Available online at www.eccomasproceedia.org
Eccomas Proceedia UNCECOMP (2021) 112-128

ECCOMAS
Proceedia

UNCECOMP 2021
4th ECCOMAS Thematic Conference on
Uncertainty Quantification in Computational Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Streamed from Athens, Greece, 28 -30 June 2021

EFFICIENT DISCRIMINATION BETWEEN BIOLOGICAL POPULATIONS VIA NEURAL-BASED ESTIMATION OF RÉNYI DIVERGENCE[1]

Anastasios Tsourtis[2], Georgios Papoutsoglou[2] and Yannis Pantazis[2]

[2]Institute of Applied and Computational Mathematics,
Foundation for Research and Technology - Hellas, Greece
e-mail: {tsourtis, papoutsoglou, pantazis}@iacm.forth.gr

[1]This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020" in the context of the project "Characterizing Population Dynamics with Applications in Biological Data" (MIS 5050686).

Abstract. *The advent of single-cell or single-molecule sampling techniques allowed the study of abnormalities in small subsets of cell populations as well as subtle differences of evolving biological phenomena. A major challenge in this effort is the quantification of statistical differences between the probability distributions of measured quantities; particularly when small perturbations or small distributional changes need to be detected. Here, we propose to use as a discriminative tool the Rényi divergence whose key advantage is its ability to highlight differences between probability tails. In addition, we describe an algorithm which is based on a variational representation formula for the Rényi divergence and implicitly estimates it by solving an optimization problem. We evaluate the discrimination performance on both synthetic and real datasets. The proposed algorithm is able to detect distributional differences which are below 0.5% and quantify the trade-off between number of samples and neural network complexity. The comparison with existing density ratio approaches reveals that the proposed method is significantly better when the dimension of the data is moderately high (e.g., larger than 10).*

Keywords: Statistical populations, Rényi divergence, Variational representation, Neural networks

ISSN:2623-3339
Peer-review under responsibility of the organizing committee of UNCECOMP 2021.
doi: 10.7712/120221.8026.19067

1 INTRODUCTION

Population datasets emerge in many scientific fields such as biology [1, 2], ecology [3], epidemiology [4] and molecular motion in biochemistry [5, 6] to name a few. Particularly in biology typical population datasets now consist of tens of thousands of samples measuring dozens of quantities of interest. For instance, high dimensional single-cell technologies, such as flow and mass cytometry [7], are able to capture the abundance of up to 40 proteins on thousands of cells simultaneously. Such moderate to high dimensional datasets cannot be screened out manually (e.g. through scatter plots) and computational approaches are required for the detection of clusters and differences in the data. Despite the recent proposal of computational tools that handle single-cell population data [2, 8], a major challenge in the characterization of cellular heterogeneity still remains. Indeed, it is often the case that rare sub-populations exist in the samples which are very difficult to be detected due to their low abundance levels. For instance, stem and progenitor cells are underrepresented in the total cell population therefore; they are rarely detected using general-purpose methodologies over large populations of cells.

Statistical quantities such as the mean value and the covariance matrix are not sufficient discriminative metrics because they do not capture the complete probabilistic characteristics of the two populations. On the other hand, a probability distance or a divergence could capture all the statistical information induced by the observed sample distributions. Additionally, probability distances which are sensitive to small perturbations and be able to detect small distributional changes are ideal choices. In this paper, we suggest using Rényi divergence as an approach to discriminate between two population datasets. Rényi divergence has the advantage that its hyper-parameter controls how much weight to put on the tails of the distributions thus it can become very sensitive to rare sub-populations inside the population datasets (see Figure 2 for three examples).

The estimation of the Rényi divergence becomes feasible with the use of a variational representation [9, 10, 11]. Variational representations essentially transform the estimation of a divergence to an optimization problem over an infinite-dimensional function space. Then, the function space is approximated by a neural network parametrized space in a similar fashion to [12, 13, 14, 15]. Thus, we present and then evaluate an algorithm that estimates the Rényi divergence which we named NERD (Neural-based Estimation of Rényi Divergence) algorithm.The utilization of neural networks offers additional advantages such as the ability to handle high dimensional data as well as any type of input data with the trade-off being the requirement for a large sample size; a limitation which is already alleviated in practice by the production of large amounts of single-cell measurements per experiment.

We first evaluate NERD algorithm on synthetic data where the ground truth is known. NERD is capable of handling high-dimensional data better than state-of-the-art methods such as ITE [16]. We assess the behavior of the estimator as a function of various hyperparameters such as the number of samples, the rarity of the sub-population and the choice of function space. We numerically show that NERD algorithm is capable of accurately estimating the Rényi divergence in high dimensions given enough sample size. We also compute the discriminative capabilities of NERD between single-cell populations. The two populations consist of cells from healthy participants as well as healthy cells contaminated by a small portion of "sick" cells. We show that NERD algorithm can discriminate confidently when the percentage of rare subpopulation is above 0.2% and the number of available samples is above 40K.

2 DEFINITION AND PROPERTIES OF THE RÉNYI DIVERGENCE

Let Q and P be two probability measures (or distributions) on a measurable space $(\Omega, \mathcal{M})$. The Rényi divergence of order $\alpha > 0$ with $\alpha \neq 1$ of Q with respect to P is defined as [17, 18]

$$\mathcal{R}_\alpha\left(Q||P\right) := \frac{1}{\alpha(\alpha-1)} \log \mathbb{E}_P\left[\left(\frac{dQ}{dP}\right)^\alpha\right] \tag{1}$$

when Q and P are mutually absolutely continuous[1] with respect to each other, otherwise, $\mathcal{R}_\alpha\left(Q||P\right) = \infty$. The 'ratio' $\frac{dQ}{dP}$ is the Radon-Nikodym derivative of Q with respect to P which always exists due to the imposed absolute continuity condition. The defining properties of a divergence are that (a) it is non-negative and (b) it equals to zero if and only if $Q = P$. Despite not being a distance since it is neither symmetric nor satisfies the triangular inequality, divergences are widely used for the comparison of probability distributions.

In some studies, the definition of Rényi divergence utilizes the factor $\frac{1}{\alpha-1}$ (cf. [19, 20, 9]) instead of $\frac{1}{\alpha(\alpha-1)}$, nevertheless, we prefer the definition (1) due to the symmetry property

$$\mathcal{R}_\alpha\left(Q \,||\, P\right) = \mathcal{R}_{1-\alpha}\left(P \,||\, Q\right)$$

when $0 < \alpha < 1$. Using this symmetry property, the definition of Rényi divergence is straightforwardly extended to $\alpha < 0$ (e.g., $\mathcal{R}_{-1}\left(Q \,||\, P\right) := \mathcal{R}_2\left(P \,||\, Q\right)$).

