SIAM J. MATH. DATA SCI.
Vol. 3, No. 4, pp. 1093–1116

Variational Representations and Neural Network Estimation of Rényi Divergences*

Jeremiah Birrell†, Paul Dupuis‡, Markos A. Katsoulakis§, Luc Rey-Bellet§, and Jie Wang§

Abstract. We derive a new variational formula for the Rényi family of divergences, $R_\alpha(Q\|P)$, between probability measures Q and P. Our result generalizes the classical Donsker–Varadhan variational formula for the Kullback–Leibler divergence. We further show that this Rényi variational formula holds over a range of function spaces; this leads to a formula for the optimizer under very weak assumptions and is also key in our development of a consistency theory for Rényi divergence estimators. By applying this theory to neural network estimators, we show that if a neural network family satisfies one of several strengthened versions of the universal approximation property, then the corresponding Rényi divergence estimator is consistent. In contrast to density estimator based methods, our estimators involve only expectations under Q and P and hence are more effective in high dimensional systems. We illustrate this via several numerical examples of neural network estimation in systems of up to 5,000 dimensions.

Key words. Rényi divergence, variational representation, neural network estimator

AMS subject classifications. 94A17, 62B10, 62G05

DOI. 10.1137/20M1368926

1. Introduction.

Information-theoretic divergences are widely used to quantify the notion of "distance" between probability measures Q and P; commonly used examples include the Kullback–Leibler divergence (i.e., KL-divergence or relative entropy), f-divergences, and Rényi divergences. The computation and estimation of divergences is important in many applications, including independent component analysis [25], medical image registration [36], feature selection [30], genomic clustering [12], the information bottleneck method [51], independence testing [29], and the analysis and design of generative adversarial networks (GANs) [23, 38, 3, 24, 40].

Estimation of divergences from data is known to be a difficult problem [39, 19]. Density estimator based methods such as those in [43, 26] are known to work best in low dimensions.

*Received by the editors October 6, 2020; accepted for publication (in revised form) July 23, 2021; published electronically October 6, 2021.
https://doi.org/10.1137/20M1368926
Funding: The research of the first, third, and fourth authors was partially supported by NSF TRIPODS CISE-1934846. The research of the third and fourth authors was also partially supported by the National Science Foundation (NSF) under the grant DMS-2008970 and by the Air Force Office of Scientific Research (AFOSR) under the grant FA-9550-18-1-0214. The research of the second author was supported in part by the NSF under the grant DMS-1904992 and by the AFOSR under the grant FA-9550-18-1-0214. The research of the fifth author was partially supported by the Defense Advanced Research Projects Agency (DARPA) EQUiPS program under the grant W911NF1520122.

†TRIPODS Institute for Theoretical Foundations of Data Science, University of Massachusetts Amherst, Amherst, MA 01003 USA (birrell@math.umass.edu).
‡Division of Applied Mathematics, Brown University, Providence, RI 02912 USA (dupuis@dam.brown.edu).
§Department of Mathematics and Statistics, University of Massachusetts Amherst, Amherst, MA 01003 USA (markos@math.umass.edu, luc@math.umass.edu, wang@math.umass.edu).

However, recent work has shown that variational representations of divergences can be used to construct statistical estimators for the KL-divergence [7], and more general f-divergences [37, 48, 9], which scale better with dimension. The family of Rényi divergences, first introduced in [46], provides means of quantifying the discrepancy between two probability measures that are especially sensitive to the relative tail behavior of the distributions. Rényi divergences are used in variational inference [32] and uncertainty quantification for rare events [17] and naturally arise in coding theory and hypothesis testing (see [53] for further discussion and references). Rényi divergences have several advantages over the commonly used KL-divergence, including the ability to compare heavy-tailed distributions and certain nonabsolutely continuous distributions. In addition, the estimation of KL-divergence can suffer from stability issues, due to the impact of rare events as well as high variance [50]—problems that we empirically find to be less pronounced for certain Rényi divergences (see the example in section 5.1 below). In this work we develop a new variational characterization for the family of Rényi divergences, $R_\alpha(Q\|P)$, and study its use in statistical estimation. More specifically, we will prove

$$R_\alpha(Q\|P) = \sup_{g\in\Gamma}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g}dP\right]\right\}, \tag{1.1}$$

where $\alpha \in \mathbb{R}$, $\alpha \neq 0, 1$, and Γ is an appropriate function space; see Theorem 3.1 below. Equation (1.1) can be viewed as an extension of the well-known Donsker–Varadhan variational formula for the relative entropy [14, 16],

$$R(Q\|P) = \sup_{g\in\mathcal{M}_b(\Omega)}\left\{\int g dQ - \log\left[\int e^g dP\right]\right\}, \tag{1.2}$$

where $\mathcal{M}_b(\Omega)$ denotes the set of bounded measurable real-valued functions on Ω. Note that (1.1) generalizes (1.2) in two directions; we generalize both the divergence, $R(Q\|P) \to R_\alpha(Q\|P)$, and the function space, $\mathcal{M}_b(\Omega) \to \Gamma$; allowed Γ's are given in Theorem 3.1, Corollary 3.2, and Lemma 4.3 below. The flexibility in choosing Γ allows us to derive a formula for the optimizer of (1.1) under very weak assumptions (see Corollary 3.2) and is also key in our development of consistent statistical estimators (see Lemma 4.3 and Theorem 4.6).

The objective functional in the optimization problem (1.1) depends on Q and P only through the expectation of certain functions of g. As a result, the objective functional can be estimated in a straightforward manner using only samples from Q and P. This property makes (1.1) a powerful tool in the construction of statistical estimators for Rényi divergences. In section 4 we provide a general framework for proving consistency of Rényi divergence estimators that are based on (1.1). In section 4.1 we apply this theory to show consistency of neural network estimators. Related methods were used to prove consistency of KL-divergence estimators in [7], though under stronger assumptions. Here we contribute a set of new technical tools that allow for a consistency proof in important cases where the prior theory did not apply—specifically when the measures Q and P have noncompact support and are light-tailed, and for neural network estimators with unbounded activation function, such as the widely used ReLU activation. Our new method involves the use of the Tietze extension theorem and new strengthened versions of the universal approximation property (see Definitions 4.1 and 4.2) to vary the function space, Γ, in the variational formula (1.1) (see Lemma 4.3) and

finally culminates in the consistency result, Theorem 4.6. Function spaces of neural networks that satisfy the required assumptions are provided in section 4.1 and are discussed further in the supplementary materials file Supplement.pdf [local/web 296KB]. Finally, in section 5 we demonstrate the effectiveness of the Rényi divergence estimators in numerical examples with systems of up to 5000 dimensions.

1.1. Related work. Our main result (1.1) can be viewed as a dual variational formula to the result in [5], generalizing the duality between the Donsker–Varadhan and Gibbs variational principles. An alternative variational formula for the Rényi divergences, using an objective functional that is a linear combination of relative entropies, can be found in Theorem 30 of [53] and also in Theorem 1 of [1]. As discussed above, our result (1.1) is advantageous for the purpose of statistical estimation, as the objective functional is straightforward to estimate using only samples from P and Q. This property was key in the use of (1.2) for the statistical estimation of KL-divergence and applications to GANs in [7], and we will similarly take advantage of this property for Rényi divergence estimation. In addition, our results on neural network estimation in section 4 provide theoretical underpinnings for cumulant GAN [40]. Finally, we note that a variational formula for quantum Rényi entropies was previously derived in [8] and agrees with (1.1) in the commutative, discrete setting.

2. Background on Rényi divergences. The Rényi divergence of order $\alpha \in (0, \infty)$, $\alpha \neq 1$, between two probability measures Q and P on a measurable space $(\Omega, \mathcal{M})$, denoted $R_\alpha(Q\|P)$, can be defined as follows: Let ν be a sigma-finite positive measure with $dQ = q d\nu$ and $dP = p d\nu$. Then

$$R_\alpha(Q\|P) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \log\left[\int_{p>0} q^\alpha p^{1-\alpha} d\nu\right] & \text{if } 0 < \alpha < 1 \text{ or} \\ & \alpha > 1 \text{ and } Q \ll P, \\ +\infty & \text{if } \alpha > 1 \text{ and } Q \not\ll P. \end{cases} \tag{2.1}$$

Such a ν always exists (e.g., $\nu = Q + P$), and it can be shown that the definition (2.1) does not depend on the choice of ν. The R_α satisfy the following divergence property: $R_\alpha(Q\|P) \geq 0$ with equality if and only if $Q = P$. In this sense, the Rényi divergences provide a notion of "distance" between probability measures. Note, however, that Rényi divergences are not symmetric, but rather they satisfy

$$R_\alpha(Q\|P) = R_{1-\alpha}(P\|Q), \quad \alpha \in (0, 1). \tag{2.2}$$

Equation (2.2) is used to extend the definition of $R_\alpha(Q\|P)$ to $\alpha < 0$. Rényi divergences are connected to the KL-divergence, $R(Q\|P)$, through the limiting formulas

$$\lim_{\alpha \to 1^-} R_\alpha(Q\|P) = R(Q\|P), \tag{2.3}$$

and if $R(Q\|P) = \infty$ or if $R_\beta(Q\|P) < \infty$ for some $\beta > 1$, then

$$\lim_{\alpha \to 1^+} R_\alpha(Q\|P) = R(Q\|P). \tag{2.4}$$

See [53] for a detailed discussion of Rényi divergences and proofs of these (and many other) properties. Note, however, that our definition of the Rényi divergences is related to theirs by

$D_\alpha(\cdot\|\cdot) = \alpha R_\alpha(\cdot\|\cdot)$. Explicit formulas for the Rényi divergence between members of many common parametric families can be found in [22]. Rényi divergences are also connected with the family of f-divergences; see [34].

3. Variational formula for the Rényi divergences. The key result in the paper is the following variational characterization of the Rényi divergences, which generalizes the Donsker–Varadhan variational formula (1.2). The proof of this theorem can be found in section 6.1.

Theorem 3.1 (Rényi–Donsker–Varadhan variational formula). *Let P and Q be probability measures on $(\Omega, \mathcal{M})$ and $\alpha \in \mathbb{R}$, $\alpha \neq 0, 1$. Then for any set of functions, Γ, with $\mathcal{M}_b(\Omega) \subset \Gamma \subset \mathcal{M}(\Omega)$ (where $\mathcal{M}(\Omega)$ denotes the set of all real-valued measurable functions on Ω), we have*

$$R_\alpha(Q\|P) = \sup_{g\in\Gamma}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g}dP\right]\right\}, \tag{3.1}$$

where we interpret $\infty - \infty \equiv -\infty$ and $-\infty + \infty \equiv -\infty$.

If in addition $(\Omega, \mathcal{M})$ is a metric space with the Borel σ-algebra, then (3.1) holds for all Γ that satisfy $\mathrm{Lip}_b(\Omega) \subset \Gamma \subset \mathcal{M}(\Omega)$, where $\mathrm{Lip}_b(\Omega)$ denotes the space of bounded Lipschitz functions on Ω (we emphasize that the Lipschitz constant is allowed to take any finite value).

Corollary 3.2 (Existence of an optimizer). *Let $\alpha \in \mathbb{R}$, $\alpha \neq 0, 1$, and suppose $Q \ll P$, $dQ/dP > 0$, $(dQ/dP)^\alpha \in L^1(P)$. Define $g^* = \log(dQ/dP)$ and suppose Γ is a function space that satisfies $g^* \in \Gamma \subset \mathcal{M}(\Omega)$. Then (3.1) holds and the supremum is achieved at g^*.*

The ability to vary the function space in (3.1) has several important consequences.

