VARIATIONAL REPRESENTATIONS AND NEURAL NETWORK ESTIMATION OF RÉNYI DIVERGENCES

A PREPRINT

Jeremiah Birrell
Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, MA 01003, USA
birrell@math.umass.edu

Paul Dupuis
Division of Applied Mathematics
Brown University
Providence, RI 02912, USA
dupuis@dam.brown.edu

Markos A. Katsoulakis
Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, MA 01003, USA
markos@math.umass.edu

Luc Rey-Bellet
Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, MA 01003, USA
luc@math.umass.edu

Jie Wang
Department of Mathematics and Statistics
University of Massachusetts Amherst
Amherst, MA 01003, USA
wang@math.umass.edu

July 21, 2021

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 5000 dimensions.

***K*eywords** Rényi divergence, variational representation, neural network estimator

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 [26], medical image registration [37], feature selection [31], genomic clustering [12], the information bottleneck method [53], independence testing [30], and in the analysis and design of generative adversarial networks (GANs) [24, 39, 3, 25, 41].

Estimation of divergences from data is known to be a difficult problem [40, 19]. Density-estimator based methods such as those in [44, 27] are known to work best in low dimensions. 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 [38, 49, 9], that scale better with dimension. The family of Rényi divergences, first introduced in [47], provide 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 [33], uncertainty quantification for rare events [17], and naturally arise in coding theory and hypothesis testing (see [54] 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 non-absolutely 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 [52], 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. Eq. (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 non-compact support, 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 Section 6.3. 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 [54] 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 Eq. (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 [41]. 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 $dQ = qd\nu$ and $dP = pd\nu$. Then

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

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

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

Eq. (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 following 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 [54] 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 [23]. Rényi divergences are also connected with the family of f-divergences; see [35].

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 Eq. (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 Eq. (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 Section 6.3.
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 Eq. (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 Eq. (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 Eq. (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 short-hands 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 [34],

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

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 [32]) 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 [30] and the development of GANs [24, 39, 3, 25, 41]. Density-estimator based methods for estimating divergences are known to be effective primarily in low-dimensions (see [44, 27] 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., [48, 10, 11]) are still applicable here.

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

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

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 Eq. (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:

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, 43]. Results for deep networks with bounded width are also known; see [28] for Eq. (4.3) and [36, 42] for Eq. (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 Section 6.3.

1. Measures with compact support: Let $\Omega \subset \mathbb{R}^m$ be compact, 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. Non-compact 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. Non-compact 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 [42] 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 is 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 non-compact 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 Eq. (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 Section 6.3 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 the proofs in Section 6.3. 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* $\alpha \in \mathbb{R}\setminus\{0,1\}$, 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 iid and* Q*-distributed,* Y_i *are iid 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* $k \ge K$ *we have*

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

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 allows us to prove consistency when Q and P to have non-compact 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 Eq. (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 non-asymptotic uniform law of large numbers-type result; see [55] for details on these tools. We include a proof of (4.12) in Section 6.4.

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) [21, 22, 56]. 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 [29, 46], an adaptive learning-rate SGD algorithm, to search for the optimum. All computations were performed in TensorFlow.

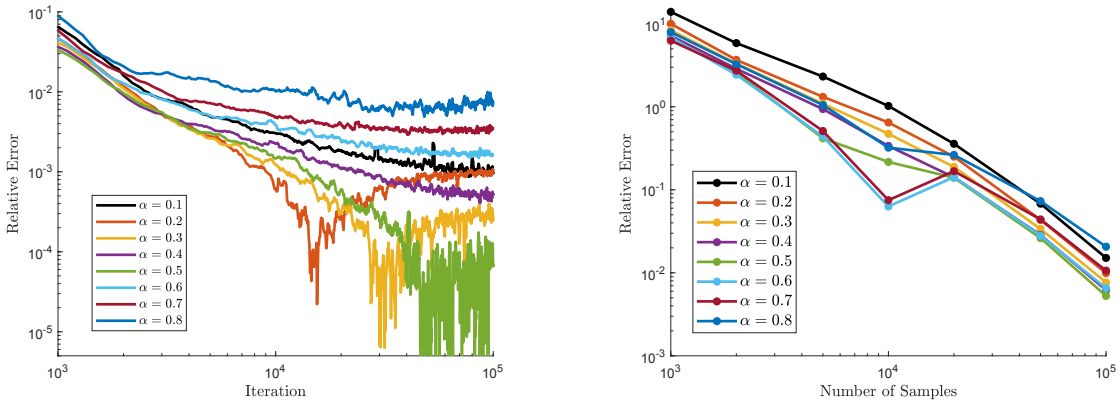

Figure 1: Left: Relative error of Rényi divergence estimators (4.9) between the distributions of $h(X)$ and $h(Y)$, where X and Y are 4-dimensional Gaussians (with means $\mu_p = 0$, $\mu_q = (2,0,0,0)$ and covariance matrices $\Sigma_p = I$, $\Sigma_q = diag(1.5, 0.7, 2, 1)$) and $h : \mathbb{R}^4 \to \mathbb{R}^{5000}$ is a nonlinear map. Specifically, we let $h_i(x) = x_i$ for $i = 1, ..., 4$ (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 $j_{k,i} \in \{1, ..., 4\}$; the parameters of A and the $c_{k,i}$'s were randomly selected at the start of each run (all components are iid $N(0,1)$). The indices $j_{k,i}$ were also randomly selected at the start of each run (iid $Unif(\{1, ..., 4\})$). 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 1000 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 10000 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.

