Expanding this expression and using the assumption that the X_i's are iid we find, after a somewhat long but straightforward calculation,

$$E[n^2 \|E_n[X] - \mu\|^4] = \frac{1}{n} E[\|X_1 - \mu\|^4] + (1 - 1/n) E[\|X_1 - \mu\|^2]^2 \tag{152}$$

$$+ 2 \sum_{j,k=1}^d (1 - 1/n) E[(X_1^j - \mu^j)(X_1^k - \mu^k)]^2,$$

which is bounded above uniformly in n. This completes the proof. ∎

We now prove the asymptotic variance result for α-divergences, Theorem 3.

Proof: First, we can rewrite the variance as

$$\alpha^2(\alpha-1)^2 n \operatorname{Var}[\widehat{H}_{scale,f_\alpha}[\phi; Q_n, P_n]] = \operatorname{Var}[\sqrt{n}h(E_{Q_n}[\phi], E_{P_n}[\psi_\alpha])] \tag{153}$$

$$= \operatorname{Var}[\sqrt{n}(h(E_{Q_n}[\phi], E_{P_n}[\psi_\alpha]) - h(\mu))],$$

where

$$h(x,y) \equiv x^\alpha y^{-(\alpha-1)} - 1, \quad \mu \equiv (E_Q[\phi], E_P[\psi_\alpha]). \tag{154}$$

The assumed bounds on ϕ imply that (ϕ, ψ_α) are valued in a compact convex set and that h has a C^1 extension from this set to all of $\mathbb{R}^2$. Let $E_n[X] = (E_{Q_n}[\phi], E_{P_n}[\psi_\alpha])$. Note that $E_n[X]$ has mean μ and covariance $\Sigma = \operatorname{diag}(\operatorname{Var}_Q[\phi], \operatorname{Var}_P[\psi_\alpha])$ (we independently sample from Q and P). The assumptions of Lemma 3 are now satisfied, and so

$$\alpha^2(\alpha-1)^2 n \operatorname{Var}[\widehat{H}_{scale,f_\alpha}[\phi; Q_n, P_n]] \to \nabla h(\mu) \cdot \Sigma \cdot \nabla h(\mu). \tag{155}$$

Simplifying this gives the claimed result (39). ∎

ACKNOWLEDGMENT

REFERENCES

[1] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1961, pp. 547–561. [Online]. Available: https://projecteuclid.org/euclid.bsmsp/1200512181

[2] 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)*, vol. 28, no. 1, pp. 131–142, 1966. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1966.tb00626.x

[3] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time. IV," *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, 1983. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160360204

[4] P. Dupuis and R. Ellis., *A Weak Convergence Approach to the Theory of Large Deviations*, ser. Wiley series in probability and statistics. New York: John Wiley & Sons, 1997, a Wiley-Interscience Publication. [Online]. Available: http://opac.inria.fr/record=b1092351

[5] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. [Online]. Available: https://books.google.gr/books?id=hV8o5R7_5tkC

[6] C. Kipnis and C. Landim, *Scaling Limits of Interacting Particle Systems*. Springer-Verlag, 1999.

[7] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Nonparametric estimation of the likelihood ratio and divergence functionals," in *2007 IEEE International Symposium on Information Theory*, 2007, pp. 2016–2020.

[8] ——, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[9] 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 Coference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, 2012, p. 1155–1162.

[10] 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*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 531–540. [Online]. Available: http://proceedings.mlr.press/v80/belghazi18a.html

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[12] S. Nowozin, B. Cseke, and R. Tomioka, "F-GAN: Training Generative Neural Samplers Using Variational Divergence Minimization," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 271–279.

[13] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 214–223.

[14] K. Chowdhary and P. Dupuis, "Distinguishing and integrating aleatoric and epistemic variation in uncertainty quantification," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 47, no. 3, p. 635–662, 2013.

[15] 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*, vol. 4, no. 1, pp. 80–111, 2016.

[16] D. A. McAllester, "PAC-Bayesian model averaging," in *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, ser. COLT '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 164–170. [Online]. Available: https://doi.org/10.1145/307400.307435

[17] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, ser. COLT '97. New York, NY, USA: Association for Computing Machinery, 1997, p. 2–9. [Online]. Available: https://doi.org/10.1145/267460.267466

[18] O. Catoni, P. Euclid, C. U. Library, and D. U. Press, *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, ser. Lecture notes-monograph series. Cornell University Library, 2008. [Online]. Available: https://books.google.gr/books?id=-EtrnQAACAAJ

[19] M. Broniatowski and A. Keziou, "Parametric estimation and tests through divergences and the duality

technique," *Journal of Multivariate Analysis*, vol. 100, no. 1, pp. 16 – 36, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0047259X08001036

[20] ——, "Minimization of divergences on sets of signed measures," *Studia Scientiarum Mathematicarum Hungarica*, vol. 43, no. 4, p. 403–442, 2006.