1. Taking $\Gamma = \mathcal{M}(\Omega)$, or some other appropriate set of unbounded functions, implies that one can use unbounded activation functions (e.g., ReLU) in neural network estimators of Rényi divergences; see section 4.1.
2. For certain activation functions, taking $\Gamma = \mathrm{Lip}_b(\Omega)$ is key to proving the consistency of neural network estimators based on (3.1); see the third example in section 4.1 along with the supplementary materials file Supplement.pdf [local/web 296KB].
3. The ability to consider unbounded functions allows for existence of an optimizer under very general assumptions; see Corollary 3.2. In some cases, the existence of an optimizer can be used to reduce the optimization to a finite dimensional problem; see section 3.1 below.

One can formally obtain the classical Donsker–Varadhan variational formula (1.2) by letting $\Gamma = \mathcal{M}_b(\Omega)$ and taking $\alpha \to 1$ in (3.1). Similarly, taking $\alpha \to 0$ and reindexing $g \to -g$, one obtains the Donsker–Varadhan variational formula for $R(P\|Q)$. Rigorously, the extension of the Donsker–Varadhan variational formula to Γ with $\mathcal{M}_b(\Omega) \subset \Gamma \subset \mathcal{M}(\Omega)$ follows from (1.2) together with Theorem 1 in [7]. The generalization to $\mathrm{Lip}_b(\Omega) \subset \Gamma \subset \mathcal{M}(\Omega)$ can be proven via the same method we use for Rényi divergences (see (6.17)–(6.19) and the surrounding discussion). This is a new result to the best of our knowledge; we omit the details.

Remark 3.3. Note that the conventions regarding infinities in Theorem 3.1 are simply convenient shorthand that allow us to consider arbitrary unbounded functions. If one wishes to avoid infinities in the objective functional, then the optimization can be restricted to

$$\widetilde{\Gamma} \equiv \{g \in \Gamma : \exp((\alpha-1)g) \in L^1(Q), \exp(\alpha g) \in L^1(P)\}, \tag{3.2}$$

and the equality (3.1) will still hold.

3.1. Variational formula for the Rényi divergences: Exponential families. If P and Q are members of a parametric family, then, by using the formula for the optimizer $g^* = \log(dQ/dP)$, the function space Γ can be further reduced to a finite dimensional manifold of functions (here we assume the conditions from Corollary 3.2 that ensure the existence of g^*). In particular, if $P = \mu_{\theta_p}$ and $Q = \mu_{\theta_q}$ are members of the same exponential family $d\mu_\theta = h(x)e^{\kappa(\theta)\cdot T(x)-\beta(\theta)}\mu(dx)$, $\theta \in \Theta$, with $T : \Omega \to \mathbb{R}^k$ the vector of sufficient statistics and μ a σ-finite positive measure, then the optimizer g^* lies in the $(k+1)$-dimensional subspace of functions

$$g_{(\Delta\kappa,\Delta\beta)} \equiv \Delta\kappa \cdot T - \Delta\beta\,, \quad (\Delta\kappa, \Delta\beta) \in \mathbb{R}^{k+1}\,. \tag{3.3}$$

Computation of the Rényi divergence therefore reduces to the following k-dimensional optimization problem (note that the Rényi objective functional is invariant under shifts, and so the $\Delta\beta$ terms cancel):

$$R_\alpha(Q\|P) = \sup_{\Delta\kappa\in\mathbb{R}^k} \left\{ \frac{1}{\alpha-1}\log\int e^{(\alpha-1)\Delta\kappa\cdot T}dQ - \frac{1}{\alpha}\log\int e^{\alpha\Delta\kappa\cdot T}dP \right\}. \tag{3.4}$$

Contrast this with an alternative parametric approach, wherein one estimates θ_p and θ_q using maximum likelihood estimation and then uses the explicit formula for the Rényi divergence between members of an exponential family found in Chapter 2 in [33],

$$R_\alpha(Q\|P) = \frac{1}{\alpha(\alpha-1)}\log\left(\frac{Z(\alpha\theta_q + (1-\alpha)\theta_p)}{Z(\theta_p)^{1-\alpha}Z(\theta_q)^\alpha}\right), \quad \alpha > 0\,,\ \alpha \neq 1\,, \tag{3.5}$$

where $Z(\theta) \equiv \exp(\beta(\theta)) = \int h(x)e^{\kappa(\theta)\cdot T(x)}\mu(dx)$ is the partition function. Using (3.5) to estimate the Rényi divergence from data requires the solution of two optimization problems (one each to find maximum likelihood estimators for θ_q and θ_p) and then the computation of three partition functions. Even if one uses a more sophisticated method such as thermodynamic integration (see [31]) to compute the partition functions in (3.5), there is still the challenge of generating data from $\mu_{\alpha\theta_q+(1-\alpha)\theta_p}$, which is required to address the partition function in the numerator of (3.5). These challenges are absent when using (3.4), which only requires the solution of one optimization problem and can be estimated directly using samples from Q and P; one does not need to generate samples from any auxiliary distribution. Therefore, we only expect (3.5) to be preferable in simpler cases where the partition function can be computed analytically. We illustrate the use of (3.4) to estimate Rényi divergences in section 5.3.

4. Statistical estimation of Rényi divergences. We now discuss how the variational formula (3.1) can be used to construct statistical estimators for Rényi divergences. The estimation of divergences in high dimensions is a difficult but important problem, e.g., for independence testing [29] and the development of GANs [23, 38, 3, 24, 40]. Density estimator based methods for estimating divergences are known to be effective primarily in low dimensions (see [43, 26] as well as Figure 1 in [7] and further references therein). In contrast, variational methods for KL- and f-divergences have proven effective in a range of medium and

high-dimensional systems [7, 9]. It should be noted that high-dimensional problems still pose a considerable challenge in general; this is due in part to the problem of sampling rare events. However, existing Monte Carlo methods for sampling rare events (see, e.g., [47, 10, 11]) are still applicable here.

The variational formula (3.1) naturally suggests estimators of the form

$$\widehat{R}_\alpha^{n,k}(Q\|P) \equiv \sup_{\phi\in\Phi_k}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi}dQ_n\right]-\frac{1}{\alpha}\log\left[\int e^{\alpha\phi}dP_n\right]\right\}, \tag{4.1}$$

where Φ_k is an appropriate family of functions (e.g., a neural network family) and Q_n, P_n are the empirical measures constructed from n independent samples from Q and P, respectively. Note that there are two levels of approximation here: we approximate the measures $Q \approx Q_n$, $P \approx P_n$, and we approximate the function space $\Gamma \approx \Phi_k$, with the approximations becoming arbitrarily good (in the appropriate senses) as $n, k \to \infty$. In Theorem 4.6 below we will give a consistency result for (4.1); under appropriate assumptions we will show that for all $\delta > 0$ there exists $K \in \mathbb{Z}^+$ such that for all $k \geq K$ we have

$$\lim_{n\to\infty}\mathbb{P}\left(\left|R_\alpha(Q\|P)-\widehat{R}_\alpha^{n,k}(Q\|P)\right|\geq\delta\right)=0. \tag{4.2}$$

Theorem 3.1 implies that the Φ_k are allowed to contain unbounded functions, an important point for practical computations. In addition, note the objective functional in (4.1) only involves the values of ϕ at the sample points; there is no need to estimate the likelihood ratio dQ/dP. In contrast, estimators of the form (4.1) perform well in high dimensions, as we demonstrate below in section 5.

4.1. Neural network estimators for Rényi divergences. While we will provide a general consistency theory for the estimator (4.1) in section 4.2, we are primarily interested in neural network estimators on $\Omega = \mathbb{R}^m$, i.e., where the Φ_k in (4.1) are neural network families. By a neural network family, we mean a collection of functions, $\phi : \mathbb{R}^m \to \mathbb{R}$ (here, $\mathbb{R}^m$ is called the input layer and $\mathbb{R}$ the output layer), that are constructed as follows: First compose some number, d, of hidden layers of the form $\sigma_j \circ B_{j-1}$, where $B_{j-1} : \mathbb{R}^{m_{j-1}} \to \mathbb{R}^{m_j}$ is affine ($m_0 \equiv m$) and $\sigma_j : \mathbb{R}^{m_j} \to \mathbb{R}^{m_j}$ is a (nonlinear) activation function. Then finish by composing with a final affine map $B_d : \mathbb{R}^{m_d} \to \mathbb{R}$. Often, the σ_j's are defined by applying a nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$ to each of the m_j components; in such a case, we will call σ the activation function. The parameters of the neural network consist of the (weight) matrices and shift (i.e., bias) vectors from all affine transformations used in the construction (for technical reasons, we will assume that the set of allowed weights and biases is closed). The number of hidden layers is called the depth of the network, and the dimension of each layer is called its width.

As we will see in Theorem 4.6 below, consistency of the estimator (4.1) will rely on the ability of $\Phi \equiv \cup_k \Phi_k$ to approximate $\Gamma = \mathrm{Lip}_b(\mathbb{R}^m)$ in the appropriate sense. Neural networks are well suited for this task, as they satisfy various versions of the universal approximation property. The two most common variants are as follows:

a. For all $g \in C(\mathbb{R}^m)$, all $\epsilon > 0$, and all compact $K \subset \mathbb{R}^m$ there exists $\phi \in \Phi$ such that

$$\sup_{x\in K}|g(x)-\phi(x)|<\epsilon\,. \tag{4.3}$$

b. Let $p \in [1, \infty)$. For all $g \in L^p(\mathbb{R}^m)$ and all $\epsilon > 0$ there exists $\phi \in \Phi$ such that

$$\int_{\mathbb{R}^m} |g(x) - \phi(x)|^p dx < \epsilon\,. \tag{4.4}$$

For example, under suitable assumptions the family of (shallow) arbitrary width neural networks satisfies (4.3) [13, 42]. Results for deep networks with bounded width are also known; see [27] for (4.3) and [35, 41] for (4.4). Here we will only work with neural networks consisting of continuous functions, i.e., those with continuous activation functions; this is true of most activation functions used in practice.

We will prove that consistency of a neural network estimator follows from one of several strengthened versions of the universal approximation property; we introduce these in Definitions 4.1 and 4.2 below. Before presenting these details, we first give three classes of networks to which our consistency result (Theorem 4.6) will apply; proofs that all required assumptions are satisfied can be found in the supplementary materials file Supplement.pdf [local/web 296KB].

1. Measures with compact support: Let $\Omega \subset \mathbb{R}^m$ be compact, let Φ be a family of neural networks that satisfy the universal approximation property (4.3), and let $\Phi_k \subset \Phi$ be the set of networks with depth and width bounded by k and with parameter values restricted to $[-a_k, a_k]$, where $a_k \nearrow \infty$. Then the estimator (4.1) is consistent.
2. Noncompact support, bounded Lipschitz activation functions: Let $\Omega = \mathbb{R}^m$ and Φ be the family of neural networks with 2 hidden layers, arbitrary width, and activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Let $\Phi_k \subset \Phi$ be the set of width-k networks with parameter values restricted to $[-a_k, a_k]$, where $a_k \nearrow \infty$ (this family of networks satisfies (4.3)). If the activation function, σ, is bounded and there exists $(c, d) \subset \mathbb{R}$ on which σ is one-to-one and Lipschitz, then the estimator (4.1) is consistent.
3. Noncompact support, unbounded Lipschitz activation functions: Let $p \in (1, \infty)$ and $\Omega = \mathbb{R}^m$. Let Q and P be probability measures on Ω with finite moment generating functions everywhere and with densities dQ/dx and dP/dx that are bounded on compact sets. Let Φ be the family of neural networks obtained by using either the ReLU activation function or the GroupSort activation with group size 2 (these satisfy variants of (4.3) and (4.4); see Theorem 1 in [41] and Theorem 3 in [2], respectively); note that these activations are unbounded; hence in this case it is critical that Theorem 3.1 applies to spaces of unbounded functions. Finally, let $\Phi_k \subset \Phi$ be the set of networks with depth and width bounded by k and with parameter values restricted to $[-a_k, a_k]$, where $a_k \nearrow \infty$. Then the estimator (4.1) is consistent. For ReLU activations our proof shows that 3 hidden layers are sufficient.