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 (non-linear) 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 non-linear map. If h is an embedding (in particular, it must be one-to-one) then the data processing inequality (see Theorem 14 in [35]) 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. 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,

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

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., [45]). 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 component-wise 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 non-parametric methods (again, see [27, 7]).

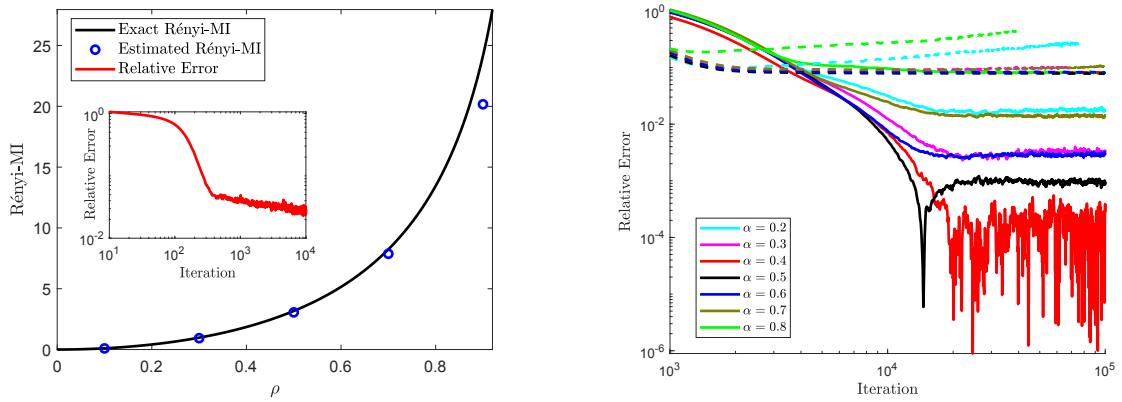

Figure 2: Left: Estimation of Rényi-based mutual information (5.1) with $\alpha = 1/2$ between 20-dimensional correlated Gaussians with component-wise correlation ρ. We used a neural network with one hidden layer of 256 nodes and training was performed with a minibatch size of 1000. We show the Rényi-MI as a function of ρ after 10000 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 1000 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 Eq. (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 Eq. (1.3) of [5]). Though the right hand side of Eq. (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]). Eq. (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 Eq. (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 Eq. (6.2) can be given by using a different variational formula for the Rényi divergences, which can be found in Theorem 30 of [54] and also in Theorem 1 of [1].

Proof of Eq. (6.2). We separate the proof into two cases.
1) $\alpha>1$: If $Q\not\ll P$ the result is trivial (see Eq. (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 $dP=pd\nu$, $dQ=qd\nu$ 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

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

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

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

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

We now use Eq. (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 Eq. (3.1) for all $\alpha > 1$ and all P, Q, then, using Eq. (2.2) and reindexing $g \to -g$ in the supremum, one finds that Eq. (3.1) also holds for all $\alpha < 0$. So we only need to consider the cases $\alpha \in (0,1)$ and $\alpha > 1$.

Eq. (6.2) immediately implies

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

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 $Q(A) \neq 0$ and define $g_n = n1_A$. The definition (6.7) implies

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

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 Eq. (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 Eq. (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 Eq. (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 Eq. (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 $f_{n,m}(q/p)^\alpha \leq f_{\infty,m}(q/p)^\alpha$. We thereby obtain

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

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

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

Therefore the claim is proven in case 3, and the proof of Eq. (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 $\|g_n\|_\infty \leq \|g\|_\infty$ and $g_n = g$ on F_n. Therefore

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

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

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

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

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

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 $\|g_n\|_\infty \leq \|g\|_\infty$ (see Box 1.5 on page 6 of [50]). The result then follows from a similar computation 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

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

for all $g \in \mathcal{M}(\Omega)$. The equality (3.1) then follows by combining Eq. (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 $g_{n,m} = -n1_{g<-n} + g1_{-n\leq g\leq m} + m1_{g>m}$. These are bounded and so Theorem 3.1 implies

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

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

1) $\alpha > 1$: If $\int e^{\alpha g} dP = \infty$ then Eq. (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 $\alpha > 1$, Eq. (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 $g_m \equiv g1_{g\leq m} + m1_{g>m}$ and $e^{cg_{n,m}} \leq e^{cm}$ for all n. Therefore the dominated convergence theorem implies

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

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

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

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

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

(note that we are in the sub-case 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 Eq. (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 $e^{cg_{n,n}} \leq 1 + e^{cg}$ and $\lim_{n\to\infty} e^{cg_{n,n}} = e^{cg}$. Therefore the dominated convergence theorem implies that