The definition of Rényi divergence is extended to $\alpha = 1$ where the limit equals to the Kullback-Leibler divergence defined by

$$D_{KL}\left(Q||P\right) := \int \log \frac{dQ}{dP} dQ \tag{2}$$

when $Q \ll P$, otherwise $D_{KL}\left(Q||P\right) = +\infty$ as well as to $\alpha = 0$ where the limit equals to the reverse[2] Kullback-Leibler divergence. Interestingly, several other divergences are linked to Rényi divergence. Rényi divergence has an one-to-one and onto correspondence with α-divergence [21] where Rényi divergence can be obtained as an affine transformation of the logarithm of the α-divergence. Rényi divergence is also related to Hellinger distance [22] as well as to χ^2-divergence [22] for particular values of α. Figure 1 summarizes those relationships. Further properties of the Rényi divergence can be found for instance in [23, 20, 9].

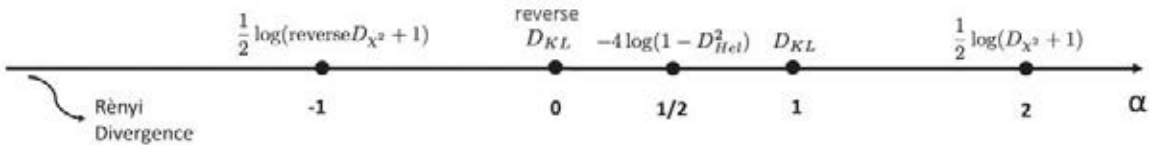

Figure 1: Rényi divergence as a function of its order α and its connections to other divergences. The case $\alpha = 0.5$ relates with the Hellinger distance while the cases $\alpha = 1$ & 2 relate with Kullback-Leibler and (Pearson's) χ^2 divergence, respectively. Rényi divergence is reverse symmetric around 0.5 thus the cases $\alpha = 0$ & -1 relate with reverse Kullback-Leibler and reverse (i.e., Neyman's) χ^2 divergence, respectively.

[1] We say that Q is absolutely continuous with respect to P if for every measurable set $A \in \Omega$, $P(A) = 0 \Rightarrow Q(A) = 0$. It is written as $Q \ll P$.

[2] In the sense that the order of Q and P has been reversed.

2.1 Rényi Divergence Highlights Distributions' Tail Differences

Existing literature has shown that Rényi divergence is capable of efficiently bounding the probability of rare events and more generally of risk-sensitive observables of a distribution [18, 24, 25] through its order parameter. Intuitively, the order parameter as a power factor of the density ratio leverages the amount of weight put on the tails of the distributions. For instance, in [24], to discriminate between rare events from distributions with infinitesimal small differences the order had to be sent to infinity.

We demonstrate this sensitivity property of the Rényi divergence through a series of examples. First, we consider two zero-centered univariate Gaussian distributions with different standard deviations[3], $Q \equiv \mathcal{N}(0, \sigma_1^2)$ and $P \equiv \mathcal{N}(0, \sigma_0^2)$. The Rényi divergence of Q with respect to P is given by [26]

$$\mathcal{R}_\alpha\left(Q||P\right) = \begin{cases} \frac{1}{\alpha}\log\frac{\sigma_0}{\sigma_1} + \frac{1}{2\alpha(\alpha-1)}\log\frac{\sigma_0^2}{\alpha\sigma_0^2+(1-\alpha)\sigma_1^2} & \text{if } \alpha\sigma_0^2 + (1-\alpha)\sigma_1^2 > 0 \\ +\infty & \text{otherwise} \end{cases} . \tag{3}$$

As α approaches to the 'finiteness' limit $\frac{\sigma_1^2}{\sigma_1^2-\sigma_0^2}$, the Rényi divergence takes exponentially-large values resulting in an unequivocal discrimination between the two distributions. Going one step further, if $\sigma_1^2 = \sigma_0^2(1+\epsilon)$ with ϵ being a small number then α should be of order $O(\epsilon^{-1})$ in order to efficiently discriminate between the two distributions. Figure 2(a) demonstrates this behavior for two values of ϵ. Analogous discriminative capacity is observed when the Rényi divergence between a Gaussian distribution with full covariance matrix and a Gaussian with diagonal covariance structure is calculated. In this second example, let Q be a zero-mean Gaussian with covariance matrix $\Sigma_1 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and P a zero-mean Gaussian with covariance matrix equal to the identity matrix (i.e., independent components). Figure 2(b) presents the Rényi divergence as a function of its order. As it is evident from the plot, there are both positive and negative α's of order $O(\rho^{-1})$ that assign very large values to the Rényi divergence implying that even very small correlations between variables could be detected when $|\alpha|$ becomes sufficiently large. Additionally, both Gaussian examples show that large values for the order may result to infinite Rényi divergence and there is a finiteness limit for α that should not be exceeded. Therefore, caution must be placed on the choice of the order value. As a rule of thumb, the "closer" the two distributions are the larger the value of α can (or must) be set.

In this paper, we suggest exploiting this sensitivity property and distinguish between statistical populations of data that differ slightly by containing samples from rare sub-populations. Rare sub-populations are hard to detect exactly because of their rarity. Therefore, we aim to search for the highest value for Rényi divergence by tuning α. As a third and more relevant motivation example, Figures 2(c) & (d) present the Rényi divergence between a mixture of two Gaussians ($Q \equiv (1-w)\mathcal{N}(\mu_0, \sigma_0^2) + w\mathcal{N}(\mu_1, \sigma_1^2)$) with w corresponding to the percentage of the less probable population and a Gaussian ($P \equiv \mathcal{N}(\mu_0, \sigma_0^2)$). Under this particular setting, there is an optimal α that maximizes the Rényi divergence. Additionally, the smaller the percentage of the less probable population the larger the value of the optimal α is. This is consistent with the two previous examples in the sense that distributions with smaller differences require larger values of α in order to obtain larger Rényi divergence values.

[3] In this paper, we always consider P to be the baseline (or unperturbed) distribution while Q is the different or perturbed one.