Note that in all cases, the Φ_k's are an increasing family of neural networks with parameter values restricted to an increasing family of compact sets. Similar boundedness assumptions on the network parameters were required in [7], which studied neural network estimators for the KL-divergence. Apart from generalizing to Rényi divergences, the primary contributions of the current work are several new approximation results which enable us to consider Q and P with noncompact support as well as unbounded activation functions. In contrast, the consistency result for KL-divergence in [7] only applies to compactly supported measures (in which case boundedness of the activation is irrelevant).

4.2. Consistency of the Rényi divergence estimators. Though we are primarily interested in neural network estimators, we will present our consistency result in terms of abstract requirements on the approximation spaces Φ_k. Intuitively, the basic requirement is that $\Phi \equiv \cup_k \Phi_k$ is "dense" in $\mathrm{Lip}_b(\Omega)$ in the appropriate sense. More precisely, we will need a space of functions, Φ, that satisfies one of the following strengthened/modified versions of the universal approximation properties from (4.3) and (4.4).

Definition 4.1. *Let Ω be a metric space and $\Phi, \Psi \subset \mathcal{M}(\Omega)$. We say that Φ has the* Ψ-bounded L^∞ approximation property *if the following two properties hold:*

1. *For all $\phi \in \Phi$ there exists $\psi \in \Psi$ with $|\phi| \leq \psi$.*
2. *For all $g \in \mathrm{Lip}_b(\Omega)$ there exists $\psi \in \Psi$ such that*
   (a) *$|g| \leq \psi$;*
   (b) *for all compact $K \subset \Omega$ and all $\epsilon > 0$ there exists $\phi \in \Phi$ with $|\phi| \leq \psi$ and $\sup_{x \in K} |g(x) - \phi(x)| < \epsilon$.*

Definition 4.2. *Let Ω be a metric space, $\mathcal{Q}$ be a collection of Borel probability measures on Ω, and $\Phi, \Psi \subset \mathcal{M}(\Omega)$. Let $p \in [1, \infty)$. We say that Φ has the* Ψ-bounded $L^p(\mathcal{Q})$ approximation property *if the following two properties hold:*

1. *For all $\phi \in \Phi$ there exists $\psi \in \Psi$ with $|\phi| \leq \psi$.*
2. *For all $g \in \mathrm{Lip}_b(\Omega)$ there exists $\psi \in \Psi$ such that*
   (a) *$|g| \leq \psi$;*
   (b) *for all compact $K \subset \Omega$ and all $\epsilon > 0$ there exists $\phi \in \Phi$ with $|\phi| \leq \psi$ and $\sup_{\mu \in \mathcal{Q}} \left(\int_K |g - \phi|^p d\mu \right)^{1/p} < \epsilon$.*

Intuitively, these definitions state that functions in Φ are able to approximate bounded Lipschitz functions on compact sets (in some norm), and with the approximating functions being uniformly bounded on the whole space by some fixed function in Ψ. For the neural network families 1 and 2 of section 4.1 we will let Ψ be the set of positive constant functions, and in case 3 we will let $\Psi = \{x \mapsto a\|x\| + b : a, b \geq 0\}$; see the supplementary materials file Supplement.pdf [local/web 296KB] for details.

Under appropriate integrability assumptions on Ψ, the ability to approximate in either of the above manners allows one to restrict the optimization in (3.1) to Φ, leading to the following result (the proof can be found in section 6.2).

Lemma 4.3. *Let Ω be a complete separable metric space, Q, P be Borel probability measures on Ω, $\alpha \in \mathbb{R} \setminus \{0, 1\}$, and $\Phi, \Psi \subset \mathcal{M}(\Omega)$. Suppose one of the following two collections of properties holds:*

1. (a) *Φ has the Ψ-bounded L^∞ approximation property.*
   (b) *$e^{\pm(\alpha-1)\psi} \in L^1(Q)$ for all $\psi \in \Psi$.*
   (c) *$e^{\pm\alpha\psi} \in L^1(P)$ for all $\psi \in \Psi$.*
2. *There exist conjugate exponents $p, q \in (1, \infty)$ such that*
   (a) *Φ has the Ψ-bounded $L^p(\mathcal{Q})$ approximation property, where $\mathcal{Q} \equiv \{Q, P\}$;*
   (b) *$e^{\pm q(\alpha-1)\psi} \in L^1(Q)$ for all $\psi \in \Psi$;*
   (c) *$e^{\pm q\alpha\psi} \in L^1(P)$ for all $\psi \in \Psi$.*

Then

$$R_\alpha(Q\|P) = \sup_{\phi\in\Phi}\left\{\frac{1}{\alpha-1}\log\int e^{(\alpha-1)\phi}dQ - \frac{1}{\alpha}\log\int e^{\alpha\phi}dP\right\}. \tag{4.5}$$

We will be able to prove consistency of the estimator (4.1) when the approximation spaces, Φ_k, increase to a function space, Φ, that satisfies the assumptions of Lemma 4.3. More specifically (and slightly more generally), we will work under the following set of assumptions.

Assumption 4.4. Suppose we have $\Phi_k, \Psi \subset \mathcal{M}(\Omega)$ that satisfy the following:

1. $$R_\alpha(Q\|P) = \lim_{k\to\infty}\sup_{\phi\in\Phi_k}\left\{\frac{1}{\alpha-1}\log\int e^{(\alpha-1)\phi}dQ - \frac{1}{\alpha}\log\int e^{\alpha\phi}dP\right\}. \tag{4.6}$$
2. Each Φ_k has the form
$$\Phi_k = \{\phi_k(\cdot,\theta) : \theta\in\Theta_k\}, \tag{4.7}$$
where $\phi_k : \Omega\times\Theta_k \to \mathbb{R}$ is continuous and Θ_k is a compact metric space.
3. For each k there exists $\psi_k \in \Psi$ with $\sup_{\theta\in\Theta_k}|\phi_k(\cdot,\theta)| \le \psi_k$.
4. $e^{\pm(\alpha-1)\psi} \in L^1(Q)$ for all $\psi\in\Psi$.
5. $e^{\pm\alpha\psi} \in L^1(P)$ for all $\psi\in\Psi$.

Our primary means of satisfying the condition (4.6) is described in the following lemma.

Lemma 4.5. *Suppose Φ satisfies the assumptions of Lemma 4.3. Take subsets $\Phi_k \subset \Phi_{k+1} \subset \Phi$, $k\in\mathbb{Z}^+$, with $\cup_k\Phi_k = \Phi$. Then the equality (4.6) holds.*

We use this lemma in the concrete examples in section 4.1 and in the proofs in the supplementary materials file Supplement.pdf [local/web 296KB]. However, we will not directly use Lemma 4.5 in the proof of the consistency result, Theorem 4.6; there we will work under the more general Assumption 4.4. We now state our consistency result.

Theorem 4.6. *Let $\alpha\in\mathbb{R}\backslash\{0,1\}$, Ω be a complete separable metric space, P,Q be Borel probability measures on Ω, and X_i, Y_i, $i\in\mathbb{Z}_+$ be Ω-valued random variables on a probability space $(N,\mathcal{N},\mathbb{P})$. Suppose X_i are i.i.d. and Q-distributed, suppose Y_i are i.i.d. and P-distributed, and let Q_n, P_n denote the corresponding n-sample empirical measures. Suppose Assumption 4.4 holds for the spaces $\Phi_k, \Psi \subset \mathcal{M}(\Omega)$, $k\in\mathbb{Z}_+$; in particular, the Φ_k's have the form*

$$\Phi_k = \{\phi_k(\cdot,\theta) : \theta\in\Theta_k\}. \tag{4.8}$$

Define the corresponding estimator

$$\widehat{R}_\alpha^{n,k}(Q\|P) = \sup_{\theta\in\Theta_k}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ_n\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta}}dP_n\right]\right\}. \tag{4.9}$$

1. *If $R_\alpha(Q\|P) < \infty$, then for all $\delta > 0$ there exists $K\in\mathbb{Z}^+$ such that for all $k \ge K$ we have*
$$\lim_{n\to\infty}\mathbb{P}\left(\left|R_\alpha(Q\|P) - \widehat{R}_\alpha^{n,k}(Q\|P)\right| \ge \delta\right) = 0. \tag{4.10}$$

2. *If* $R_\alpha(Q\|P) = \infty$, *then for all* $M > 0$ *there exists* $K \in \mathbb{Z}^+$ *such that for all* $k \geq K$ *we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{R}_\alpha^{n,k}(Q\|P) \leq M\right) = 0. \tag{4.11}$$

The proof of Theorem 4.6, which can be found in section 6.2, is inspired by the work in [7], which used the Donsker–Varadhan variational formula (1.2) to estimate the KL-divergence. However, as mentioned above, we have developed new techniques that allow us to prove consistency when Q and P have noncompact support. This is accomplished by introducing the space Ψ in both Lemma 4.3 and Theorem 4.6, which allows the use of ϕ's that are Ψ-bounded, as opposed to simply being bounded.

If $\Theta_k \subset \mathbb{R}^{d_k} \cap \{\theta : \|\theta\| \leq K_k\}$ and ϕ_k is bounded by M_k and is L_k-Lipschitz (i.e., Lipschitz continuous with constant L_k) in $\theta \in \Theta_k$, then one can derive sample complexity bounds for the estimator (4.1) by using the same technique that was used in [7] to study KL-divergence estimators. To obtain an α-divergence estimator error less than ϵ with probability at least $1 - \delta$, it is sufficient to have the number of samples, n, satisfy

$$n \geq \frac{32 D_{\alpha,k}^2}{\epsilon^2}\left(d_k \log(16 L_k K_k \sqrt{d_k}/\epsilon) + 2 d_k M_k \max\{|\alpha|, |\alpha - 1|\} + \log(4/\delta)\right), \tag{4.12}$$

where $D_{\alpha,k} \equiv \max\{e^{2|\alpha| M_k}/|\alpha|, e^{2|\alpha-1| M_k}/|\alpha - 1|\}$. The qualitative behavior of (4.12) in ϵ, δ, and d_k is the same as the KL result from [7], though some modifications to the proof are necessary. The derivation uses the same techniques as the proof of Theorem 3 in [7]. In particular, it relies on a combination of concentration inequalities and covering theorems to obtain a nonasymptotic uniform law of large numbers–type result; see [54] for details on these tools. We include a proof of (4.12) in the supplementary materials file Supplement.pdf [local/web 296KB].

5. Numerical examples.

In this section we present several numerical examples of using the estimator (4.9); in practice, we search for the optimum in (4.9) via stochastic gradient descent (SGD) [20, 21, 55]. We take the function space, Φ, to be a neural network family ϕ_θ, $\theta \in \Theta$, with ReLU activation function, $\sigma(x) = \mathrm{ReLU}(x) \equiv \max\{x, 0\}$. We used the AdamOptimizer method [28, 45], an adaptive learning-rate SGD algorithm, to search for the optimum. All computations were performed in TensorFlow.