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

This proves Eq. (6.20) in case 3 and thus completes the proof of Eq. (3.1) when $\Gamma = \mathcal{M}(\Omega)$. Eq. (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 Eq. (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 1a and 2a imply that for $\phi \in \Phi$ there exists $\psi \in \Psi$ with $|\phi| \leq \psi$. Either of the integrability assumptions 1b - 1c or 2b - 2c then imply that all expectations on the right hand side of Eq. (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 1a and 2a imply that there exists $\psi_g \in \Psi$ such that $|g| \leq \psi_g$ and for all $\delta, \epsilon > 0$ there exists $\phi_{\delta,\epsilon} \in \Phi$ with $|\phi_{\delta,\epsilon}| \leq \psi_g$ and, in the case of 1a,

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

while in the case of 2a 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

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

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

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

In either case, we find

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

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

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

under assumption 1 and by

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

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 $Q(K_\delta^c),P(K_\delta^c)\leq 2\delta$ 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 implies $\widehat{R}_\alpha^{n,k}(Q\|P)$ 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

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

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 $\eta>0$:

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

For all $\theta\in\Theta_k$ we have $|e^{(\alpha-1)\phi_k(X_i,\theta)}|\leq e^{|\alpha-1|\psi_k(X_i)}\in L^1(\mathbb{P})$ and $|e^{\alpha\phi_k(Y_i,\theta)}|\leq e^{|\alpha|\psi_k(Y_i)}\in L^1(\mathbb{P})$, therefore the uniform law of large numbers (see Lemma 3.10 in [20]) implies convergence in probability:

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

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

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

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 $k\geq K$, Eq. (6.44) implies

$$
\mathbb{P}(|R_\alpha(Q\|P)-\widehat{R}_\alpha^{n,k}(Q\|P)|\geq\delta)\leq\mathbb{P}(|R_\alpha^k(Q\|P)-\widehat{R}_\alpha^{n,k}(Q\|P)|\geq\delta/2)\to0 \qquad (6.45)
$$

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

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

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 $R_\alpha(Q\|P) = \infty$.

□

6.3 Applying Theorem 4.6 to Several Classes of Neural Networks

Here we prove consistency of the neural network estimators that were discussed in Section 4.1. Specifically, we show they satisfy all of the properties required to apply Theorem 4.6.

1. Measures with compact support: Let $\Omega \subset \mathbb{R}^m$ be compact, 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 parameter values restricted to $[-a_k, a_k]$, where $a_k \nearrow \infty$. Let Ψ be the set of positive constants. Then the assumptions of Theorem 4.6 are satisfied and hence the estimator (4.1) is consistent.

   Proof. To see this, first note that compactness of Ω implies that every $\phi \in \Phi$ is bounded and so property 1 of Definition 4.1 is trivial. Property 2 of Definition 4.1 easily follows from the universal approximation property (4.3) applied to the compact set Ω. Therefore Φ has the Ψ-bounded L^∞ approximation property. Assumptions 1b and 1c of Lemma 4.3 are trivial, as $\psi \in \Psi$ are bounded, and so we have (4.5). Eq. (4.6) then follows from the fact that Φ_k increase to Φ. The remaining items 2 - 5 in Assumption 4.4 then follow from compactness of K and boundedness of $\psi \in \Psi$. □

2. Non-compact 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 $a_k \nearrow \infty$, and let Ψ be the set of positive constants. 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.

   Proof. To prove this, first note that boundedness of σ implies boundedness of every $\phi \in \Phi$. Therefore 1 of Definition 4.1 holds. For any $g \in \mathrm{Lip}_b(\mathbb{R}^m)$ we can find $b \in \mathbb{R}$, $a > 0$ such that the range of $(g - b)/a$ is contained in (c, d), and hence $\sigma^{-1}((g - b)/a)$ is well-defined and continuous. Let L be the Lipschitz constant for σ on (c, d) and define $\psi \equiv \max\{|b| + a\|\sigma\|_\infty, \|g\|_\infty\} \in \Psi$. Then $|g| \leq \psi$ and, by the universal approximation property in [43], for any compact $K \subset \mathbb{R}^m$ and any $\epsilon > 0$ there exists a network with one hidden layer, ϕ_ϵ, that satisfies

   $$\sup_K |\sigma^{-1}((g - b)/a) - \phi_\epsilon| \leq \epsilon/(aL)\,. \tag{6.48}$$

   Therefore

   $$\sup_K |g - (b + a\sigma(\phi_\epsilon))| \leq aL \sup_K |\sigma^{-1}((g - b)/a) - \phi_\epsilon| \leq \epsilon\,. \tag{6.49}$$