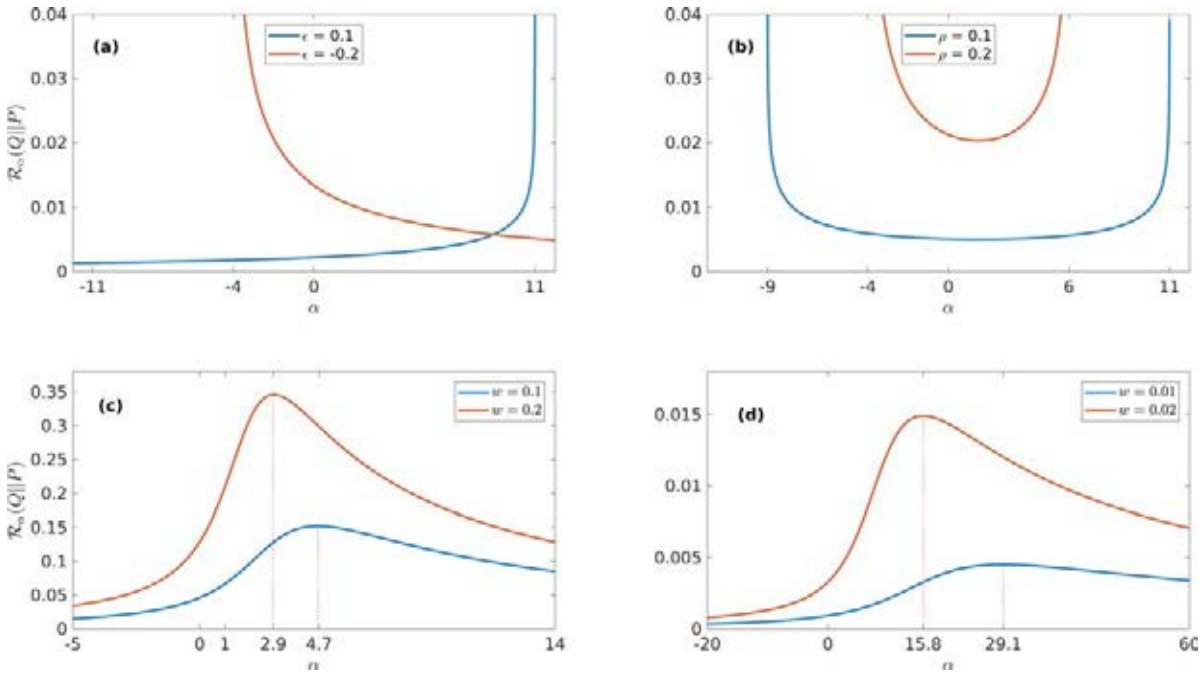

Figure 2: Rényi divergence of Q (perturbed) with respect to P (unperturbed) as a function of α between: **(a)** two 1D zero-mean Gaussian distributions with different variances ($\sigma_0^2 = 1, \sigma_1^2 = 1 + \epsilon$), **(b)** two 2D zero-mean Gaussian distributions with different covariance structure (ρ: correlation coefficient between the two elements of the perturbed Gaussian) and **(c)-(d)** between a mixture of two Gaussians and a Gaussian distribution (w: the percentage of the second mode in the mixture).

2.2 A Variational Representation Formula for the Rényi Divergence

By definition, the estimation of Rényi divergence requires either the knowledge of the densities of the probabilities involved or an approximation of their ratio. An alternative approach is to transform the estimation problem into an optimization problem via the utilization of a variational representation. A variational representation formula is essentially a lower bound of the divergence for which the optimal solution gives rise to the divergence value. It consists of two mathematical ingredients: the function space where the optimal solution will be searched for and, the representation expression, called here the 'objective functional', whose optimization leads to the value of the divergence.

The following theorem, proved in [11], states that Rényi divergence is the solution of a variational optimization with an objective functional which is the difference of two risk-sensitive observables (i.e., the expression inside the curly brackets in (4)).

Theorem 1. *Let P, Q be two probability measures on $(\Omega, \mathcal{M})$ and $\alpha \in \mathbb{R} \setminus \{0, 1\}$. Then,*

$$\mathcal{R}_\alpha(Q||P) = \sup_{g \in \mathcal{M}_b(\Omega)} \left\{ \frac{1}{\alpha - 1} \log \mathbb{E}_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha} \log \mathbb{E}_P[e^{\alpha g}] \right\}, \tag{4}$$

where $\mathcal{M}_b(\Omega)$ is the space of all (real-valued) measurable and bounded functions from Ω to $\mathbb{R}$ and we assume the conventions $+\infty - \infty = -\infty$ and $-\infty + \infty = -\infty$.

The optimal solution under appropriate conditions provided in [11] can be explicitly written as $g^* = \log \frac{dQ}{dP}$, and, the aim of this paper is to approximate g^* as accurately as possible. Finally, taking $\alpha \to 1$, we recover the Donsker-Varadhan variational formula for the Kullback-Leibler

divergence [27] which is given by

$$D_{KL}\left(Q||P\right) = \sup_{g \in \mathcal{M}_b(\Omega)} \left\{\mathbb{E}_Q[g] - \log \mathbb{E}_P[e^g]\right\} \ . \tag{5}$$

Hence, equation (4) can be seen as a generalization of the Donsker-Varadhan formula to the Rényi divergence. Similarly, the Donsker-Varadhan formula with the order of Q and P reversed is obtained when $\alpha \to 0$.

3 NEURAL-BASED ESTIMATION OF RÉNYI DIVERGENCE (NERD)

The variational formula in (4) is still not fully practical because (a) the expectations cannot be explicitly computed since Q and P are not known, and (b) the infinite dimensional space of test functions (i.e., of g's) needs to be restricted to a parametric representation that can be handled by a computer. Regarding the first issue, the expectations are replaced by their statistical averages using a finite number of samples. This approximation fits well with our setting since we only have access to samples from the distributions of interest. Moreover, as the number of samples tends to infinity, the statistical averages converge to the respective expectation values.

For the latter issue, we concentrate to $\Omega = \mathbb{R}^d$ and parametrize the space of all measurable and bounded functions with neural networks of bounded activation function for the output layer. Letting $\theta \in \mathbb{R}^p$ be the parameter vector with the weights and biases of the neural network, our aim is to optimize $g_\theta : \mathbb{R}^d \to [-M, M]$ where M is a user-defined clipping factor. In our experiments we enforce the boundedness condition via the use of $M \tanh\left(\frac{\cdot}{M}\right)$ as the activation function of the output layer. The error induced by this second approximation can be controlled using (a) Lusin's theorem [28] where the space of all measurable and bounded functions is replaced with all continuous and bounded functions with arbitrary accuracy and (b) the fact that a large enough neural network is a universal approximator of continuous and bounded functions [29, 30].

The parameters of the neural network are estimated using *stochastic gradient ascent* because we are searching for the solution that maximizes the objective functional. The pseudo-code of the neural-based Rényi divergence estimator is provided in Algorithm 1. When $\alpha = 1$ or 0, we apply the finite sampling approximation formulas stemming from the Donsker-Varadhan variational representation (5).

3.1 Statistical Properties of NERD

The asymptotic consistency of NERD has been shown in [11, Theorem 2]. However, stability and consistency results as well as bias-variance trade-offs for finite number of samples is an open and active problem even for the Kullback-Leibler case [31, 32, 33]. The main encumbrance stems from the fact that both terms in the objective function of Rényi's variational representation are sensitive to tail events and the variance of the estimator could grow exponentially with the true value of the divergence [33]. A partial solution proposed in [33] sets a small clipping factor M applied to the output of the final layer which results in reduced variance for the estimator at the cost of larger bias especially when the value of Rényi divergence is high since the maximum possible value for the estimator is $2M$. As it is already presented, NERD algorithm has adopted the clipping operator. In the following section, we propose a different approach to reduce the variance by utilizing a different, more regularized function space for the optimization problem.