5.1. Example: Estimating Rényi divergences in high dimensions.

Estimators of divergences based on variational formulas are especially powerful in high-dimensional systems with hidden low-dimensional (nonlinear) structure, a setting that, again, is challenging for likelihood-ratio based methods. We illustrate the effectiveness of the estimator (4.9) in such a setting by estimating the Rényi divergence between the distributions of $h(X)$ and $h(Y)$, where X and Y are both 4-dimensional Gaussians and $h : \mathbb{R}^4 \to \mathbb{R}^{5000}$ is a nonlinear map. If h is an embedding (in particular, it must be one-to-one), then the data processing inequality (see Theorem 14 in [34]) implies $R_\alpha(P_{h(X)}\|P_{h(Y)}) = R_\alpha(P_X\|P_Y)$, with the latter being easily computable (we use P_Z to denote the distribution of a random variable Z). Hence we have an exact value with which we can compare our numerical estimate of $R_\alpha(P_{h(X)}\|P_{h(Y)})$. In Figure 1 we show the relative error, comparing the results of our method to the exact values

of the Rényi divergences. The left panel shows the error as a function of the number of SGD iterations, and the right panel shows the error as a function of the size of the data set.

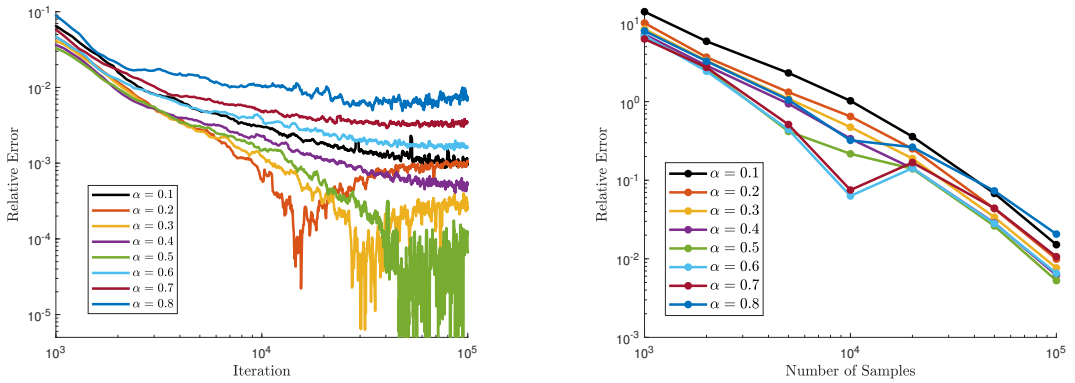

Figure 1. *Left: Relative error of Rényi divergence estimators (4.9) between the distributions of $h(X)$ and $h(Y)$, where X and Y are 4-dimensional Gaussians (with means $\mu_p = 0$, $\mu_q = (2,0,0,0)$ and covariance matrices $\Sigma_p = I$, $\Sigma_q = diag(1.5, 0.7, 2, 1)$) and $h : \mathbb{R}^4 \to \mathbb{R}^{5000}$ is a nonlinear map. Specifically, we let $h_i(x) = x_i$ for $i = 1, \dots, 4$ (to ensure it is an embedding), and then for $i > 4$ we define $h_i(x) = A_i(x) + c_{1,i} \cos(c_{2,i} x_{j_{1,i}}) \sin(c_{3,i} x_{j_{2,i}}) + c_{4,i} x_{j_{3,i}} x_{j_{4,i}}$, where A is an affine function and $j_{k,i} \in \{1, \dots, 4\}$; the parameters of A and the $c_{k,i}$'s were randomly selected at the start of each run (all components are i.i.d. $N(0,1)$). The indices $j_{k,i}$ were also randomly selected at the start of each run (i.i.d. $Unif(\{1, \dots, 4\})$). Computations were done using a neural network with 1 hidden layer of 128 nodes. On the left we show the relative error as a function of the number of SGD iterations; SGD was performed using a minibatch size of 1,000 and an initial learning rate of 2×10^{-4}. We show the moving average over the last 10 data points, with results averaged over 20 runs. The behavior of the $\alpha = 0.2, 0.3$ curves is due to the estimates crossing above and converging to a result slightly above the true values. On this problem the method failed to converge when $\alpha = 0.9$ and when using the KL-divergence. Right: The relative error as a function of the number of samples, N. We used a fixed number of 10,000 SGD iterations, with the other parameters being as in the left panel. Results were averaged over 100 runs. The error is well approximated by a power-law decay of $N^{-1.4}$, and this behavior appears insensitive to the value of α.*

Our choice of nonlinear map h is detailed in the caption. We emphasize that the estimator (4.1) is effective in high dimensions, with no preprocessing (i.e., dimensional reduction) of the data required; the results shown in Figure 1 were obtained by applying the algorithm directly to the 5000-dimensional data. Note that here, and as a general rule, the estimation becomes more difficult as $\alpha \to 0, 1$ (i.e., the KL limits), regimes where the importance of rare events increases. The method failed to converge when $\alpha = 1$ (i.e., when using the KL objective functional), and numerical estimation is even more challenging when $\alpha > 1$.

5.2. Example: Estimating Rényi-based mutual information. Next we demonstrate the use of (4.9) in the estimation of Rényi mutual information,

(5.1) (Rényi-MI) $$R_\alpha(P_{(X,Y)} \| P_X \times P_Y),$$

between random variables X and Y; this should be compared with [7], which used the Donsker–Varadhan variational formula to estimate KL mutual information, and [9], which considered f-divergences. (Mutual information is typically defined in terms of the KL-divergence, but

one can consider many alternative divergences; see, e.g., [44].) In the left panel of Figure 2 we show the results of estimating the Rényi-MI where $\alpha = 1/2$ and X and Y are correlated 20-dimensional Gaussians with componentwise correlation ρ (the same case that was considered in [7, 9]). This is a moderate dimensional problem (specifically, 40-dimensional) with no low-dimensional structure. Our method is capable of accurately estimating the Rényi-MI over a wide range of correlations, something not achievable with likelihood-ratio based nonparametric methods (again, see [26, 7]).

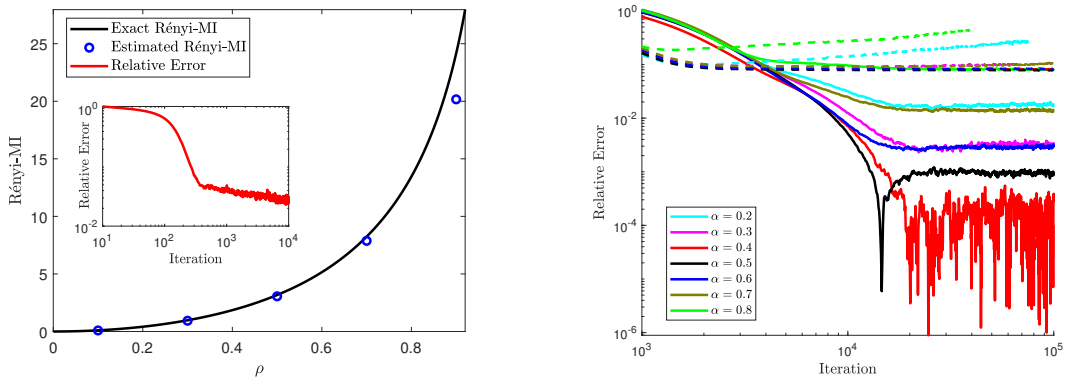

Figure 2. *Left: Estimation of Rényi-based mutual information (5.1) with $\alpha = 1/2$ between 20-dimensional correlated Gaussians with componentwise correlation ρ. We used a neural network with one hidden layer of 256 nodes, and training was performed with a minibatch size of 1,000. We show the Rényi-MI as a function of ρ after 10,000 steps of SGD and averaged over 20 runs. The inset shows the relative error for a single run with $\rho = 0.5$, as a function of the number of SGD iterations. Right: Estimation of the Rényi divergence between two 25-dimensional distributions of the form $\prod_{i=1}^{25} Beta(a_i, b_i)$. The exponential family estimator (5.2) (solid curves) outperformed the neural-network estimator (4.9) (dashed curves) with a comparable number of parameters (one hidden layer with 4 nodes). Training was performed with a minibatch size of 1,000 and an initial learning rate of 0.001. Results were averaged over 20 runs, and the values of the a and b parameters for each distribution were randomly selected at the start of each run. Again, the estimation becomes more difficult as $\alpha \to 0, 1$.*

5.3. Example: Estimating Rényi divergence for exponential families. As discussed in section 3.1, when working with an exponential family, the formula for the optimizer (see Corollary 3.2) reduces the Rényi variational formula to a finite dimensional optimization problem (see (3.4)). Using the corresponding estimator,

$$\widehat{R}^n_\alpha(Q\|P) = \sup_{\Delta\kappa\in\mathbb{R}^k} \left\{ \frac{1}{\alpha-1} \log \int e^{(\alpha-1)\Delta\kappa\cdot T(x)} dQ_n - \frac{1}{\alpha} \log \int e^{\alpha\Delta\kappa\cdot T(x)} dP_n \right\}, \tag{5.2}$$

can yield a substantial computational benefit over a general-purpose neural network estimator (4.9), as we now demonstrate. Here we estimate the divergence between products of Beta distributions; this is another moderate dimensional problem (specifically, 25-dimensional) with no low-dimensional structure. The results are shown in the right panel of Figure 2. The solid curves show the relative error that resulted from using (5.2), while the dashed curves show the result of using a neural network estimator (4.9) with a comparable number of parameters (specifically, one hidden layer with 4 nodes, and hence on the order of 100 parameters). The

former achieves high accuracy over a range of α's, while the latter performs poorly and fails to converge in several cases. To achieve comparable accuracy with a neural network estimator would require a much larger network, leading to a much greater computational cost.

6. Proofs.

6.1. Proof of the Rényi–Donsker–Varadhan variational formula.

The starting point for the proof of Theorem 3.1 is the following variational formula, proven in [5]: Let P be a probability measure on $(\Omega, \mathcal{M})$, $g \in \mathcal{M}_b(\Omega)$, and $\alpha > 0$, $\alpha \neq 1$. Then

$$\frac{1}{\alpha} \log \left[\int e^{\alpha g} dP \right] = \sup_Q \left\{ \frac{1}{\alpha - 1} \log \left[\int e^{(\alpha-1)g} dQ \right] - R_\alpha(Q\|P) \right\}, \tag{6.1}$$

where the optimization is over all probability measures, Q, on $(\Omega, \mathcal{M})$. (Let $\gamma = \alpha$, $\beta = \alpha - 1$ in equation (1.3) of [5]). Though the right-hand side of (6.1) is not a Legendre transform, (6.1) is still in some sense a "dual" version of (3.1); this is reminiscent of the duality between the Donsker–Varadhan variational formula (1.2) and the Gibbs variational principle (see Proposition 1.4.2 in [16]). Equation (6.1) was previously used in [5, 17, 4] to derive uncertainty quantification bounds on risk-sensitive quantities (e.g., rare events or large deviations estimates) and in [6] to derive PAC-Bayesian bounds.