   We have $b + a\sigma(\phi_\epsilon) \in \Phi$ (as we have simply added a second hidden layer to the network ϕ_ϵ) and $|b + a\sigma(\phi_\epsilon)| \leq \psi$. This completes the proof of the Ψ-bounded L^∞ approximation property. Properties 1b and 1c of Lemma 4.3 are trivial and so we can conclude (4.5). The sets Φ_k increase to Φ and so, combined with (4.5), we can conclude (4.6). The remaining items in Assumption 4.4 hold due to boundedness of $\psi \in \Psi$, uniform boundedness of the parameter values for $\phi \in \Phi_k$, and boundedness of the activation function. □

3. Non-compact support, unbounded Lipschitz activation function: Let $p \in (1, \infty)$ and $\Omega = \mathbb{R}^m$, equipped with the ℓ^p-norm. Let Q and P be probability measures on Ω that have finite moment generating functions everywhere and have densities dQ/dx and dP/dx that are bounded on compact sets. Define Φ be the family of neural networks obtained by using either the ReLU activation function or the GroupSort activation with group size 2 (see [2]). Let $\Phi_k \subset \Phi$ be the set of networks with depth and width bounded by k (for ReLU, one can alternatively use networks with depth equal to 3) and with parameter values restricted to $[-a_k, a_k]$, where $a_k \nearrow \infty$. Finally, let $\Psi = \{x \mapsto a\|x\| + b : a, b \geq 0\}$, where $\|\cdot\|$ denotes the ℓ^p-norm. Then the assumptions of Theorem 4.6 are satisfied, and hence the estimator (4.1) is consistent.

Proof. (a) First consider the case of ReLU activation functions. We will show that Φ has the Ψ-bounded L^∞ approximation property (Definition 4.1). First, all $\phi \in \Phi$ are Lipschitz, hence are bounded by $x \mapsto a\|x\| + b$ for some $a, b \geq 0$. Next, fix $g \in \mathrm{Lip}_b(\Omega)$. By the universal approximation property (4.3) (see [13, 43]), for all compact $K \subset \Omega$ and all $\epsilon > 0$ there exists a neural network, $\phi_{\epsilon,K}$, with one hidden layer and ReLU activation that satisfies

$$\sup_{x\in K} |g(x) - \phi_{\epsilon,K}(x)| < \min\{\epsilon, 1\}\,. \tag{6.50}$$

Define $\psi = \|g\|_\infty + 1 \in \Psi$ and

$$\widetilde{\phi}_{\epsilon,K} = \mathrm{ReLU}(-\mathrm{ReLU}(\|g\|_\infty + 1 - \phi_{\epsilon,K}) + 2(\|g\|_\infty + 1)) - (\|g\|_\infty + 1)\,. \tag{6.51}$$

Note that $\widetilde{\phi}_{\epsilon,K} \in \Phi$ has depth equal to 3. We have $|g| \leq \psi$, $|\widetilde{\phi}_{\epsilon,K}| \leq \psi$, and

$$\sup_{x\in K} |g(x) - \widetilde{\phi}_{\epsilon,K}(x)| = \sup_{x\in K} |g(x) - \phi_{\epsilon,K}(x)| < \epsilon\,. \tag{6.52}$$

This proves the Ψ-bounded L^∞ approximation property. Properties 1b - 1c of Lemma 4.3 follow from the assumption that Q and P have finite moment generating functions everywhere. Therefore we conclude (4.5). Items 1 and 2 in Assumption 4.4 follows from the definition of Φ_k, as in the previous cases. Item 3 follows from the fact that the activation is Lipschitz and the network parameters, depth, and width of $\phi \in \Phi_k$ are uniformly bounded. Finally, 4 and 5 are implied by 1b and 1c from Lemma 4.3, which were shown above.

(b) Finally, we consider the GroupSort case. We start by showing the Ψ-bounded $L^p(\mathcal{Q})$ approximation property (Definition 4.2), where $\mathcal{Q} = \{Q, P\}$. Item 1 of Definition 4.2 follows from the fact that every $\phi \in \Phi$ is L-Lipschitz for some $L \geq 0$ and hence $|\phi(x)| \leq L\|x\| + |\phi(0)| \in \Psi$. Let $g \in \mathrm{Lip}_b(\Omega)$ with Lipschitz constant L and define $\psi(x) = L\|x\| + \|g\|_\infty + L + 1$. Then $\psi \in \Psi$, $|g| \leq \psi$, and, using Theorem 3 in [2], we see that for any compact $K \subset \mathbb{R}^m$ and any $j \in \mathbb{Z}^+$ there exists $\phi_j \in \Phi$ that is L-Lipschitz and satisfies

$$\left(\int_{K\cup\overline{B_1(0)}} |\phi_j - g|^p dx\right)^{1/p} \leq 1/j\,, \tag{6.53}$$

i.e., ϕ_j converges to g in $L^p(K \cup \overline{B_1(0)}, dx)$ ($\overline{B_1(0)}$ denotes the closed ball of ℓ^p-radius 1 centered at 0). Take a subsequence ϕ_{j_i} that converges to g a.e. on $K \cup \overline{B_1(0)}$. In particular, there exists x_0 with $\|x_0\| \leq 1$ and $\phi_{j_i}(x_0) \to g(x_0)$. Hence for $x \in \mathbb{R}^m$ we have

$$\begin{aligned} |\phi_{j_i}(x)| \leq& |\phi_{j_i}(x) - \phi_{j_i}(x_0)| + |\phi_{j_i}(x_0) - g(x_0)| + |g(x_0)| \\ \leq& L\|x\| + L + \|g\|_\infty + |\phi_{j_i}(x_0) - g(x_0)|\,. \end{aligned} \tag{6.54}$$