Algorithm 1 Neural Estimation of Rényi Divergence **(NERD)**

Input: Sample matrix $X \in \mathbb{R}^{N\times d} \sim Q$, sample matrix $Y \in \mathbb{R}^{N\times d} \sim P$, order parameter α, neural network $g_\theta(\cdot)$, batch size m and learning rate λ_{lr}

Output: Rényi divergence estimate: $\hat{R}_\alpha^N$

1: $\theta \leftarrow$ Initialize_Neural_Network()

2: **while** not converged **do**

3: Choose randomly m samples from X: $\{x_i\}_{i=1}^m$ and from Y: $\{y_i\}_{i=1}^m$

4: Compute the variational expression:

$$R(\theta) = \frac{1}{\alpha-1}\log\frac{1}{m}\sum_{i=1}^{m} e^{(\alpha-1)g_\theta(x_i)} - \frac{1}{\alpha}\log\frac{1}{m}\sum_{i=1}^{m} e^{\alpha g_\theta(y_i)} \tag{6}$$

5: Update the neural net's parameters:

$$\theta \leftarrow \theta + \lambda_{lr}\nabla_\theta R(\theta)$$

6: **end while**

7: Compute the variational estimate using all samples:

$$\hat{\mathcal{R}}_\alpha^N = \frac{1}{\alpha-1}\log\frac{1}{N}\sum_{i=1}^{N} e^{(\alpha-1)g_\theta(x_i)} - \frac{1}{\alpha}\log\frac{1}{N}\sum_{i=1}^{N} e^{\alpha g_\theta(y_i)} \tag{7}$$

3.2 Using Lipschitz Continuous Functions as Test Functions

The space of test functions can be selected differently. The cost of choosing a subset of $\mathcal{M}_b(\Omega)$ is that a lower bound –but not necessarily strictly lower– for the divergence is obtained. Given that we are interested in alleviating the impact of finite sampling on the approximated risk-sensitive observables, we propose to use Lipschitz continuous functions with Lipschitz constant K as the function space over which the optimal solution will be sought for. The 1-Wasserstein distance, which is also defined on the Lipschitz function space but uses a different objective functional, has shown significantly better stability and convergence properties during the training of GANs [14, 15]. Thus, we anticipate improved statistical properties such as reduced variance in our experiments. Theoretically, it has also been shown that the function space replacement from measurable to Lipschitz functions retains the divergence property for the α-divergence [34] hence it is also retained for the Rényi divergence.

From an implementation perspective, the only difference for NERD algorithm is the removal of the clipping function and the addition of a gradient penalty term in (6). The new formula is given by

$$\begin{aligned} R(\theta) = &\frac{1}{\alpha-1}\log\frac{1}{m}\sum_{i=1}^{m} e^{(\alpha-1)g_\theta(x_i)} - \frac{1}{\alpha}\log\frac{1}{m}\sum_{i=1}^{m} e^{\alpha g_\theta(y_i)} \\ &+ \lambda_{GP}\frac{1}{m}\sum_{i=1}^{m}\max\left(0, ||\nabla_x g_\theta(z_i)||^2 - K\right) , \end{aligned} \tag{8}$$

where $z_i = u_i x_i + (1-u_i)y_i$ and $u_i \sim \mathcal{U}(0,1)$ for $i = 1, ..., m$. We remark that this is the one-sided gradient penalty and it is only activated when the square of the gradient's norm is above K. The two-sided gradient penalty which is valid for the Wasserstein distance is not applicable

for the Rényi divergence since the norm of the gradient is not everywhere equal to one for the optimal test function.

4 RESULTS

In this Section, we test the accuracy of the proposed algorithm on two synthetic examples as well as its discriminative efficacy on one real biological dataset. Our aim is to numerically evaluate the performance of NERD algorithm on the statistical estimation of Rényi divergence and also explore the Rényi's order parameter that leads to the most efficient discrimination between two sample distributions with small sub-population differences. Our results are compared against a state-of-the-art density ratio approximation algorithm implemented by the Information Theoretical Estimators (ITE) toolbox [16].

4.1 Experimental setup

In Section 3, we presented two variants of the NERD algorithm depending on the chosen space of test functions: i) the space of continuous and bounded test functions referred to as NERD$_{C_b}$, and ii) the space of Lipschitz continuous test functions referred to as NERD$_{Lip}$. The boundedness condition of NERD$_{C_b}$ is enforced through a bounding factor $M > 0$ on the activation function of the final layer. In contrast, the Lipschitz continuous condition is enforced through the addition of a regularization term that depends on two hyper-parameters: the Lipschitz constant $K > 0$ and the regularization coefficient λ_{GP}. Both M and K are capable of affecting the trade-off between estimation bias and estimation variance with smaller values favoring reduced variance with the cost of increased bias. Table 1 summarizes the value ranges of all (hyper-)parameters that appear in our numerical experiments.

Neural network hyperparameters were set following a similar rationale as in [12] where the Kullback-Leibler case was studied and the implementation is carried out using TensorFlow2[4]. Specifically, we employ fully-connected feed-forward neural networks with $l = 3$ layers, variable number of units per layer and $\tanh(\cdot)$ as activation function for the hidden layers. The number of units per layer is primarily dependent on the dimension of the data while the number of trainable parameters is typically of order $\mathcal{O}(10^3)$. Given that the sample size of the the studied datasets is between $\mathcal{O}(10^4) - \mathcal{O}(10^5)$ we anticipate no overfitting. For the sake of fairness, both NERD variants share the same architecture (i.e., hidden layers, number of units per layer, activation function) except the activation function of the output layer which is different.

We apply Adam optimizer [35] as the training algorithm with its default hyperparameter values. The learning rate is set to $\lambda_{lr} = 0.0005$ for the synthetic examples while the number of iterations was $N_{it} = 20000$. Due to slower convergence, the respective values for the real dataset are $\lambda_{lr} = 0.01$ and $N_{it} = 60000$. Moreover, we set a large value to the batch size so that samples from the tails are included in the statistical average with high probability at each step.

We also set the hyperparameter of the ITE-based estimator that defines the number, k, of nearest neighbors. Since the computational cost increases non-linearly with k, ITE becomes prohibitive for high dimensions and large sample sizes. We found that setting $k = 20$ is a balanced choice for approximating the Rényi divergence in our experiments.

Finally, our primal goal in the synthetic examples is to assess the accuracy of the estimator, hence, we fix the order of the Rényi divergence to $\alpha = 0.5$. Such order value provides a stable statistical behavior with low variance for the estimator relative to the other values of α. In contrast, for the real dataset example, we provide results for a range of α values excluding

[4]Code will be available upon acceptance.

Table 1: Parameters' symbols, their categorization and range in our experiments.