In fact, we will not require the full strength of (6.1). We will only need the following bound for $g \in \mathcal{M}_b(\Omega)$, $\alpha > 0$, $\alpha \neq 1$:

$$\frac{1}{\alpha - 1} \log \left[\int e^{(\alpha-1)g} dQ \right] \leq \frac{1}{\alpha} \log \left[\int e^{\alpha g} dP \right] + R_\alpha(Q\|P)\,. \tag{6.2}$$

To keep our argument self-contained, we include a proof of (6.2) below. Our proof is adapted from the proof of (6.1) found in section 4 of [5]. We note that an alternative proof of (6.2) can be given by using a different variational formula for the Rényi divergences, which can be found in Theorem 30 of [53] and also in Theorem 1 of [1].

Proof of (6.2). We separate the proof into two cases.

(1) $\alpha > 1$: If $Q \not\ll P$, the result is trivial (see (2.1)), so assume $Q \ll P$. For $g \in \mathcal{M}_b(\Omega)$ we can use Hölder's inequality with conjugate exponents $\alpha/(\alpha - 1)$ and α to obtain

$$\begin{aligned} \frac{1}{\alpha - 1} \log \int e^{(\alpha-1)g} dQ \leq & \frac{1}{\alpha - 1} \log \left[\left(\int (e^{(\alpha-1)g})^{\frac{\alpha}{\alpha-1}} dP \right)^{\frac{\alpha-1}{\alpha}} \left(\int \left(\frac{dQ}{dP} \right)^{\alpha} dP \right)^{\frac{1}{\alpha}} \right] \\ = & \frac{1}{\alpha} \log \int e^{\alpha g} dP + \frac{1}{\alpha(\alpha - 1)} \log \int (dQ/dP)^{\alpha} dP. \end{aligned} \tag{6.3}$$

In this case the definition (2.1) implies $R_\alpha(Q\|P) = \frac{1}{\alpha(\alpha-1)} \log \int (dQ/dP)^\alpha dP$, and so we have proven the claimed bound (6.2).

(2) $\alpha \in (0, 1)$: Let $dP = p d\nu$, $dQ = q d\nu$ as in definition (2.1), and define $h = e^{-g} q$. Then

$$R_\alpha(Q\|P) = \frac{1}{\alpha(\alpha - 1)} \log \int q^\alpha p^{1-\alpha} d\nu = \frac{1}{\alpha(\alpha - 1)} \log \int_{p,q>0} (h/p)^{\alpha-1} e^{(\alpha-1)g} dQ. \tag{6.4}$$

Using Hölder's inequality for the measure $e^{(\alpha-1)g}dQ$, the conjugate exponents $1/\alpha$ and $1/(1-\alpha)$, and the functions 1 and $1_{q,p>0}(h/p)^{\alpha-1}$, we find

$$\begin{aligned}\int_{q,p>0}(h/p)^{\alpha-1}e^{(\alpha-1)g}dQ &\le \left(\int e^{(\alpha-1)g}dQ\right)^{\alpha}\left(\int_{q,p>0}(h/p)^{-1}e^{(\alpha-1)g}dQ\right)^{1-\alpha}\\ &= \left(\int e^{(\alpha-1)g}dQ\right)^{\alpha}\left(\int_{q,p>0}e^{\alpha g}dP\right)^{1-\alpha}\\ &\le \left(\int e^{(\alpha-1)g}dQ\right)^{\alpha}\left(\int e^{\alpha g}dP\right)^{1-\alpha}.\end{aligned} \tag{6.5}$$

Taking the logarithm of both sides, dividing by $\alpha(\alpha-1)$ (which is negative), and using (6.4), we arrive at

$$R_\alpha(Q\|P) \ge \frac{1}{\alpha-1}\log\int e^{(\alpha-1)g}dQ - \frac{1}{\alpha}\log\int e^{\alpha g}dP. \tag{6.6}$$

This implies the claimed bound (6.2) and completes the proof. ∎

We now use (6.2) to derive the variational formula (3.1). The argument is inspired by the proof of the Donsker–Varadhan variational formula from Appendix C.2 in [16].

Proof of Theorem 3.1. First let $\Gamma = \mathcal{M}_b(\Omega)$. If one can show (3.1) for all $\alpha > 1$ and all P, Q, then, using (2.2) and reindexing $g \to -g$ in the supremum, one finds that (3.1) also holds for all $\alpha < 0$. So we only need to consider the cases $\alpha \in (0,1)$ and $\alpha > 1$.

Inequality (6.2) immediately implies

$$\begin{aligned}R_\alpha(Q\|P) \ge& \sup_{g\in\mathcal{M}_b(\Omega)}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g}dP\right]\right\}\\ \equiv& \widetilde{R}_\alpha(Q\|P).\end{aligned} \tag{6.7}$$

If $Q \ll P$ and $g^* \equiv \log(dQ/dP) \in \mathcal{M}_b(\Omega)$, then the reverse inequality easily follows from an explicit calculation. However, $g^* \in \mathcal{M}_b(\Omega)$ is a very strong assumption which we do not make here. Our general proof will therefore require several limiting arguments but will still be based on this intuition.

We separate the proof of the reverse inequality into three cases.

(1) $\alpha > 1$ and $Q \not\ll P$: We will show $\widetilde{R}_\alpha(Q\|P) = \infty$, which will prove the desired inequality. To do this, take a measurable set A with $P(A) = 0$ but $Q(A) \ne 0$, and define $g_n = n1_A$. The definition (6.7) implies

$$\begin{aligned}\widetilde{R}_\alpha(Q\|P) \ge& \frac{1}{\alpha-1}\log\int e^{(\alpha-1)g_n}dQ - \frac{1}{\alpha}\log\int e^{\alpha g_n}dP\\ =& \frac{1}{\alpha-1}\log\left[e^{(\alpha-1)n}Q(A) + Q(A^c)\right] - \frac{1}{\alpha}\log P(A^c).\end{aligned} \tag{6.8}$$

The lower bound goes to $+\infty$ as $n \to \infty$ (here it is key that $\alpha > 1$) and therefore we have the claimed result.

(2) $\alpha > 1$ and $Q \ll P$: In this case we can take $\nu = P$ in (2.1) and write

$$R_\alpha(Q\|P) = \frac{1}{\alpha(\alpha-1)} \log\left[\int (dQ/dP)^\alpha dP\right]. \tag{6.9}$$

Define

$$f_{n,m}(x) = x1_{1/m<x<n} + n1_{x\geq n} + 1/m1_{x\leq 1/m} \tag{6.10}$$

and $g_{n,m} = \log(f_{n,m}(dQ/dP))$. These are bounded, and so (6.7) implies

$$\begin{aligned}\widetilde{R}_\alpha(Q\|P) \geq& \frac{1}{\alpha-1}\log\int e^{(\alpha-1)g_{n,m}}dQ - \frac{1}{\alpha}\log\int e^{\alpha g_{n,m}}dP \\ =& \frac{1}{\alpha-1}\log\int f_{n,m}(dQ/dP)^{(\alpha-1)}\frac{dQ}{dP}dP - \frac{1}{\alpha}\log\int f_{n,m}(dQ/dP)^\alpha dP.\end{aligned} \tag{6.11}$$

Define $f_{n,\infty}(x) = x1_{x<n} + n1_{x\geq n}$. Using the dominated convergence theorem to take $m \to \infty$ in (6.11), we find

$$\begin{aligned}\widetilde{R}_\alpha(Q\|P) \geq& \frac{1}{\alpha-1}\log\int f_{n,\infty}(dQ/dP)^{(\alpha-1)}\frac{dQ}{dP}dP - \frac{1}{\alpha}\log\int f_{n,\infty}(dQ/dP)^\alpha dP \\ \geq& \frac{1}{\alpha(\alpha-1)}\log\int f_{n,\infty}(dQ/dP)^\alpha dP.\end{aligned} \tag{6.12}$$

To obtain the last line we used $xf_{n,\infty}(x)^{\alpha-1} \geq f_{n,\infty}(x)^\alpha$. Next, we have $0 \leq f_{n,\infty}(dQ/dP) \nearrow dQ/dP$ as $n \to \infty$, and so the monotone convergence theorem implies

$$\widetilde{R}_\alpha(Q\|P) \geq \frac{1}{\alpha(\alpha-1)}\log\int (dQ/dP)^\alpha dP = R_\alpha(Q\|P). \tag{6.13}$$

This proves the claimed result for case (2).

(3) $\alpha \in (0,1)$: In this case definition (2.1) becomes

$$R_\alpha(Q\|P) = \frac{1}{\alpha(\alpha-1)}\log\left[\int_{p>0} q^\alpha p^{1-\alpha} d\nu\right], \tag{6.14}$$

where ν is any sigma-finite positive measure for which $dQ = qd\nu$ and $dP = pd\nu$. Define $f_{n,m}(x)$ via (6.10), and let $g_{n,m} = \log(f_{n,m}(q/p))$, where q/p is defined to be 0 if $q = 0$ and $+\infty$ if $p = 0$ and $q \neq 0$. The functions $g_{n,m}$ are bounded; hence (6.7) implies

$$\begin{aligned}\widetilde{R}_\alpha(Q\|P) \geq& -\frac{1}{1-\alpha}\log\int e^{(\alpha-1)g_{n,m}}dQ - \frac{1}{\alpha}\log\int e^{\alpha g_{n,m}}dP \\ =& -\frac{1}{1-\alpha}\log\int f_{n,m}(q/p)^{\alpha-1}qd\nu - \frac{1}{\alpha}\log\int f_{n,m}(q/p)^\alpha pd\nu.\end{aligned} \tag{6.15}$$

Define $f_{\infty,m}(x) = x1_{x>1/m} + 1/m1_{x\leq 1/m}$. We have the bound $f_{n,m}(q/p)^{\alpha-1} \leq (1/m)^{\alpha-1}$ (here it is critical that $\alpha \in (0,1)$), and so the dominated convergence theorem can be used to

compute the $n \to \infty$ limit of the first term on the right-hand side of (6.15), while the second term can be bounded using $f_{n,m}(q/p)^\alpha \le f_{\infty,m}(q/p)^\alpha$. We thereby obtain

$$\begin{aligned}\widetilde{R}_\alpha(Q\|P) &\ge -\frac{1}{1-\alpha}\log\int f_{\infty,m}(q/p)^{\alpha-1}q d\nu - \frac{1}{\alpha}\log\int f_{\infty,m}(q/p)^\alpha p d\nu \\ &\ge -\frac{1}{1-\alpha}\log\int_{q>0,p>0} q^\alpha p^{1-\alpha} d\nu - \frac{1}{\alpha}\log\int_{p>0} f_{\infty,m}(q/p)^\alpha p d\nu,\end{aligned} \tag{6.16}$$

where we used $f_{\infty,m}(x) \ge x$ to obtain the second line. Using the dominated convergence theorem on the second term (which is always finite), we find

$$\begin{aligned}\widetilde{R}_\alpha(Q\|P) &\ge -\frac{1}{1-\alpha}\log\int_{p>0} q^\alpha p^{1-\alpha} d\nu - \frac{1}{\alpha}\log\int_{p>0} q^\alpha p^{1-\alpha} d\nu \\ &= \frac{1}{\alpha(\alpha-1)}\log\int_{p>0} q^\alpha p^{1-\alpha} d\nu = R_\alpha(Q\|P).\end{aligned}$$

Therefore, the claim is proven in case (3), and the proof of (3.1) is complete.