Therefore, if $\epsilon > 0$ then for all i sufficiently large we have $|\phi_{j_i}| \leq \psi$ and

$$\begin{aligned} &\sup_{\mu\in\mathcal{Q}} \left(\int_K |g - \phi_{j_i}|^p d\mu\right)^{1/p} \\ &\leq \max\{\sup_K |dQ/dx|, \sup_K |dP/dx|\}^{1/p} \left(\int_K |g - \phi_{j_i}|^p dx\right)^{1/p} < \epsilon\,. \end{aligned} \tag{6.55}$$

This proves the Ψ-bounded $L^p(\mathcal{Q})$-approximation property. Properties 2b - 2c of Lemma 4.3 follow from the assumption that Q and P have finite moment generating functions everywhere. Therefore we conclude (4.5). Items 1 and 2 in Assumption 4.4 follows from the definition of Φ_k, as in the previous cases. Item 3 follows from the fact that the activation function is Lipschitz and the network parameters, depth, and width of $\phi \in \Phi_k$ are uniformly bounded. Finally, 4 and 5 are implied by 2b and 2c from Lemma 4.3, which were shown above.

□

Remark 6.1. *For ReLU activation, our proof shows that neural networks with 3 hidden layers are sufficient to obtain consistency of the estimator; see Eq. (6.51). To the best of the authors' knowledge, it is an open question as to whether the Ψ-bounded L^∞ approximation property (or the Ψ-bounded $L^p(\mathcal{Q})$ approximation property) holds for networks with only one or two hidden layers. If so then consistency for one or two layer networks would follow by the same argument as above. The numerical results in Section 5 do suggest that shallow networks yield consistent estimators.*

Remark 6.2. *In the case of the GroupSort activation, it was crucial that the variational formula from Theorem 3.1 holds when the optimization is performed over the space $\Gamma = \mathrm{Lip}_b(\Omega)$. The required uniform bounds on the sequence of approximating functions by a fixed $\psi \in \Psi$ would not hold without the Lipschitz restriction.*

6.4 Complexity Proof

Here we will derive the complexity result (4.12) for Rényi divergence estimation; we use the same notation as in Theorem 4.6. Specifically, we derive finite sample bounds on how well the estimator $\widehat{R}_\alpha^{n,k}(Q\|P)$ (see Eq. (4.9)) approximates

$$R_\alpha^k(Q\|P) \equiv \sup_{\theta\in\Theta_k}\left\{\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta}}dP\right]\right\}. \tag{6.56}$$

One should compare this with the corresponding result for KL divergence estimation, Theorem 6 in [7], which has the same qualitative behavior in ϵ, δ, and d_k.

Theorem 6.3. *Let $\alpha \in \mathbb{R}\setminus\{0,1\}$, $\Theta_k \subset \mathbb{R}^{d_k}\cap\{\theta : \|\theta\| \le K_k\}$, and suppose $\phi_{k,\theta}$ is bounded by M_k and is L_k-Lipschitz in $\theta\in\Theta_k$. Then for all $\epsilon>0$, $\delta>0$ we have*

$$\mathbb{P}\left(|\widehat{R}_\alpha^{n,k}(Q\|P) - R_\alpha^k(Q\|P)| > \epsilon\right) \le \delta \tag{6.57}$$

whenever

$$n \ge \frac{32D_{\alpha,k}^2}{\epsilon^2}\left(d_k\log(16L_kK_k\sqrt{d_k}/\epsilon) + 2d_kM_k\max\{|\alpha|,|\alpha-1|\} + \log(4/\delta)\right), \tag{6.58}$$

where $D_{\alpha,k}\equiv\max\{e^{2|\alpha|M_k}/|\alpha|, e^{2|\alpha-1|M_k}/|\alpha-1|\}$.

Proof. First note that

$$\begin{aligned}
&|\widehat{R}_\alpha^{n,k}(Q\|P) - R_\alpha^k(Q\|P)| \qquad (6.59)\\
\le&\sup_{\theta\in\Theta_k}\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ_n\right]\right|\\
&+\sup_{\theta\in\Theta_k}\left|\frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta}}dP\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta}}dP_n\right]\right|.
\end{aligned}$$

Given $\eta>0$, take an open cover $B_\eta(\theta_j)$ of Θ_k. It is known that the minimal covering number satisfies [51]

$$N_\eta(\Theta_k) \le \left(\frac{2K_k\sqrt{d_k}}{\eta}\right)^{d_k}. \tag{6.60}$$

