A VARIANCE REDUCTION METHOD FOR NEURAL-BASED DIVERGENCE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A central problem in machine learning is the computation of similarity or closeness between two (data) distributions. The applications span from generative modelling via adversarial training, representation learning, and robustness in out-of-distribution settings, to name a few. A palette of divergences, mutual information, and integral probability metrics are indispensable tools for measuring the "distance" between distributions and these are made tractable in high dimensional settings through variational representation formulas. Indeed, such formulas transform an estimation problem into an optimization problem. Unfortunately, the approximation of expectations that are inherent in variational formulas by statistical averages can be problematic due to high statistical variance, e.g., exponential for the Kullback-Leibler divergence and certain estimators. In this paper, we propose a new variance penalty term that acts directly on the variance of each component of the statistical estimator. The power of the variance penalty is controlled by a penalty coefficient which trades off bias and variance. We tested the proposed approach on several variational formulas and synthetic examples and showed that the overall error is decreased about an order of magnitude relative to the baseline statistical estimator. Impressive results are obtained for Rényi divergence with large order values due to the improved stability of the proposed estimator. Furthermore, in real biological datasets we are able to detect very rare sub-populations with a moderate sample size. Finally, we obtain improved (in terms of objective measures) disentangled representation of speech signals into text, speaker, and style components via variance-penalized mutual information minimization.

1 INTRODUCTION

Divergences such as Kullback-Leibler (KL) divergence, f-divergences, Hellinger divergence, α-divergences and Rényi divergences, which were initially developed in the fields of information theory and statistical physics, are indispensable tools in a growing number of machine learning applications. They have been used in adversarial training of generative models (Goodfellow et al., 2014; Nowozin et al., 2016), in the estimation of generalization errors (Esposito et al., 2021) and hypothesis testing (Broniatowski & Keziou, 2009), to name a few. Mutual information (MI), in particular, which is defined as the KL divergence between the joint distribution of a pair of variables and their marginals (and can be generalized to divergences other than KL), plays a crucial role in Bayesian networks and (conditional) independence (Cheng et al., 2002), self-supervised learning via contrastive losses (van den Oord et al., 2018; Le-Khac et al., 2020) as well as in representation learning (Hjelm et al., 2019; Chen et al., 2016).

Classical divergence estimators perform reasonably well for low dimensional cases, however they scale poorly to large, high dimensional datasets which are typically encountered in modern machine learning. The most compelling estimation approach of a divergence is via the optimization of a lower variational bound parametrized by neural networks. These lower bounds, which are likelihood-free approximations, are maximized in order to compute the divergence value at the optimizer. Well-known variational representations are the Legendre transformation of an f-divergence (Broniatowski & Keziou, 2006; Nguyen et al., 2010) as well as the Donsker-Varadhan (DV) variational formula (Donsker & Varadhan, 1983) for KL divergence and its extension to Rényi divergence (Birrell et al., 2020b). Their tractability stems from their objective functionals, which are computed from expected values and approximated using statistical averages from the available or generated samples.

Despite the scalability and tractability, the estimation of a divergence based on variational formulas is a notoriously difficult problem. One challenge stems from the potentially high bias, since any approximation for the worst case scenario requires an exponential number of samples in order to attain the true divergence value (McAllester & Stratos, 2020). Additionally, the statistical variance, which scales exponentially with respect to the divergence's value for certain variational estimators (Song & Ermon, 2019), is often prohibitively high. Focusing on the elevated MI, there are several further lower bounds (Barber & Agakov, 2003; Belghazi et al., 2018; van den Oord et al., 2018; Poole et al., 2019; Guo et al., 2021) and a few upper bounds (Cheng et al., 2020; Poole et al., 2019) which aim to provide more reliable estimates of MI in the low sample size regime. However, the majority of these MI estimators are not transferable to the general estimation of divergences and frequently produce instabilities during training which are further magnified by the small batch and/or sample size.

In this paper, we propose to reduce a divergence estimator's variance via an explicit variance penalty (VP) which is added to the objective functional. Our contributions are summarized as follows:

- We present a novel variance reduction penalty for f-divergence and expand it via the delta method to the nonlinear setting, including the DV formula for KL divergence as well as the variational formula for the Rényi divergences. The proposed VP is able to flexibly trade off bias and variance.
- We present numerical evidence on synthetic datasets that the proposed approach improves both mean squared error (MSE) and median absolute error (MedAE) in a range of sample sizes and types of divergences. Furthermore, we implemented the proposed VP in several other lower and upper bounds of MI, showing that our variance reduction approach is not restricted to particular variational formulas but it is generic and applicable to the majority of existing variational representations.
- When applied to real datasets, we demonstrate the ability of the proposed approach to reduce the variance of the estimated Rényi divergence, thus enabling the detection of rare biological sub-populations which are otherwise difficult to identify. Interestingly, the baseline estimator is unstable when the order value is above one, but it becomes stable when the VP is added.
- We also applied the VP to the disentangled representation learning of speech into its text, speaker, and style components. Results on objective evaluation metrics showed that the addition of the VP generally improves the training performance, as much as 18% relative to the baseline systems.