In addition, now suppose that $(\Omega, \mathcal{M})$ is a metric space with the Borel σ-algebra. We will next show that (3.1) holds with $\Gamma = C_b(\Omega)$, the space of bounded continuous functions on Ω. Define the probability measure $\mu = (P+Q)/2$, and let $g \in \mathcal{M}_b(\Omega)$. Lusin's theorem (see, e.g., Appendix D in [15]) implies that for all $n \in \mathbb{Z}^+$ there exists a closed set $F_n \subset \Omega$ such that $\mu(F_n^c) < 1/n$ and $g|_{F_n}$ is continuous. By the Tietze extension theorem (see, e.g., Theorem 4.16 in [18]) there exists $g_n \in C_b(\Omega)$ with $\|g_n\|_\infty \le \|g\|_\infty$ and $g_n = g$ on F_n. Therefore,

$$\begin{aligned}\left|\int e^{(\alpha-1)g_n} dQ - \int e^{(\alpha-1)g} dQ\right| &\le (\|e^{(\alpha-1)g_n}\|_\infty + \|e^{(\alpha-1)g}\|_\infty) Q(F_n^c) \\ &\le 4e^{|\alpha-1|\|g\|_\infty}/n \to 0\end{aligned} \tag{6.17}$$

as $n \to \infty$. Similarly, we have $\lim_{n\to\infty}\int e^{\alpha g_n} dP = \int e^{\alpha g} dP$. Hence

$$\begin{aligned}&\sup_{g\in C_b(\Omega)}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g} dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g} dP\right]\right\} \\ &\ge \lim_{n\to\infty}\left(\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g_n} dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g_n} dP\right]\right) \\ &= \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g} dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g} dP\right].\end{aligned} \tag{6.18}$$

$g \in \mathcal{M}_b(\Omega)$ was arbitrary, and so we have proven

$$\begin{aligned}&\sup_{g\in C_b(\Omega)}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g} dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g} dP\right]\right\} \\ &\ge \sup_{g\in \mathcal{M}_b(\Omega)}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g} dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g} dP\right]\right\}.\end{aligned} \tag{6.19}$$

The reverse inequality is trivial. Therefore, we have shown that (3.1) holds with $\Gamma = C_b(\Omega)$. To see that (3.1) holds when $\Gamma = \mathrm{Lip}_b(\Omega)$, use the fact that every $g \in C_b(\Omega)$ is the pointwise

limit of Lipschitz functions, g_n, with $\|g_n\|_\infty \le \|g\|_\infty$ (see Box 1.5 on page 6 of [49]). The result then follows from a computation similar to the above, this time using the dominated convergence theorem.

Finally, we prove (3.1) with $\Gamma = \mathcal{M}(\Omega)$. To do this we need to show

$$R_\alpha(Q\|P) \ge \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g}dP\right] \tag{6.20}$$

for all $g \in \mathcal{M}(\Omega)$. The equality (3.1) then follows by combining (6.20) with Theorem 3.1. To prove the bound (6.20) we start by fixing $g \in \mathcal{M}(\Omega)$ and defining the truncated functions $g_{n,m} = -n1_{g<-n} + g1_{-n\le g\le m} + m1_{g>m}$. These are bounded, and so Theorem 3.1 implies

$$R_\alpha(Q\|P) \ge \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g_{n,m}}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g_{n,m}}dP\right]. \tag{6.21}$$

We now consider three cases, based on the value of α.

(1) $\alpha > 1$: If $\int e^{\alpha g}dP = \infty$, then (6.20) is trivial (due to our convention that $\infty-\infty = -\infty$, this is true even if $\int e^{(\alpha-1)g}dQ = \infty$), so suppose $\int e^{\alpha g}dP < \infty$. When $\alpha > 1$, (6.21) involves integrals of the form $\int e^{cg_{n,m}}d\mu$, where $c > 0$ and μ is a probability measure. We have $\lim_{n\to\infty} e^{cg_{n,m}} = e^{cg_m}$, where $g_m \equiv g1_{g\le m} + m1_{g>m}$ and $e^{cg_{n,m}} \le e^{cm}$ for all n. Therefore, the dominated convergence theorem implies

$$\lim_{n\to\infty}\int e^{cg_{n,m}}d\mu = \int e^{cg_m}d\mu. \tag{6.22}$$

We have $0 \le e^{cg_m} \nearrow e^{cg}$ as $m \to \infty$, and hence the monotone convergence theorem yields

$$\lim_{m\to\infty}\lim_{n\to\infty}\int e^{cg_{n,m}}d\mu = \lim_{m\to\infty}\int e^{cg_m}d\mu = \int e^{cg}d\mu. \tag{6.23}$$

Therefore, we can take the iterated limit of (6.21) to obtain

$$R_\alpha(Q\|P) \ge \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g}dP\right] \tag{6.24}$$

(note that we are in the subcase where the second term is finite, and so this is true even if $\int e^{(\alpha-1)g}dQ = \infty$). This proves the claim in case (1).

(2) $\alpha < 0$: Use (2.2) and apply the result of case (1) to the function $-g$ to obtain (6.20).

(3) $0 < \alpha < 1$: If either $\int e^{(\alpha-1)g}dQ = \infty$ or $\int e^{\alpha g}dP = \infty$, then the bound (6.20) is again trivial, so suppose they are both finite. For $c \in \mathbb{R}$ we can bound $e^{cg_{n,n}} \le 1 + e^{cg}$ and $\lim_{n\to\infty} e^{cg_{n,n}} = e^{cg}$. Therefore, the dominated convergence theorem implies that

$$\begin{aligned} R_\alpha(Q\|P) &\ge \lim_{n\to\infty}\left(\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g_{n,n}}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g_{n,n}}dP\right]\right) \\ &= \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)g}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha g}dP\right]. \end{aligned} \tag{6.25}$$

This proves (6.20) in case (3) and thus completes the proof of (3.1) when $\Gamma = \mathcal{M}(\Omega)$. Equation (3.1) for the spaces between $\mathcal{M}_b(\Omega)$ (or $\mathrm{Lip}_b(\Omega)$) and $\mathcal{M}(\Omega)$ then easily follows. ∎

We end this subsection by deriving a formula for the optimizer.

Proof of Corollary 3.2. If $Q \ll P$, $dQ/dP > 0$, and $(dQ/dP)^\alpha \in L^1(P)$, then we also have $P \ll Q$. By taking $\nu = P$ in (2.1) (and for $\alpha < 0$, using the definition (2.2)), we find

$$R_\alpha(Q\|P) = \frac{1}{\alpha(\alpha-1)} \log \int (dQ/dP)^\alpha \, dP. \tag{6.26}$$

Letting $g^* = \log dQ/dP$, it is straightforward to show by direct calculation that

$$\frac{1}{\alpha-1} \log\left[\int e^{(\alpha-1)g^*} dQ\right] - \frac{1}{\alpha} \log\left[\int e^{\alpha g^*} dP\right] = \frac{1}{\alpha(\alpha-1)} \log \int (dQ/dP)^\alpha \, dP. \tag{6.27}$$

This, together with Theorem 3.1, implies that (3.1) holds for any Γ with $g^* \in \Gamma \subset \mathcal{M}(\Omega)$ and g^* is an optimizer. This completes the proof. ∎

6.2. Consistency proof. In this subsection we prove consistency of the Rényi divergence estimator (4.9).

Proof of Lemma 4.3. Both assumptions 1(a) and 2(a) imply that for $\phi \in \Phi$ there exists $\psi \in \Psi$ with $|\phi| \le \psi$. Either of the integrability assumptions 1(b)–1(c) or 2(b)–2(c) then imply that all expectations on the right-hand side of (4.5) are finite. Define the probability measure $\mu = (P+Q)/2$. Ω is a complete separable metric space; hence μ is inner regular. In particular, for any $\delta > 0$ there exists a compact set K_δ such that $\mu(K_\delta) > 1 - \delta$. Fix $g \in \mathrm{Lip}_b(\Omega)$. Assumptions 1(a) and 2(a) imply that there exists $\psi_g \in \Psi$ such that $|g| \le \psi_g$, and for all $\delta, \epsilon > 0$ there exists $\phi_{\delta,\epsilon} \in \Phi$ with $|\phi_{\delta,\epsilon}| \le \psi_g$, and, in the case of 1(a),

$$\sup_{x \in K_\delta} |g(x) - \phi_{\delta,\epsilon}(x)| < \epsilon\,, \tag{6.28}$$

while in the case of 2(a) we have

$$\max\left\{\left(\int_{K_\delta} |g - \phi_{\delta,\epsilon}|^p dQ\right)^{1/p}, \left(\int_{K_\delta} |g - \phi_{\delta,\epsilon}|^p dP\right)^{1/p}\right\} < \epsilon. \tag{6.29}$$

The fact that g and $\phi_{\delta,\epsilon}$ are bounded by ψ_g implies

$$\int e^{(\alpha-1)\phi_{\delta,\epsilon}} dQ, \int e^{(\alpha-1)g} dQ \in [M_{g,-}, M_{g,+}], \quad \int e^{\alpha\phi_{\delta,\epsilon}} dP, \int e^{\alpha g} dP \in [N_{g,-}, N_{g,+}], \tag{6.30}$$

where $M_{g,\pm} \equiv \int e^{\pm|\alpha-1|\psi_g} dQ \in (0,\infty)$, $N_{g,\pm} \equiv \int e^{\pm|\alpha|\psi_g} dP \in (0,\infty)$. Using the fact that log

is $1/c$-Lipschitz on $[c,\infty)$ for all $c>0$, we can compute

$$\begin{aligned}(6.31)\qquad &\left|\frac{1}{\alpha-1}\log\int e^{(\alpha-1)g}dQ-\frac{1}{\alpha}\log\int e^{\alpha g}dP\right.\\ &\left.-\left(\frac{1}{\alpha-1}\log\int e^{(\alpha-1)\phi_{\delta,\epsilon}}dQ-\frac{1}{\alpha}\log\int e^{\alpha\phi_{\delta,\epsilon}}dP\right)\right|\\ \le&\frac{1}{|\alpha-1|M_{g,-}}\left|\int e^{(\alpha-1)g}dQ-\int e^{(\alpha-1)\phi_{\delta,\epsilon}}dQ\right|+\frac{1}{|\alpha|N_{g,-}}\left|\int e^{\alpha g}dP-\int e^{\alpha\phi_{\delta,\epsilon}}dP\right|\\ \le&\frac{1}{|\alpha-1|M_{g,-}}\int_{K_\delta}\left|e^{(\alpha-1)g}-e^{(\alpha-1)\phi_{\delta,\epsilon}}\right|dQ+\frac{2}{|\alpha-1|M_{g,-}}\int e^{|\alpha-1|\psi_g}1_{K_\delta^c}dQ\\ &+\frac{1}{|\alpha|N_{g,-}}\int_{K_\delta}\left|e^{\alpha g}-e^{\alpha\phi_{\delta,\epsilon}}\right|dP+\frac{2}{|\alpha|N_{g,-}}\int e^{|\alpha|\psi_g}1_{K_\delta^c}dP.\end{aligned}$$