We let $\eta = \frac{\epsilon}{8L_k}e^{-2M_k\max\{|\alpha|,|\alpha-1|\}}$. For $\theta\in B_\eta(\theta_j)$ we can use the triangle inequality, the uniform bound on $\phi_{k,\theta}$, and the Lipschitz bounds on log, exp, and $\theta\mapsto\phi_{k,\theta}$ to compute

$$\begin{aligned}
&\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ_n\right]\right| \qquad (6.61)\\
\le&\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ\right]\right|\\
&+\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ_n\right]\right|\\
&+\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ_n\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ_n\right]\right|\\
\le&2L_ke^{2|\alpha-1|M_k}\eta + \left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ_n\right]\right|\\
\le&\frac{\epsilon}{4} + \left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ_n\right]\right|.
\end{aligned}$$

A similar bound applies to the second term on the right hand side of (6.59). Using a union bound and a Lipschitz bound we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{j=1,\dots,N_\eta(\Theta_k)}\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ_n\right]\right| > \epsilon/4\right) \qquad (6.62)\\
\le&\sum_j\mathbb{P}\left(\frac{1}{|\alpha-1|}e^{|\alpha-1|M_k}\left|\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ - \int e^{(\alpha-1)\phi_{k,\theta_j}}dQ_n\right| > \epsilon/4\right). \qquad (6.63)
\end{aligned}$$

Using Hoeffding's inequality we can bound

$$\mathbb{P}\left(\left|\int e^{(\alpha-1)\phi_{k,\theta_j}}\,dQ - \int e^{(\alpha-1)\phi_{k,\theta_j}}\,dQ_n\right| > c\right) \leq 2\exp\left(-\frac{nc^2}{2\exp(2|\alpha-1|M_k)}\right) \tag{6.64}$$

for all $c > 0$ and all j, hence

$$\begin{aligned}&\mathbb{P}\left(\max_{j=1,\ldots,N_\eta(\Theta_k)}\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}\,dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}\,dQ_n\right]\right| > \epsilon/4\right) \qquad (6.65)\\ \leq& 2N_\eta(\Theta_k)\exp\left(-\frac{n\epsilon^2}{32D_{\alpha,k}^2}\right).\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\mathbb{P}\left(\max_{j=1,\ldots,N_\eta(\Theta_k)}\left|\frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta_j}}\,dP\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta_j}}\,dP_n\right]\right| > \epsilon/4\right) \qquad (6.66)\\ \leq& 2N_\eta(\Theta_k)\exp\left(-\frac{n\epsilon^2}{32D_{\alpha,k}^2}\right).\end{aligned}$$

Combining these we can compute

$$\begin{aligned}&\mathbb{P}\left(|\widehat{R}_\alpha^{n,k}(Q\|P) - R_\alpha^k(Q\|P)| > \epsilon\right) \qquad (6.67)\\ \leq&\mathbb{P}\left(\sup_{\theta\in\Theta_k}\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta}}dQ_n\right]\right| > \epsilon/2\right)\\ &+\mathbb{P}\left(\sup_{\theta\in\Theta_k}\left|\frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta}}dP\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta}}dP_n\right]\right| > \epsilon/2\right)\\ \leq&\mathbb{P}\left(\max_j\left|\frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ\right] - \frac{1}{\alpha-1}\log\left[\int e^{(\alpha-1)\phi_{k,\theta_j}}dQ_n\right]\right| > \epsilon/4\right)\\ &+\mathbb{P}\left(\max_j\left|\frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta_j}}dP\right] - \frac{1}{\alpha}\log\left[\int e^{\alpha\phi_{k,\theta_j}}dP_n\right]\right| > \epsilon/4\right)\\ \leq&4N_\eta(\Theta_k)\exp\left(-\frac{n\epsilon^2}{32D_{\alpha,k}^2}\right)\\ \leq&4\left(\frac{2K_k\sqrt{d_k}}{\eta}\right)^{d_k}\exp\left(-\frac{n\epsilon^2}{32D_{\alpha,k}^2}\right).\end{aligned}$$

Finally, it is straightforward to show that

$$4\left(\frac{2K_k\sqrt{d_k}}{\eta}\right)^{d_k}\exp\left(-\frac{n\epsilon^2}{32D_{\alpha,k}^2}\right) \leq \delta \tag{6.68}$$

whenever n satisfies (6.58). □

Remark 6.4. *Though the general techniques for proving Theorem 6.3 are the same as those used in [7] to study KL divergence estimators, there are some technical errors in [7] that we have corrected in the above derivation. They have minimal impact on the qualitative behavior, with the exception of the behavior in the bound M_k; the correct behavior of the prefactor is exponential in M_k (in [7] it was stated to be M_k^2). This impacts both the KL and Rényi results. Specifically, the use of Hoeffding's inequality to obtain Eq. (49) in [7] must employ a bound on e^{T_θ} instead of a bound on T_θ, hence the right hand side of that bound should read $2N_\eta(\Theta)\exp(-\frac{\epsilon^2 n}{32\exp(2M)})$ (in their notation, there is no subscript k on M, Θ, etc.). This in turn implies that the KL complexity result, Eq. (45) in [7], should also have a factor of e^{2M} in place of M^2. Similar exponential behavior is also present in the result for Rényi divergences* (6.58)*.*

Acknowledgments

The research of J.B., M.K., and L. R.-B. was partially supported by NSF TRIPODS CISE-1934846. The research of M. K. and L. R.-B. was 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 P.D. was supported in part by the National Science Foundation (NSF) under the grant DMS-1904992 and by the Air Force Office of Scientific Research (AFOSR) under the grant FA-9550-18-1-0214. The research of J.W. was partially supported by the Defense Advanced Research Projects Agency (DARPA) EQUiPS program under the grant W911NF1520122.