Parameter	Explanation	Association	Range
N	No. samples	Data set	$\mathcal{O}(10^4) - \mathcal{O}(10^5)$
d	Dimension	Data set	$[1, 50]$
w	Sub-population proportion	Data set	$[0.002, 0.2]$
ρ	Correlation coefficient	Data set	$[0, 0.9]$
α	Order	Rényi divergence	$[0.1, 0.9]$
k	No. nearest neighbors	ITE	20
M	Boundedness const.	NERD (bounded)	$[1, 50]$
K	Lipshcitz const.	NERD (Lipschitz)	$[1, 10]$
λ_{GP}	Gradient penalty	NERD (Lipschitz)	0.1
N_{it}	No. iterations (training steps)	Training alg.	$[10000, 150000]$
m	Batch size	Training alg.	4000
λ_{lr}	Learning rate	Training alg.	$[0.0005, 0.01]$
l	No. hidden layers	Neural network	3
θ	Vector w/ weights & biases	Neural network	—
p	Dimension of θ	Neural network	$\mathcal{O}(10^2) - \mathcal{O}(10^3)$

$\alpha \in \{0, 1\}$ wherein the variance of the estimator might become very large [32, 33].

4.2 Rényi Divergence Estimation on Synthetic Data

4.2.1 Between a Gaussian Mixture Model (GMM) and a Gaussian

In our first example we consider Q to be a 1-D bimodal distribution (mixture of two Gaussians) and P a 1-D Gaussian distribution. The first mode of Q will be referred to as the 'main' mode and the second one as the 'rare' mode, inspired by biological data terminology of main and rare cell sub-populations. The mean and variance of P and of the main mode of Q were set equal to one another. Specifically,

$$\begin{aligned} Q &= (1-w)\mathcal{N}(\mu_0, \sigma_0) + w\mathcal{N}(\mu_1, \sigma_1) \\ P &= \mathcal{N}(\mu_0, \sigma_0) \end{aligned} \tag{9}$$

where the μ's and the σ's denote the means and variances of the distributions and w is the probability of the rare mode. In our simulations, we set $\mu_0 = 0$, $\sigma_0 = 1$ for the main mode and $\mu_1 = 1$, $\sigma_1 = \frac{1}{4}$ for the rare mode. The upper panels of Figure (3) illustrate the convergence of all estimators as the sample size increases. As expected, larger sample sizes reduce the variance of the estimators and $N = 50000$ is sufficient for this example. Additional experimentation not shown here revealed that similar results are obtained for other values of μ_1, σ_1 and α.

The lower panels in Figure 3 depict the effect of the probability of the rare mode w on the estimation of the Rényi divergence. Since the percentage of samples between the main and the rare mode of Q is controlled by w, we consider the range between 0.3% and 10% as being representative of the frequencies found in cases of rare cell populations in disease-like situations. Apparently, as w decreases it becomes harder to differentiate between Q and P. Interestingly, both NERD variants are more accurate both in terms of variance and bias in discriminating between slightly different sample distributions for the same sample size. On the other hand, the ITE estimator has undeniable difficulties with small values for w due to its large variance.

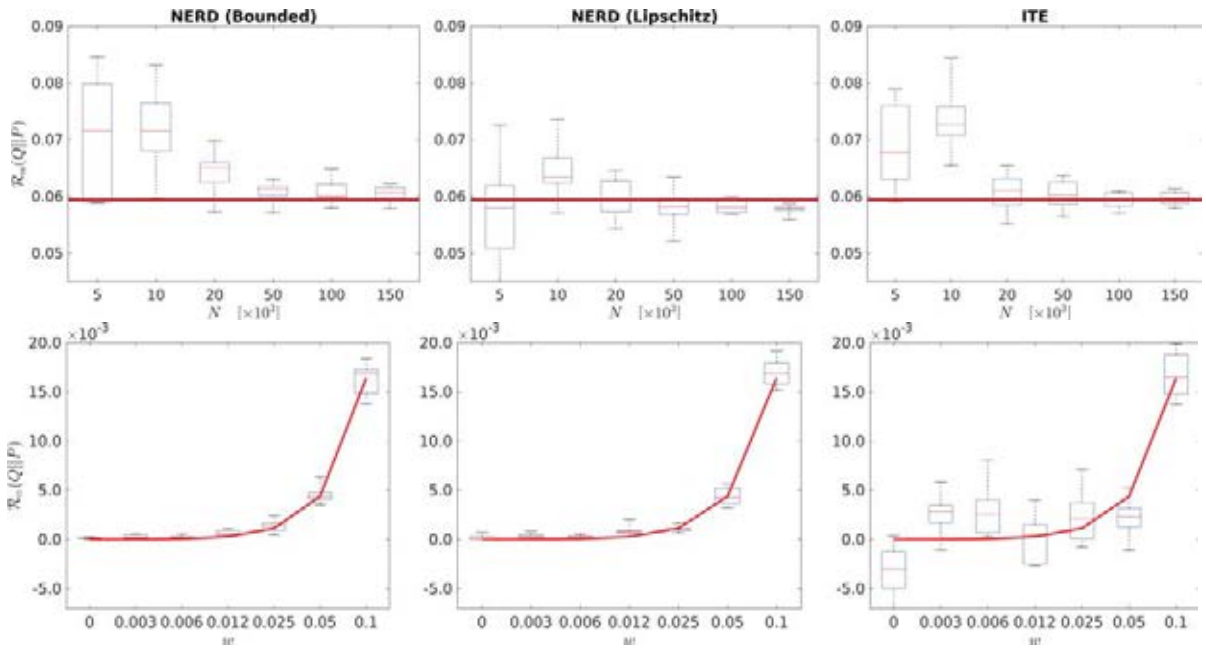

Figure 3: Estimated Rényi divergence between a GMM with two modes and a Gaussian distribution using both variants of NERD algorithm and ITE with the order fixed at $\alpha = 0.5$. *Upper panels:* As the sample size N increases, all methods converge to the exact Rényi divergence value (red solid line) with decreasing variance. Here, we set $w = 0.2$ and quantiles in the box-plots are estimated over 10 independent runs. *Lower panels:* Estimated Rényi divergence as a function of the sub-population proportion w. Here, the sample size is $N = 40000$. Evidently, both variants of NERD exhibit less bias and reduced variance relative to ITE-based estimator for $w < 0.05$.

4.2.2 Between Two High-dimensional Gaussians

In this example, we let Q and P be two zero-mean multivariate (standardized) Gaussian random variables of dimension d with different covariance matrices. We impose the element-wise correlation $corr(x_i, x_{\frac{d}{2}+j}) = \delta_{i,j}\rho$ to the samples $x \sim Q$ where $i, j = 1, \ldots, \frac{d}{2}$ and $\delta_{i,j}$ is Kronecker's delta. In contrast, no correlation is assumed for the samples $y \sim P$. We test how the estimation accuracy of both variants of NERD changes with increasing dimension as well as correlation coefficient. This setting is quite challenging because as dimension increases the probability mass concentrates in a ball around the origin whose radius is exponentially decreasing with respect to the dimension. Moreover, as the correlation coefficient increases the intersection between the supports of the sample distributions is significantly reduced which could result in large estimation errors.