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

[22] R. Dudley, R. Dudley, B. Bollobas, W. Fulton, A. Katok, F. Kirwan, P. Sarnak, and B. Simon, *Uniform Central Limit Theorems*, ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999. [Online]. Available: https://books.google.com/books?id=URvbWBtY4XQC

[23] A. Ben-Tal and M. Teboulle, "An old-new concept of convex risk measures: The optimized certainty equivalent," *Mathematical Finance*, vol. 17, no. 3, pp. 449–476, 2007. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9965.2007.00311.x

[24] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964. [Online]. Available: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1964.tb00553.x

[25] J. Song and S. Ermon, "Bridging the gap between f-GANs and Wasserstein GANs," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 9078–9087. [Online]. Available: https://proceedings.mlr.press/v119/song20a.html

[26] R. Keener, *Theoretical Statistics: Topics for a Core Course*, ser. Springer Texts in Statistics. Springer New York, 2010. [Online]. Available: https://books.google.com/books?id=aVJmcega44cC

[27] C. Tao, L. Chen, R. Henao, J. Feng, and L. C. Duke, "Chi-square generative adversarial network," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 4887–4896. [Online]. Available: http://proceedings.mlr.press/v80/tao18b.html

[28] T. Minka, "Technical Report MSR-TR-2005-173," Microsoft Research, Tech. Rep., 2005.

[29] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *J Stat Phys*, vol. 52, p. 479–487, 1988.

[30] S. Amari, *Information Geometry and Its Applications*, ser. Applied Mathematical Sciences. Springer Japan, 2016. [Online]. Available: https://books.google.com/books?id=UkSFCwAAQBAJ

[31] J. Birrell, P. Dupuis, M. A. Katsoulakis, L. Rey-Bellet, and J. Wang, "Variational representations and neural network estimation of Rényi divergences," *SIAM Journal on Mathematics of Data Science*, vol. 3, no. 4, pp. 1093–1116, 2021. [Online]. Available: https://doi.org/10.1137/20M1368926

[32] Y. Pantazis, D. Paul, M. Fasoulakis, Y. Stylianou, and M. A. Katsoulakis, "Cumulant GAN," *arXiv:2006.06625*, 2020.

[33] R. Atar, K. Chowdhary, and P. Dupuis, "Robust bounds on risk-sensitive functionals via Rényi divergence," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 3, no. 1, pp. 18–33, 2015.

[34] G. B. Folland, *Real analysis:*, 2nd ed. New York:: Wiley,, 1999.

[35] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x62TNtDS

[36] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015. [Online]. Available: http://dx.doi.org/10.1561/2200000050

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[38] E. Weinan, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden, "Heterogeneous multiscale methods: A review." *Communications in Computational Physics*, 2007.

[39] S. Rahman, "The f-sensitivity index," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 130–162, 2016.

[40] Y. LeCun, "The MNIST database of handwritten digits," 1998. [Online]. Available: http://yann.lecun.com/exdb/mnist/

[41] M. S. Shell, "The relative entropy is fundamental to multiscale and inverse thermodynamic problems," *The Journal of Chemical Physics*, vol. 129, no. 14, p. 144108, 2008.

[42] A. J. Majda and B. Gershgorin, "Improving model fidelity and sensitivity for complex systems through empirical information theory," *Proceedings of the National Academy of Sciences*, vol. 108, no. 25, pp. 10 044–10 049, 2011. [Online]. Available: https://www.pnas.org/content/108/25/10044

[43] M. A. Katsoulakis and P. Plechac, "Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems," *Journal of Chemical Physics*, vol. 139, no. 7, Aug 21 2013.

[44] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, ser. Springer Texts in Statistics. Springer New York, 2013. [Online]. Available: https://books.google.com/books?id=qrcuBAAAQBAJ

[45] A. Gut, *Probability: A Graduate Course*, ser. Springer Texts in Statistics. Springer New York, 2006. [Online]. Available: https://books.google.com/books?id=CnBxaLaOcBAC

[46] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. [Online]. Available: https://books.google.com/books?id=H6QmKBcLlgkC