References

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

[2] C. ANIL, J. LUCAS, AND R. GROSSE, *Sorting out Lipschitz function approximation*, vol. 97 of Proceedings of Machine Learning Research, Long Beach, California, USA, 09–15 Jun 2019, PMLR, 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, D. Precup and Y. W. Teh, eds., vol. 70 of Proceedings of Machine Learning Research, International Convention Centre, Sydney, Australia, 06–11 Aug 2017, PMLR, 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*, 2020, https://arxiv.org/abs/2001.02110.

[5] R. ATAR, K. CHOWDHARY, AND P. DUPUIS, *Robust bounds on risk-sensitive functionals via Rényi divergence*, SIAM/ASA Journal on Uncertainty Quantification, 3 (2015), pp. 18–33.

[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, A. Gretton and C. C. Robert, eds., vol. 51 of Proceedings of Machine Learning Research, Cadiz, Spain, 09–11 May 2016, PMLR, 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, J. Dy and A. Krause, eds., vol. 80 of Proceedings of Machine Learning Research, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018, PMLR, 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), p. 2239–2265.

[9] J. BIRRELL, M. A. KATSOULAKIS, AND Y. PANTAZIS, *Optimizing variational representations of divergences and accelerating their statistical estimation*, arXiv e-prints, (2020), arXiv:2006.08781, https://arxiv.org/abs/2006.08781.

[10] J. BUCKLEW, *Introduction to Rare Event Simulation*, Springer Series in Statistics, Springer New York, 2013.

[11] A. BUDHIRAJA AND P. DUPUIS, *Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods*, Probability Theory and Stochastic Modelling, Springer US, 2019.

[12] A. BUTTE AND K. IS., *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements*, Pac Symp Biocomput., (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*, Communications on Pure and Applied Mathematics, 36 (1983), pp. 183–212, https://doi.org/10.1002/cpa.3160360204, https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.3160360204, https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.3160360204.

[15] R. DUDLEY, *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, Cambridge University Press, 2014.

[16] P. DUPUIS AND R. ELLIS., *A Weak Convergence Approach to the Theory of Large Deviations*, Wiley series in probability and statistics, John Wiley & Sons, New York, 1997, http://opac.inria.fr/record=b1092351. A Wiley-Interscience Publication.

[17] P. DUPUIS, M. A. KATSOULAKIS, Y. PANTAZIS, AND L. REY-BELLET, *Sensitivity analysis for rare events based on Rényi divergence*, The Annals of Applied Probability, 30 (2020), pp. 1507 – 1533, https://doi.org/10.1214/19-AAP1468, https://doi.org/10.1214/19-AAP1468.

[18] G. FOLLAND, *Real Analysis: Modern Techniques and Their Applications*, Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts, Wiley, 2013.

[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, G. Lebanon and S. V. N. Vishwanathan, eds., vol. 38 of Proceedings of Machine Learning Research, San Diego, California, USA, 09–12 May 2015, PMLR, pp. 277–286, http://proceedings.mlr.press/v38/gao15.html.

[20] S. GEER, S. VAN DE GEER, R. GILL, B. RIPLEY, S. ROSS, B. SILVERMAN, D. WILLIAMS, AND M. STEIN, *Empirical Processes in M-Estimation*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2000.

[21] S. GHADIMI AND G. LAN, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization, 23 (2013), pp. 2341–2368, https://doi.org/10.1137/120880811, https://doi.org/10.1137/120880811, https://arxiv.org/abs/https://doi.org/10.1137/120880811.

[22] S. GHADIMI AND G. LAN, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), p. 59–99.

[23] M. GIL, F. ALAJAJI, AND T. LINDER, *Rényi divergence measures for commonly used univariate continuous distributions*, Information Sciences, 249 (2013), pp. 124 – 131, https://doi.org/10.1016/j.ins.2013.06.018.

[24] 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, NIPS'14, Cambridge, MA, USA, 2014, MIT Press, p. 2672–2680.

[25] 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, Red Hook, NY, USA, 2017, Curran Associates Inc., p. 5769–5779.

[26] A. HYVÄRINEN, J. KARHUNEN, AND E. OJA, *Independent Component Analysis*, Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control, Wiley, 2004.

[27] 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, Inc., 2015, pp. 397–405.

[28] 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., vol. 125 of Proceedings of Machine Learning Research, PMLR, 09–12 Jul 2020, pp. 2306–2327, http://proceedings.mlr.press/v125/kidger20a.html.