1.1 Related Work

There are several general-purpose variance reduction techniques in Monte Carlo stochastic sampling, with the most popular approaches being antithetic sampling or more broadly coupling methods, control of variates and importance sampling (Robert & Casella, 2005; Glasserman, 2004; Srinivasan, 2013). These methods have not been explicitly applied for the variational divergence estimation problem. We speculate that either they are not applicable due to the unavailability of analytical probability density formulas or they are inefficient (e.g., the control of variates approach requires a second estimator and potentially a second parametric model in order to be applied).

Another way to reduce the variance is to restrict the function space to more smooth and/or controlled test (or critic) functions, balancing again between bias and variance. For instance, the restriction to Lipschitz continuous functions has the potential to reduce the variance since there exist favorable concentration inequality results for the Lipschitz space (Wainwright, 2019). In the GAN literature, Wasserstein GAN (Gulrajani et al., 2017) and spectral normalization (Miyato et al., 2018) impose Lipschitz continuity which resulted in signigicant gains in terms of training stability. Similarly, the restriction of test functions to an appropriately designed reproducing kernel Hilbert space could reduce the variance (Sreekar et al., 2020). Such approaches can be combined with our proposed variance penalties, as our formulation allows for general test-function spaces. However, we do not focus on this point here.

Given the importance of MI, several estimators aim towards improved statistical properties. Lower bounds such as MINE (Belghazi et al., 2018), which uses the DV variational formula with an expo-

nential moving average, NWJ estimator (Nguyen et al., 2010) and BA estimator (Barber & Agakov, 2003) as well as upper bounds such as CLUB (Cheng et al., 2020) still have high variance. InfoNCE (van den Oord et al., 2018) is one of the few MI estimators that has low variance, but at the cost of either high bias or high computational cost due to the need for many negative samples and thus large batch size. Poole et al. (2019) and Guo et al. (2021) aim to clarify the relationships and trade-offs between those variational bounds. A different approach to reducing variance is by appropriately working on the gradients of the objective function (Wen et al., 2020; 2021).

Finally, we discuss the approach of truncating the test function inside a bounded region as proposed in (Song & Ermon, 2019). The determination of the truncation threshold is quite difficult since it requires an a priori understanding of the log-likelihood ratio. Moreover, a high truncation threshold will not affect the estimation since a high threshold implies no real benefit in terms of variance reduction. On the other hand, a low threshold will result in large bias. Overall, using a high truncation threshold in order to avoid extreme values is a good practice even though it will have a limited impact on variance reduction.

2 Background on Variational Formulas for Rényi and f-Divergences.

While our variance reduction method can be applied to any divergence that possesses a variational formula, here our focus will be on the Rényi and f-divergences, including the KL divergence. For Rényi divergences an appropriate objective functional can be constructed from a difference of cumulant generating functions (Birrell et al., 2020b)

$$R_\alpha(Q\|P) = \sup_{g\in\mathcal{M}_b(\Omega)} \left\{ \frac{1}{\alpha-1}\log E_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_P\left[e^{\alpha g}\right] \right\}, \quad \alpha \neq 0, 1\,. \tag{1}$$

Here Q and P are probability distributions on the set Ω, E_Q and E_P denote the expectations with respect to Q and P respectively, and $\mathcal{M}_b(\Omega)$ is the space of bounded measurable real-valued functions on Ω. For f divergences, f being a lower semicontinuous convex function with $f(1) = 0$, one has the well-known Legendre transform variational formula (Broniatowski & Keziou, 2006; Nguyen et al., 2010)

$$D_f(Q\|P) = \sup_{g\in\mathcal{M}_b(\Omega)} \{E_Q[g] - E_P[f^*(g)]\}\,, \tag{2}$$

where $f^*(y) = \sup_{x\in\mathbb{R}}\{yx - f(x)\}$ is the Legendre transform of f. Here and in the following, the function of g that is being optimized will be called the objective functional. Equation (2) can be generalized to the (f,Γ)-divergences (Birrell et al., 2020a), where $\Gamma \subset \mathcal{M}_b(\Omega)$ is a restricted test-function space

$$D_f^\Gamma(Q\|P) = \sup_{g\in\Gamma}\{E_Q[g] - \Lambda_f^P[g]\}\,, \tag{3}$$

$$\Lambda_f^P[g] = \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\}\,. \tag{4}$$