Under assumption 1(a) and restricting to $\epsilon\le 1$, we can use (6.28) to bound $|\phi_{\delta,\epsilon}|\le\|g\|_\infty+1$ on K_δ, and so $|e^{cg}-e^{c\phi_{\delta,\epsilon}}|1_{K_\delta}\le|c|e^{|c|(\|g\|_\infty+1)}\epsilon$ for $c\in\mathbb{R}$. Under assumption 2(a) we can use (6.29) and Hölder's inequality to bound

$$\begin{aligned}(6.32)\qquad &\frac{1}{|\alpha-1|M_{g,-}}\int_{K_\delta}\left|e^{(\alpha-1)g}-e^{(\alpha-1)\phi_{\delta,\epsilon}}\right|dQ+\frac{1}{|\alpha|N_{g,-}}\int_{K_\delta}\left|e^{\alpha g}-e^{\alpha\phi_{\delta,\epsilon}}\right|dP\\ \le&\frac{1}{M_{g,-}}\int_{K_\delta}e^{|\alpha-1|\psi_g}|g-\phi_{\delta,\epsilon}|dQ+\frac{1}{N_{g,-}}\int_{K_\delta}e^{|\alpha|\psi_g}|g-\phi_{\delta,\epsilon}|dP\\ \le&\frac{1}{M_{g,-}}\left(\int e^{q|\alpha-1|\psi_g}dQ\right)^{1/q}\epsilon+\frac{1}{N_{g,-}}\left(\int e^{q|\alpha|\psi_g}dP\right)^{1/q}\epsilon.\end{aligned}$$

In either case, we find

$$\begin{aligned}(6.33)\qquad &\frac{1}{\alpha-1}\log\int e^{(\alpha-1)g}dQ-\frac{1}{\alpha}\log\int e^{\alpha g}dP\\ \le&\sup_{\phi\in\Phi}\left\{\frac{1}{\alpha-1}\log\int e^{(\alpha-1)\phi}dQ-\frac{1}{\alpha}\log\int e^{\alpha\phi}dP\right\}+D_{\delta,\epsilon},\\ &D_{\delta,\epsilon}\equiv D_g\epsilon+\frac{2}{|\alpha-1|M_{g,-}}\int_{K_\delta^c}e^{|\alpha-1|\psi_g}dQ+\frac{2}{|\alpha|N_{g,-}}\int_{K_\delta^c}e^{|\alpha|\psi_g}dP,\end{aligned}$$

where $D_g\in(0,\infty)$ is given by

$$(6.34)\qquad D_g=M_{g,-}^{-1}e^{|\alpha-1|(\|g\|_\infty+1)}+N_{g,-}^{-1}e^{|\alpha|(\|g\|_\infty+1)}$$

under assumption 1 and by

$$(6.35)\qquad D_g=M_{g,-}^{-1}\left(\int e^{q|\alpha-1|\psi_g}dQ\right)^{1/q}+N_{g,-}^{-1}\left(\int e^{q|\alpha|\psi_g}dP\right)^{1/q}$$

under assumption 2. Under either set of assumptions, we have $e^{|\alpha-1|\psi_g}\in L^1(Q)$ and $e^{|\alpha|\psi_g}\in L^1(P)$. Combining this fact with $Q(K_\delta^c),P(K_\delta^c)\le 2\delta$, we can use the dominated convergence

theorem for convergence in measure to compute

$$\lim_{\delta \searrow 0} \int_{K_\delta^c} e^{|\alpha-1|\psi_g} dQ = 0 = \lim_{\delta \searrow 0} \int_{K_\delta^c} e^{|\alpha|\psi_g} dP \tag{6.36}$$

(here it is important that ψ_g is independent of δ). Therefore, taking $\epsilon, \delta \searrow 0$, we obtain

$$\begin{aligned} &\frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} dQ - \frac{1}{\alpha} \log \int e^{\alpha g} dP \\ &\leq \sup_{\phi \in \Phi} \left\{ \frac{1}{\alpha-1} \log \int e^{(\alpha-1)\phi} dQ - \frac{1}{\alpha} \log \int e^{\alpha\phi} dP \right\}. \end{aligned} \tag{6.37}$$

This holds for all $g \in \mathrm{Lip}_b(\Omega)$, and so

$$\begin{aligned} &\sup_{g \in \mathrm{Lip}_b(\Omega)} \left\{ \frac{1}{\alpha-1} \log \int e^{(\alpha-1)g} dQ - \frac{1}{\alpha} \log \int e^{\alpha g} dP \right\} \\ &\leq \sup_{\phi \in \Phi} \left\{ \frac{1}{\alpha-1} \log \int e^{(\alpha-1)\phi} dQ - \frac{1}{\alpha} \log \int e^{\alpha\phi} dP \right\}. \end{aligned} \tag{6.38}$$

Using Theorem 3.1 with $\Gamma = \mathrm{Lip}_b(\Omega)$, we see that the left-hand side of (6.38) equals $R_\alpha(Q\|P)$. Theorem 3.1 with $\Gamma = \mathcal{M}(\Omega)$ implies that the right-hand side of (6.38) is bounded above by $R_\alpha(Q\|P)$. This proves the claim. ∎

Proof of Theorem 4.6. Compactness of Θ_k and continuity of ϕ_k in θ imply $\widehat{R}_\alpha^{n,k}(Q\|P)$ are real-valued and measurable. For $k \in \mathbb{Z}^+$ define

$$R_\alpha^k(Q\|P) \equiv \sup_{\theta \in \Theta_k} \left\{ \frac{1}{\alpha-1} \log \int e^{(\alpha-1)\phi_{k,\theta}} dQ - \frac{1}{\alpha} \log \int e^{\alpha\phi_{k,\theta}} dP \right\}. \tag{6.39}$$

By using the bound

$$\begin{aligned} &\left| R_\alpha^k(Q\|P) - \widehat{R}_\alpha^{n,k}(Q\|P) \right| \\ &\leq \frac{1}{|\alpha-1|} \sup_{\theta \in \Theta_k} \left| \log \left[\int e^{(\alpha-1)\phi_{k,\theta}} dQ \right] - \log \left[\frac{1}{n} \sum_{i=1}^n e^{(\alpha-1)\phi_k(X_i,\theta)} \right] \right| \\ &\quad + \frac{1}{|\alpha|} \sup_{\theta \in \Theta_k} \left| \log \left[\int e^{\alpha\phi_{k,\theta}} dP \right] - \log \left[\frac{1}{n} \sum_{i=1}^n e^{\alpha\phi_k(Y_i,\theta)} \right] \right| \end{aligned} \tag{6.40}$$

together with the facts that

$$\int e^{(\alpha-1)\phi_{k,\theta}} dQ \geq \int e^{-|\alpha-1|\psi_k} dQ, \quad \int e^{\alpha\phi_{k,\theta}} dP \geq \int e^{-|\alpha|\psi_k} dP, \ \ \theta \in \Theta_k, \tag{6.41}$$

and log is $1/c$-Lipschitz on $[c,\infty)$ for all $c>0$, we can compute the following for all $\eta>0$:

$$
\begin{aligned}
(6.42)\qquad & \left\{\left|R_\alpha^k(Q\|P)-\widehat{R}_\alpha^{n,k}(Q\|P)\right|\ge\eta\right\}\\
&\subset\left\{\sup_{\theta\in\Theta_k}\left|\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right]-\log\left[\frac{1}{n}\sum_{i=1}^n e^{(\alpha-1)\phi_k(X_i,\theta)}\right]\right|\ge|\alpha-1|\eta/2\right.\\
&\qquad\left.\text{and }\sup_{\theta\in\Theta_k}\left|\frac{1}{n}\sum_{i=1}^n e^{(\alpha-1)\phi_k(X_i,\theta)}-\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right|\le E_Q[e^{-|\alpha-1|\psi_k}]/2\right\}\\
&\cup\left\{\sup_{\theta\in\Theta_k}\left|\frac{1}{n}\sum_{i=1}^n e^{(\alpha-1)\phi_k(X_i,\theta)}-\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right|>E_Q[e^{-|\alpha-1|\psi_k}]/2\right\}\\
&\cup\left\{\sup_{\theta\in\Theta_k}\left|\log\left[\int e^{\alpha\phi_{k,\theta}}dP\right]-\log\left[\frac{1}{n}\sum_{i=1}^n e^{\alpha\phi_k(Y_i,\theta)}\right]\right|\ge|\alpha|\eta/2\right.\\
&\qquad\left.\text{and }\sup_{\theta\in\Theta_k}\left|\frac{1}{n}\sum_{i=1}^n e^{\alpha\phi_k(Y_i,\theta)}-\int e^{\alpha\phi_{k,\theta}}dP\right|\le E_P[e^{-|\alpha|\psi_k}]/2\right\}\\
&\cup\left\{\sup_{\theta\in\Theta_k}\left|\frac{1}{n}\sum_{i=1}^n e^{\alpha\phi_k(Y_i,\theta)}-\int e^{\alpha\phi_{k,\theta}}dP\right|>E_P[e^{-|\alpha|\psi_k}]/2\right\}\\
&\subset\left\{\sup_{\theta\in\Theta_k}\left|\frac{1}{n}\sum_{i=1}^n(e^{(\alpha-1)\phi_k(X_i,\theta)}-E_{\mathbb{P}}[e^{(\alpha-1)\phi_k(X_i,\theta)}])\right|\ge\epsilon_1\right\}\\
&\cup\left\{\sup_{\theta\in\Theta_k}\left|\frac{1}{n}\sum_{i=1}^n(e^{\alpha\phi_k(Y_i,\theta)}-E_{\mathbb{P}}[e^{\alpha\phi_k(Y_i,\theta)}])\right|\ge\epsilon_2\right\},\\
&\epsilon_1\equiv\min\{|\alpha-1|\eta E_Q[e^{-|\alpha-1|\psi_k}]/4,E_Q[e^{-|\alpha-1|\psi_k}]/2\},\\
&\epsilon_2\equiv\min\{|\alpha|\eta E_P[e^{-|\alpha|\psi_k}]/4,E_P[e^{-|\alpha|\psi_k}]/2\}.
\end{aligned}
$$

For all $\theta\in\Theta_k$ we have $|e^{(\alpha-1)\phi_k(X_i,\theta)}|\le e^{|\alpha-1|\psi_k(X_i)}\in L^1(\mathbb{P})$ and $|e^{\alpha\phi_k(Y_i,\theta)}|\le e^{|\alpha|\psi_k(Y_i)}\in L^1(\mathbb{P})$; therefore, the uniform law of large numbers (see Lemma 3.10 in [52]) implies convergence in probability:

$$
\begin{aligned}
(6.43)\qquad &\lim_{n\to\infty}\mathbb{P}\left(\sup_{\theta\in\Theta_k}\left|n^{-1}\sum_{i=1}^n\left(e^{(\alpha-1)\phi_k(X_i,\theta)}-E_{\mathbb{P}}\left[e^{(\alpha-1)\phi_k(X_i,\theta)}\right]\right)\right|\ge\epsilon\right)=0,\\
&\lim_{n\to\infty}\mathbb{P}\left(\sup_{\theta\in\Theta_k}\left|n^{-1}\sum_{i=1}^n\left(e^{\alpha\phi_k(Y_i,\theta)}-E_{\mathbb{P}}\left[e^{\alpha\phi_k(Y_i,\theta)}\right]\right)\right|\ge\epsilon\right)=0
\end{aligned}
$$

for all $\epsilon>0$. Combined with (6.42), this implies

$$
(6.44)\qquad \lim_{n\to\infty}\mathbb{P}\left(\left|R_\alpha^k(Q\|P)-\widehat{R}_\alpha^{n,k}(Q\|P)\right|\ge\eta\right)=0.
$$

To finish, consider the following two cases.