[29] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, arXiv:1412.6980, (2014), https://arxiv.org/abs/1412.6980.

[30] J. B. KINNEY AND G. S. ATWAL, *Equitability, mutual information, and the maximal information coefficient*, Proceedings of the National Academy of Sciences, 111 (2014), pp. 3354–3359, https://doi.org/10.1073/pnas.1309933111, https://www.pnas.org/content/111/9/3354, https://arxiv.org/abs/https://www.pnas.org/content/111/9/3354.full.pdf.

[31] N. KWAK AND CHONG-HO CHOI, *Input feature selection by mutual information based on Parzen window*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002), pp. 1667–1671.

[32] T. LELIÉVRE, G. STOLTZ, AND M. ROUSSET, *Free Energy Computations: A Mathematical Perspective*, Imperial College Press, 2010, https://books.google.com/books?id=SqJGgfPq_ZUC.

[33] Y. LI AND R. E. TURNER, *Rényi divergence variational inference*, in Advances in Neural Information Processing Systems, 2016, pp. 1073–1081.

[34] F. LIESE AND I. VAJDA, *Convex Statistical Distances*, Teubner-Texte zur Mathematik, Teubner, 1987, https://books.google.com/books?id=PBOoAAAAIAAJ.

[35] F. LIESE AND I. VAJDA, *On divergences and informations in statistics and information theory*, IEEE Transactions on Information Theory, 52 (2006), pp. 4394–4412, https://doi.org/10.1109/TIT.2006.881731.

[36] 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, Red Hook, NY, USA, 2017, Curran Associates Inc., p. 6232–6240.

[37] 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.

[38] X. NGUYEN, M. J. WAINWRIGHT, AND M. I. JORDAN, *Estimating divergence functionals and the likelihood ratio by convex risk minimization*, IEEE Transactions on Information Theory, 56 (2010), pp. 5847–5861.

[39] 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, Red Hook, NY, USA, 2016, Curran Associates Inc., p. 271–279.

[40] L. PANINSKI, *Estimation of entropy and mutual information*, Neural Computation, 15 (2003), pp. 1191–1253, https://doi.org/10.1162/089976603321780272, https://doi.org/10.1162/089976603321780272, https://arxiv.org/abs/https://doi.org/10.1162/089976603321780272.

[41] Y. PANTAZIS, D. PAUL, M. FASOULAKIS, Y. STYLIANOU, AND M. KATSOULAKIS, *Cumulant GAN*, arXiv e-prints, (2020), arXiv:2006.06625, https://arxiv.org/abs/2006.06625.

[42] S. PARK, C. YUN, J. LEE, AND J. SHIN, *Minimum width for universal approximation*, in International Conference on Learning Representations, 2021, https://openreview.net/forum?id=O-XJwyoIF-k.

[43] A. PINKUS, *Approximation theory of the MLP model in neural networks*, Acta Numerica, 8 (1999), p. 143–195, https://doi.org/10.1017/S0962492900002919.

[44] 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, UAI'11, Arlington, Virginia, United States, 2011, AUAI Press, pp. 599–608, http://dl.acm.org/citation.cfm?id=3020548.3020618.

[45] S. RAHMAN, *The f-sensitivity index*, SIAM/ASA Journal on Uncertainty Quantification, 4 (2016), pp. 130–162.

[46] S. J. REDDI, S. KALE, AND S. KUMAR, *On the Convergence of Adam and Beyond*, arXiv e-prints, (2019), arXiv:1904.09237, https://arxiv.org/abs/1904.09237.

[47] A. RÉNYI, *On measures of entropy and information*, tech. report, HUNGARIAN ACADEMY OF SCIENCES Budapest Hungary, 1961.

[48] G. RUBINO, B. TUFFIN, ET AL., *Rare event simulation using Monte Carlo methods*, vol. 73, Wiley Online Library, 2009.

[49] 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, ICML'12, Madison, WI, USA, 2012, Omnipress, p. 1155–1162.

[50] F. SANTAMBROGIO, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Progress in Nonlinear Differential Equations and Their Applications, Springer International Publishing, 2015.

[51] S. SHALEV-SHWARTZ AND S. BEN-DAVID, *Understanding Machine Learning: From Theory to Algorithms*, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014, https://books.google.com/books?id=ttJkAwAAQBAJ.

[52] J. SONG AND S. ERMON, *Understanding the limitations of variational mutual information estimators*, 2020, https://arxiv.org/abs/1910.06222.

[53] N. TISHBY, F. C. PEREIRA, AND W. BIALEK, *The information bottleneck method*, arXiv e-prints, (2000), physics/0004057, https://arxiv.org/abs/physics/0004057.

[54] T. VAN ERVEN AND P. HARREMOS, *Rényi divergence and Kullback-Leibler divergence*, IEEE Transactions on Information Theory, 60 (2014), pp. 3797–3820.

[55] R. VERSHYNIN, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018, https://doi.org/10.1017/9781108231596.

[56] 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, 2018, p. 2955–2961.