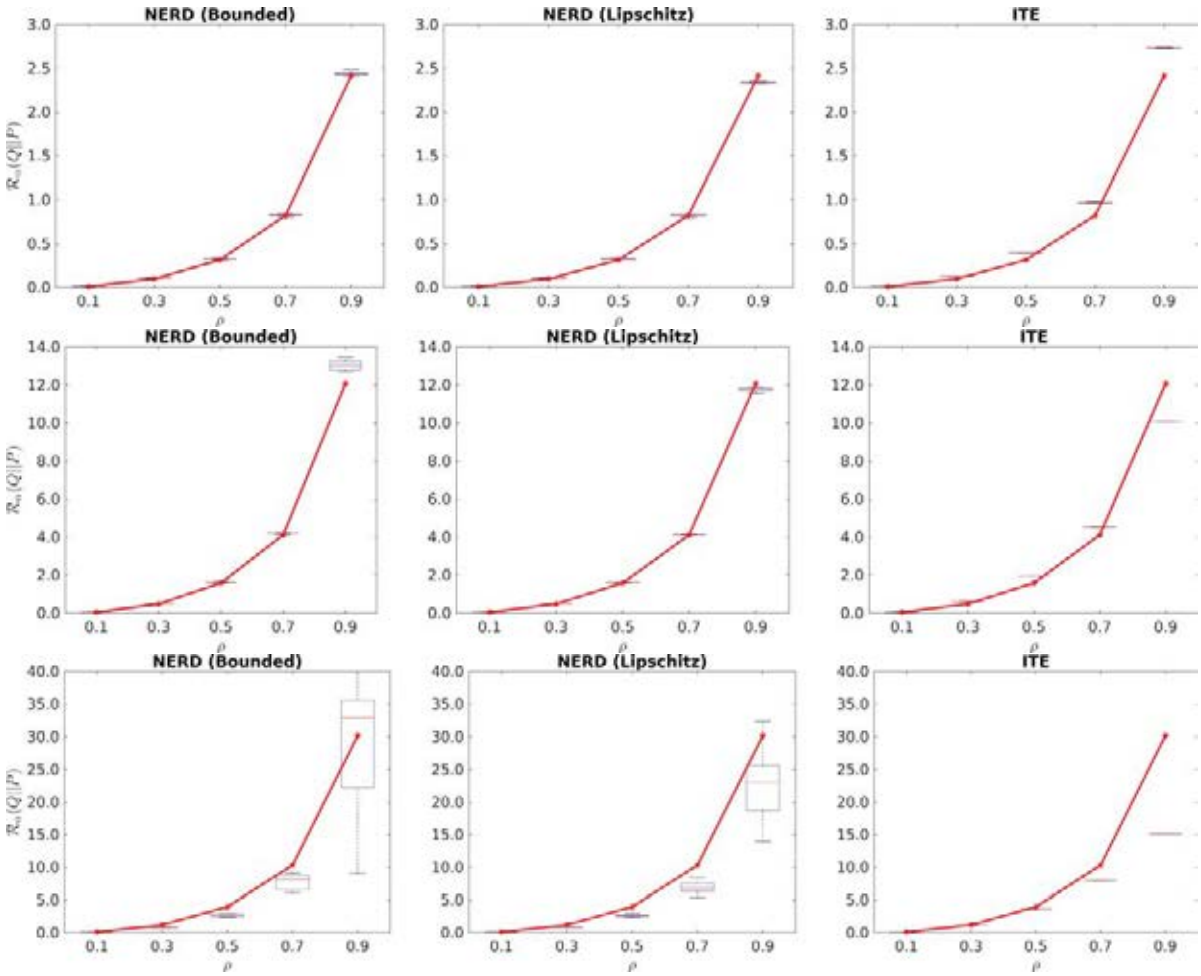

Figure 4: Rényi divergence between two multivariate Gaussian distributions as a function of the correlation coefficient ρ and the dimension d. The red solid line represents the exact Rényi divergence, whereas each boxplot corresponds to the 2nd and 3rd quantiles of the estimator over 10 independent iterations. In all cases, we set $M = 50$ and $K = 5$. *Upper panels:* We compare two $d = 4$-dimensional Gaussians and draw $N = 50000$ samples for each. Both NERD variations are close to the exact Rényi divergence value, whereas ITE slightly overestimates it. *Middle panels:* With dimension being $d = 20$ and sample size being $N = 150000$, NERD with Lipschitz continuous functions provide the most accurate estimates relative to the others. $N_{it} = 20k$]. *Lower panels:* For $d = 50$ and $N = 300000$ samples, the variance for both NERD variations is high and increases as the correlation coefficient increases. ITE-based estimator has no variability but its estimate is entirely inaccurate.

Figure 4 presents the estimation results for the three methods considered in this paper as a function of dimension and correlation coefficient ρ. We consider three values for the dimension: $d = 4$ (upper row of panels), $d = 20$ (middle row of panels), and $d = 50$ (lower row of panels) while we range $\rho \in [0.1, 0.9]$. When $\rho < 0.5$ our results indicate that all three methods provide satisfactory divergence estimations. When ρ increases, both NERD variants accurately estimate the Rényi divergence for $d \leq 20$ with small variance. In contrast, the variance increases significantly as ρ increases (lower panels) revealing that even larger sample size is required. This finding is in partial agreement with the results in [33] where it is shown that variance of this variational-based estimators may grow exponentially with the value of the Rényi divergence. On the other hand, we found that the NERD_{Lip} estimator has less variance relative to NERD_{C_b} especially for $\rho = 0.9$ revealing that the restriction of the function space to Lipschitz continuous functions as presented in Section 3.2 is beneficial from a statistical estimation perspective.

Finally, despite having very low variance in all cases, ITE-based estimation is inaccurate for large values of ρ even for $d = 4$.

Concerning the computational cost in CPU time, the most important factors are the sample size N and the dimension d. The training of neural network's parameters scales linearly with N and in most cases with d too while ITE approach which is based on k-nearest neighbors does not scale efficiently neither with N nor with d. Despite being architecture and dimension dependent, we advocate that the break even point in terms of computational cost between the NERD and ITE approaches is for sample size N approximately between $5 \cdot 10^4$ and 10^5. We also remark that NERD_{Lip} is approximately twice as slow as NERD_{C_b} due to the additional computational cost induced by the regularization term.