1. $R_\alpha(Q\|P) < \infty$: Fix $\delta > 0$. The assumption (4.6) implies that there exists K such that for $k \geq K$ we have $R_\alpha(Q\|P) - \delta/2 \leq R_\alpha^k(Q\|P) \leq R_\alpha(Q\|P)$. Hence, for $k \geq K$, (6.44) implies

$$\mathbb{P}(|R_\alpha(Q\|P) - \widehat{R}_\alpha^{n,k}(Q\|P)| \geq \delta) \leq \mathbb{P}(|R_\alpha^k(Q\|P) - \widehat{R}_\alpha^{n,k}(Q\|P)| \geq \delta/2) \to 0 \tag{6.45}$$

as $n \to \infty$. This proves the claimed result when $R_\alpha(Q\|P) < \infty$.

2. $R_\alpha(Q\|P) = \infty$: Fix $M > 0$ and $\delta > 0$. The assumption (4.6) implies that there exists K such that for all $k \geq K$ we have

$$R_\alpha^k(Q\|P) \equiv \sup_{\theta \in \Theta_k} \left\{ \frac{1}{\alpha - 1} \log \int e^{(\alpha-1)\phi_{k,\theta}} dQ - \frac{1}{\alpha} \log \int e^{\alpha \phi_{k,\theta}} dP \right\} \geq M + \delta. \tag{6.46}$$

Hence for $k \geq K$ we can use (6.44) to obtain

$$\mathbb{P}(\widehat{R}_\alpha^{n,k}(Q\|P) \leq M) \leq \mathbb{P}\left(|R_\alpha^k(Q\|P) - \widehat{R}_\alpha^{n,k}(Q\|P)| \geq \delta\right) \to 0 \tag{6.47}$$

as $n \to \infty$. This proves the claimed result when $R_\alpha(Q\|P) = \infty$. ∎

REFERENCES

[1] V. Anantharam, *A variational characterization of Rényi divergences*, in Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, Washington, DC, 2017, pp. 893–897.

[2] C. Anil, J. Lucas, and R. Grosse, *Sorting out Lipschitz function approximation*, in Proceedings of Machine Learning Research 97 (Long Beach, CA), PMLR, 2019, pp. 291–301, http://proceedings.mlr.press/v97/anil19a.html.

[3] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in Proceedings of the 34th International Conference on Machine Learning (Sydney, Australia), D. Precup and Y. W. Teh, eds., Proceedings of Machine Learning Research 70, PMLR, 2017, pp. 214–223.

[4] R. Atar, A. Budhiraja, P. Dupuis, and R. Wu, *Robust Bounds and Optimization at the Large Deviations Scale for Queueing Models via Rényi Divergence*, preprint, https://arxiv.org/abs/2001.02110, 2020.

[5] R. Atar, K. Chowdhary, and P. Dupuis, *Robust bounds on risk-sensitive functionals via Rényi divergence*, SIAM/ASA J. Uncertain. Quantif., 3 (2015), pp. 18–33, https://doi.org/10.1137/130939730.

[6] L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy, *PAC-Bayesian bounds based on the Rényi divergence*, in Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (Cadiz, Spain), A. Gretton and C. C. Robert, eds., Proceedings of Machine Learning Research 51, PMLR, 2016, pp. 435–444, http://proceedings.mlr.press/v51/begin16.html.

[7] 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 (Stockholm, Sweden), J. Dy and A. Krause, eds., Proceedings of Machine Learning Research 80, PMLR, 2018, pp. 531–540, http://proceedings.mlr.press/v80/belghazi18a.html.

[8] M. Berta, O. Fawzi, and M. Tomamichel, *On variational expressions for quantum relative entropies*, Lett. Math. Phys., 107 (2017), pp. 2239–2265.

[9] J. Birrell, M. A. Katsoulakis, and Y. Pantazis, *Optimizing Variational Representations of Divergences and Accelerating Their Statistical Estimation*, preprint, https://arxiv.org/abs/2006.08781, 2020.

[10] J. Bucklew, *Introduction to Rare Event Simulation*, Springer Series in Statistics, Springer, New York, 2004.

[11] A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods*, Probab. Theory Stoch. Model. 94, Springer, New York, 2019.

[12] A. Butte and K. IS., *Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements*, in Proceedings of the Pacific Symposium on Biocomputing 2000, World Scientific, Hackensack, NJ, 2000, pp. 418–429.

[13] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signal Systems, 2 (1989), pp. 303–314.
[14] M. D. Donsker and S. R. S. Varadhan, *Asymptotic evaluation of certain Markov process expectations for large time* IV, Comm. Pure Appl. Math., 36 (1983), pp. 183–212, https://doi.org/10.1002/cpa.3160360204.
[15] R. Dudley, *Uniform Central Limit Theorems*, 2nd ed., Cambridge Stud. Adv. Math. 142, Cambridge University Press, New York, 2014.
[16] P. Dupuis and R. Ellis, *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley Ser. Probab. Stat., John Wiley & Sons, New York, 1997.
[17] P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet, *Sensitivity analysis for rare events based on Rényi divergence*, Ann. Appl. Probab., 30 (2020), pp. 1507–1533, https://doi.org/10.1214/19-AAP1468.
[18] G. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., Pure Appl. Math. (N. Y.), John Wiley & Sons, New York, 1999.
[19] S. Gao, G. V. Steeg, and A. Galstyan, *Efficient estimation of mutual information for strongly dependent variables*, in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (San Diego, CA), G. Lebanon and S. V. N. Vishwanathan, eds., Proceedings of Machine Learning Research 38, PMLR, 2015, pp. 277–286, http://proceedings.mlr.press/v38/gao15.html.
[20] S. Ghadimi and G. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368, https://doi.org/10.1137/120880811.
[21] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99.
[22] M. Gil, F. Alajaji, and T. Linder, *Rényi divergence measures for commonly used univariate continuous distributions*, Inform. Sci., 249 (2013), pp. 124–131, https://doi.org/10.1016/j.ins.2013.06.018.
[23] 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, NIPS'14, Volume 2, MIT Press, Cambridge, MA, 2014, pp. 2672–2680.
[24] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, *Improved training of Wasserstein GANs*, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates, Red Hook, NY, 2017, pp. 5769–5779.
[25] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley, New York, 2004.
[26] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, and J. M. Robins, *Nonparametric von Mises estimators for entropies, divergences and mutual informations*, in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, Red Hook NY, 2015, pp. 397–405.
[27] P. Kidger and T. Lyons, *Universal Approximation with Deep Narrow Networks*, in Proceedings of Thirty Third Conference on Learning Theory, J. Abernethy and S. Agarwal, eds., Proceedings of Machine Learning Research 125, PMLR, 2020, pp. 2306–2327, http://proceedings.mlr.press/v125/kidger20a.html.
[28] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, preprint, https://arxiv.org/abs/1412.6980, 2014.
[29] J. B. Kinney and G. S. Atwal, *Equitability, mutual information, and the maximal information coefficient*, Proc. Natl. Acad. Sci. USA, 111 (2014), pp. 3354–3359, https://doi.org/10.1073/pnas.1309933111.
[30] N. Kwak and C.-H. Choi, *Input feature selection by mutual information based on Parzen window*, IEEE Trans. Pattern Anal. Mach. Intell., 24 (2002), pp. 1667–1671.
[31] T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations: A Mathematical Perspective*, Imperial College Press, London, 2010.
[32] Y. Li and R. E. Turner, *Rényi divergence variational inference*, in Advances in Neural Information Processing Systems, NeurIPS, San Diego, CA, 2016, pp. 1073–1081.
[33] F. Liese and I. Vajda, *Convex Statistical Distances*, Teubner-Texte Math. 95, Teubner, Leipzig, Germany, 1987.
[34] F. Liese and I. Vajda, *On divergences and informations in statistics and information theory*, IEEE

Trans. Inform. Theory, 52 (2006), pp. 4394–4412, https://doi.org/10.1109/TIT.2006.881731.
[35] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, *The expressive power of neural networks: A view from the width*, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates., Red Hook, NY, 2017, pp. 6232–6240.
[36] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, *Multimodality image registration by maximization of mutual information*, IEEE Trans. Med. Imaging, 16 (1997), pp. 187–198.
[37] X. Nguyen, M. J. Wainwright, and M. I. Jordan, *Estimating divergence functionals and the likelihood ratio by convex risk minimization*, IEEE Trans. Inform. Theory, 56 (2010), pp. 5847–5861.
[38] S. Nowozin, B. Cseke, and R. Tomioka, *F-GAN: Training generative neural samplers using variational divergence minimization*, in Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, Curran Associates, Red Hook, NY, 2016, pp. 271–279.
[39] L. Paninski, *Estimation of entropy and mutual information*, Neural Comput., 15 (2003), pp. 1191–1253, https://doi.org/10.1162/089976603321780272.
[40] Y. Pantazis, D. Paul, M. Fasoulakis, Y. Stylianou, and M. Katsoulakis, *Cumulant GAN*, preprint, https://arxiv.org/abs/2006.06625, 2020.
[41] S. Park, C. Yun, J. Lee, and J. Shin, *Minimum width for universal approximation*, in International Conference on Learning Representations, Vienna, Austria, 2021, https://openreview.net/forum?id=O-XJwyoIF-k.
[42] A. Pinkus, *Approximation theory of the MLP model in neural networks*, Acta Numer., 8 (1999), pp. 143–195, https://doi.org/10.1017/S0962492900002919.
[43] B. Póczos, L. Xiong, and J. Schneider, *Nonparametric divergence estimation with applications to machine learning on distributions*, in Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (Arlington, VA), UAI'11, AUAI Press, 2011, pp. 599–608, http://dl.acm.org/citation.cfm?id=3020548.3020618.
[44] S. Rahman, *The f-sensitivity index*, SIAM/ASA J. Uncertain. Quantif., 4 (2016), pp. 130–162, https://doi.org/10.1137/140997774.
[45] S. J. Reddi, S. Kale, and S. Kumar, *On the Convergence of Adam and Beyond*, preprint, https://arxiv.org/abs/1904.09237, 2019.
[46] A. Rényi, *On Measures of Entropy and Information*, Tech. report, Hungarian Academy of Sciences, Budapest, Hungary, 1961.
[47] G. Rubino and B. Tuffin, eds., *Rare Event Simulation using Monte Carlo Methods*, Appl. Probab. Stat. 73, Wiley, New York, 2009.
[48] A. Ruderman, M. D. Reid, D. García-García, and J. Petterson, *Tighter variational representations of f-divergences via restriction to probability measures*, in Proceedings of the 29th International Conference on Machine Learning (Madison, WI), ICML'12, 2012, Omnipress, pp. 1155–1162.
[49] F. Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Progr. Nonlinear Differential Equations Appl. 87, Springer, Cham, 2015.
[50] J. Song and S. Ermon, *Understanding the Limitations of Variational Mutual Information Estimators*, preprint, https://arxiv.org/abs/1910.06222, 2020.
[51] N. Tishby, F. C. Pereira, and W. Bialek, *The Information Bottleneck Method*, preprint, https://arxiv.org/abs/physics/0004057, 2000.
[52] S. van de Geer, R. Gill, B. Ripley, S. Ross, B. Silverman, D. Williams, and M. Stein, *Empirical Processes in M-Estimation*, Cambridge Ser. Stat. Probab. Math., Cambridge University Press, Cambridge, UK, 2000.
[53] T. Van Erven and P. Harremos, *Rényi divergence and Kullback-Leibler divergence*, IEEE Trans. Inform. Theory, 60 (2014), pp. 3797–3820.
[54] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Camb. Ser. Stat. Probab. Math. 47, Cambridge University Press, Cambridge, UK, 2018, https://doi.org/10.1017/9781108231596.
[55] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang, *A unified analysis of stochastic momentum methods for deep learning*, in Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI'18, AAAI Press, Palo Alto, CA, 2018, pp. 2955–2961.