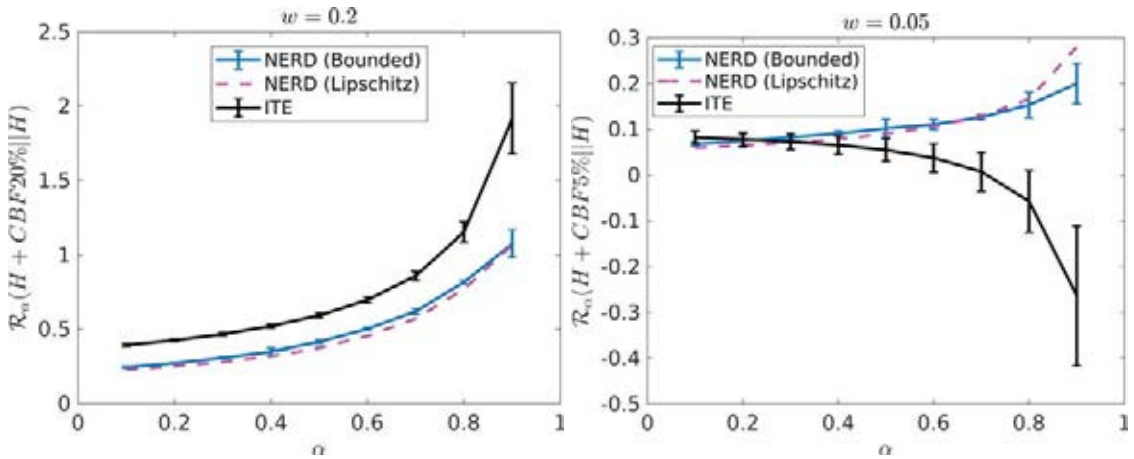

Figure 5: Rényi divergence estimates of NERD_{Lip} ($K = 1$), NERD_{C_b} and ITE when comparing healthy against disease-contaminated distributions, over varying α when the abundance of disease samples is set to $w = 0.2$ (left panel) and $w = 0.05$ (right panel). Errorbars are computed over 5 iterations. Both NERD variants generate similar consistent results while ITE estimates have shift and scaling issues making it untrustworthy.

4.3 Detecting Sub-populations in Single-Cell Datasets

Using data from [36], we test the efficacy of NERD to discriminate distributions from real biological settings[5]. Specifically, we consider single cell mass cytometry measurements on 16 bone marrow protein markers (d=16) coming from healthy and disease individuals with acute myeloid leukemia. The dataset consists of more than 150K healthy and 25K disease cell samples. Before analysis, data were transformed using the inverse hyperbolic sine $arcsinh()$ transformation with a cofactor of 5, which is typical in order to have comparable supports across dimensions. Following [8] we mix healthy and disease samples at decreasing frequencies. For this, we first split the healthy samples randomly into two equally sized subsets X and Y. Then, we replace a predefined percentage of samples in X with disease samples; that is, {20%, 5%, 1%, 0.5% and 0.2%} of cells. The resulting distributions Q_X and P_Y reflect the properties of settings where rare, disease-associated cell populations must be detected from otherwise healthy samples.

Figure 5 shows the Rényi divergence estimates for various $\alpha \in (0, 1)$ and two values for the sick cells percentage. Both NERD variants and ITE estimate positive values for the divergence thus they do discriminate the healthy distribution P_X from Q_Y when $w = 0.2$ (left panel). When

[5]Data were accessed from `https://community.cytobank.org/cytobank/experiments/46098/illustrations/121588`

$w = 0.05$ (right panel), NERD algorithm continues to produce positive values and be able to discriminate between the two distributions, however, ITE estimates are negative for several values of α. It seems that ITE approach has shift and scaling issues. Those estimation errors can be partially reduced by increasing k at the cost of significantly higher computational effort. We additionally remark that the curve of the NERD-estimated Rényi divergence as a function of its order is in accordance with Figure 2(c)-(d) in the sense that as α increases the value of the divergence does also increase. Even though the distributions of the biological data are not normal, this consistency in the behavior of Rényi divergence suggests that NERD algorithm correctly estimates the divergence value.

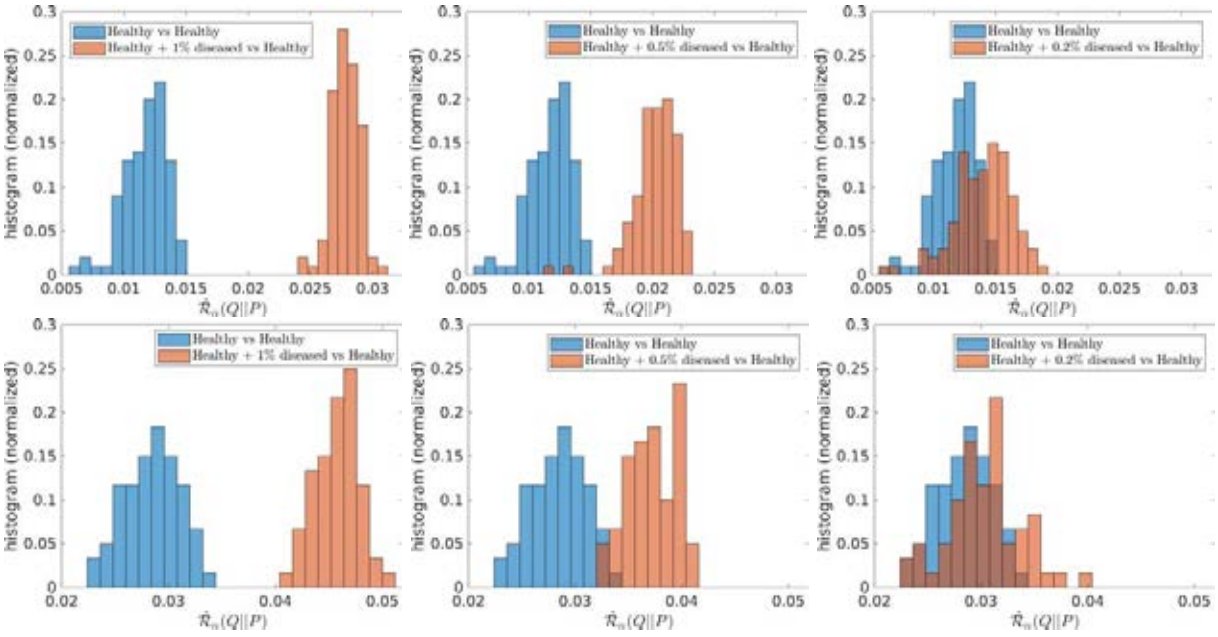

Figure 6: Histograms of repeated NERD estimates (60 runs) when trying to discriminate between two healthy datasets $\hat{\mathcal{R}}_\alpha(P_Y||P_Y)$ (blue color) and between a healthy dataset and a dataset contaminated with sick sub-population with proportion $w\%$: $\hat{\mathcal{R}}_\alpha(Q_X||P_Y)$ (orange color). We consider two sample sizes: $N = 78797$ samples per distribution (upper panels) and $\frac{N}{2}$ samples (lower panels). As the rare sub-population of sick cells decreases in numbers, it is harder to discriminate between the healthy and diseased distributions.

Finally, we investigate the limits of NERD algorithm in discriminating between healthy and sick cell contaminated distributions. Apparently, as w approaches 0, it becomes more difficult to distinguish the two distributions because the number of sick cells amounts to few dozens. Similarly, sample size is an important factor in order to generate statistically significant outcomes. Figure 6 presents histograms of repeated estimates of Rényi divergence with $\alpha = 0.5$ computed via NERD$_{Lip}$ algorithm for the healthy vs healthy case (blue color) and diseased vs healthy case (orange color). In the upper row of panels, we use all available data while the sample size is halved in the lower row of panels. As it is evident from the x-axis, the estimates of Rényi divergence for the healthy vs healthy case is doubled for the lower panels revealing that sample size is indeed a crucial factor for accurately estimating the divergence which is zero for this case. Moreover, histograms are separated for values of w above 0.005 for both sample sizes. On the other hand, there is overlap of the histograms when $w = 0.002$ especially for the halved sample size. This is also evident from the Kolmogorov-Smirnov (KS) test computed on the histograms. Table 2 reports the p-value and the statistic of the KS test using build-in

function `kstest2`. Overall, we conclude that sick cell proportion below $w = 0.002$ cannot be detected with NERD for $\alpha = 0.5$ and sample size below $N = 40000$. Nevertheless, when we increase α to 0.8, differences between the histograms start to emerge (last row in Table 2) showing that larger values for α could assist in discriminating even rarer sub-populations.

Table 2: p-values and statistic for the KS test. Apart from one case, KS test suggests that the two distributions are different.

α	w	p-value	KS stat	samples
$\alpha = 0.5$	1%	1.5×10^{-45}	1	$N \approx 79K$
$\alpha = 0.5$	0.5%	9.4×10^{-44}	0.98	N
$\alpha = 0.5$	0.2%	3.6×10^{-12}	0.51	N
$\alpha = 0.5$	1%	7.8×10^{-28}	1	$N/2$
$\alpha = 0.5$	0.5%	4.9×10^{-26}	0.96	$N/2$
$\alpha = 0.5$	0.2%	0.0068	0.3	$N/2$
$\alpha = 0.8$	0.2%	1.4×10^{-5}	0.433	$N/2$

5 CONCLUSIONS

In this paper, we propose an efficient discrimination approach based on Rényi divergence to quantify the difference between population datasets and answer whether or not two sample population come from the same distribution. The estimation of Rényi divergence is performed via the optimization of functions parametrized by neural networks. We investigated the performance of the presented algorithm (NERD) on several synthetic and real biological datasets. We showed that both NERD variants accurately estimate the Rényi divergence and discriminate rare sub-populations in the data given sufficiently-large number of samples. Therefore, its potential to be used as a screening and/or detection tool in single-cell applications is high. As future work, we target towards devising novel techniques that reduce the estimator's variance and require smaller sample sizes.

REFERENCES

[1] Sean C. Bendall, Erin F. Simonds, Peng Qiu, El-ad D. Amir, Peter O. Krutzik, Rachel Finck, Robert V. Bruggner, Rachel Melamed, Angelica Trejo, Olga I. Ornatsky, Robert S. Balderas, Sylvia K. Plevritis, Karen Sachs, Dana Pe'er, Scott D. Tanner, and Garry P. Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, 332(6030):687–696, 2011.

[2] Smita Krishnaswamy, Matthew H. Spitzer, Michael Mingueneau, Sean C. Bendall, Oren Litvin, Erica Stone, Dana Pe'er, and Garry P. Nolan. Conditional density-based analysis of t cell signaling in single-cell data. *Science*, 346(6213), 2014.

[3] Valery Tereshko. Reaction-diffusion model of a honeybee colony's foraging behaviour. In Marc Schoenauer, Kalyanmoy Deb, Günther Rudolph, Xin Yao, Evelyne Lutton, Juan Julian Merelo, and Hans-Paul Schwefel, editors, *Parallel Problem Solving from Nature PPSN VI*, pages 807–816, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[4] Sebastian Anita and Vincenzo Capasso. Reaction-diffusion systems in epidemiology, 2017.

[5] Eugenio Marco, Robert L. Karp, Guoji Guo, Paul Robson, Adam H. Hart, Lorenzo Trippa, and Guo-Cheng Yuan. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, 2014.

[6] Hong Zhan, Ramunas Stanciauskas, Christian Stigloher, Kevin K Dizon, Maelle Jospin, Jean-Louis Bessereau, and Fabien Pinaud. In vivo single-molecule imaging identifies altered dynamics of calcium channels in dystrophin-mutant c. elegans. *Nature communications*, 5:4974, 2014.

[7] Laura de Vargas Roditi and Manfred Claassen. Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. *Current Opinion in Biotechnology*, 34:9–15, 2015. Systems biology • Nanobiotechnology.

[8] Lukas M Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology*, 2(183), 2019.

[9] Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[10] Venkat Anantharam. A variational characterization of rényi divergences. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 893–897, 2017.

[11] Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational Representations and Neural Network Estimation for Rényi Divergences. *arXiv:2007.03814*, 2020.

[12] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[13] XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[14] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[15] Yannis Pantazis, Dipjyoti Paul, Michail Fasoulakis, Yannis Stylianou, and Markos A. Katsoulakis. Cumulant gan. *arXiv:2006.06625v2*, 2020.

[16] Barnabás Póczos, Zoltán Szabó, and Jeff Schneider. Nonparametric divergence estimators for independent subspace analysis. In *2011 19th European Signal Processing Conference*, pages 1718–1722, 2011.

[17] Friedrich Liese and Igor Vajda. *Convex statistical distances*, volume 95 of *Teubner-Texte zur Mathematik [Teubner Texts in Mathematics]*. BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1987. With German, French and Russian summaries.

[18] Rami Atar, Kamaljit Chowdhary, and Paul Dupuis. Robust bounds on risk-sensitive functionals via rényi divergence. *SIAM/ASA Journal on Uncertainty Quantification*, 3:18–33, 2015.

[19] Alfréd Rényi. On measures of entropy and information. In *Proc. of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume I, pages 547–561, Berkeley, CA, 1961. University of California Press.

[20] Igor Vajda. Distances and discrimination rates for stochastic processes. *Stochastic Processes and Applications*, 35:47–57, 1990.

[21] Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, January 2005.

[22] Alexandre B Tsybakov. Introduction to nonparametric estimation. *Springer Science & Business Media*, 2008.

[23] Leila Golshani, Einollah Pasha, and Gholamhossein Yari. Some properties of Rényi entropy and Rényi entropy rate. *Information Sciences*, 179:2426–2433, 2009.

[24] Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. Sensitivity analysis for rare events based on Rényi divergence. *Ann. Appl. Probab.*, 30(4):1507–1533, 08 2020.

[25] Rami Atar, Amarjit Budhiraja, Paul Dupuis, and Ruoyu Wu. Robust bounds and optimization at the large deviations scale for queueing models via rényi divergence, 2020.

[26] Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249(Complete):124–131, 2013.

[27] Monroe D. Donsker and S. R. Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[28] Gerald B. Folland. *Real analysis: Modern Techniques and Their Applications*. Wiley, New York, 1999.

[29] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359 – 366, 1989.

[30] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei 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, page 6232–6240, Red Hook, NY, USA, 2017. Curran Associates Inc.

[31] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR, 09–15 Jun 2019.

[32] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884. PMLR, 26–28 Aug 2020.

[33] Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020.

[34] Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, γ)-divergences: Interpolating between f-divergences and integral probability metrics, 2021.

[35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *The International Conference on Learning Representations (ICLR)*, 2015.

[36] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El Ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Pe'er, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, 2015.