In particular, if $f_{\text{KL}}(x) = x\log(x)$ corresponds to the KL divergence then

$$\Lambda_{f_{\text{KL}}}^P[g] = \log(E_P[\exp(g)]) \equiv \Lambda^P[g] \tag{5}$$

is the classical cumulant generating function and equation (3) (with $\Gamma = \mathcal{M}_b(\Omega)$) becomes the Donsker-Varadhan variational formula (Dupuis & Ellis., 1997, Appendix C.2)

$$D_{\text{KL}}(Q\|P) = \sup_{g\in\mathcal{M}_b(\Omega)} \{E_Q[g] - \log E_P[e^g]\}\,. \tag{6}$$

For general f, we will often write equation (3) as

$$D_f^\Gamma(Q\|P) = \sup_{g\in\Gamma,\nu\in\mathbb{R}} \{E_Q[g-\nu] - E_P[f^*(g-\nu)]\} \tag{7}$$

and if Γ is closed under the shifts $g \mapsto g-\nu$, $\nu\in\mathbb{R}$ then we can write it simply as

$$D_f^\Gamma(Q\|P) = \sup_{g\in\Gamma}\{E_Q[g] - E_P[f^*(g)]\}\,. \tag{8}$$

In particular, if $\Gamma = \mathcal{M}_b(\Omega)$ then $D_f^\Gamma = D_f$. The generalizations of Rényi and KL divergence obtained by using a restricted space Γ in place of $\mathcal{M}_b(\Omega)$ in equation (1) or equation (6) will be denoted by R_α^Γ and D_{KL}^Γ respectively (Dupuis & Mao, 2019).

3 STATISTICAL ESTIMATORS AND VARIANCE REDUCTION

Variational representations of divergences are especially useful for creating statistical estimators in a data-driven setting; a naive estimator is obtained by simply replacing expectations with the corresponding statistical averages in any of the equations (1), (2), (3), etc. More formally, the naive estimators can be written as $D_f^\Gamma(Q_n\|P_n)$, $R_\alpha^\Gamma(Q_n\|P_n)$, etc., where Γ is some parameterized space of functions (e.g., a neural network), Q_n and P_n are the n-sample empirical measures from Q and P respectively (i.e., $E_{P_n}[g] = \frac{1}{n}\sum_{j=1}^n g(X_j)$ where X_j are i.i.d. samples from P and similarly for E_{Q_n}; we also assume that the samples from Q and P are independent of one another), and the divergences are expressed in terms of the variational formulas from Section 2. However, in practice these naive methods often suffer from high variance (Song & Ermon, 2019; Birrell et al., 2020b). We address this via variance-penalized divergences, which are constructed by introducing a variance penalty into the objective functional of the variational representation, e.g.,

$$D_f^\lambda(Q\|P) \equiv \sup_{g\in\mathcal{M}_b(\Omega)} \left\{E_Q[g] - E_P[f^*(g)] - \lambda V[g;Q,P]\right\}, \tag{9}$$

where the variance penalty, λV, is proportional to the variance of $E_{Q_n}[g]-E_{P_n}[f^*(g)]$ with strength $\lambda > 0$. Using this, we construct the following divergence estimator

$$\sup_\eta \{E_{Q_n}[g_\eta] - E_{P_n}[f^*(g_\eta)] - \lambda V[g_\eta;Q_n,P_n]\}, \tag{10}$$

where g_η is a neural network with parameters η; see the pseudocode in Algorithm 1 below. A similar method applies to other divergences with variational representations.

3.1 VARIANCE PENALTY

In this subsection we provide details on the variance penalty for (f,Γ)-divergences, the KL-divergence, and Rényi divergences. The same technique can be applied to other divergences with a variational representation.

To introduce the variance penalty, first consider the (f,Γ)-divergence representation equation (8). Our goal is to penalize g's for which $E_{Q_n}[g]$ or $E_{P_n}[f^*(g)]$ have large variance, hence we introduce a penalty term proportional to (Var_Q denotes variance with respect to Q, etc.)

$$\mathbb{V}\mathrm{ar}\left[E_{Q_n}[g] + E_{P_n}[f^*(g)]\right] = \frac{1}{n}\left(\mathrm{Var}_Q[g] + \mathrm{Var}_P[f^*(g)]\right). \tag{11}$$

Specifically, for $\lambda > 0$ we define the variance-penalized (f,Γ)-divergence

$$D_f^{\Gamma,\lambda}(Q\|P) \equiv \sup_{g\in\Gamma,\nu\in\mathbb{R}} \{E_Q[g-\nu] - E_P[f^*(g-\nu)] - \lambda(\mathrm{Var}_Q[g-\nu] + \mathrm{Var}_P[f^*(g-\nu)])\}. \tag{12}$$

As noted above, if Γ is invariant under constant shifts then the optimization over ν can be omitted. A similar result to equation (12) can be derived for any objective functional that is a linear combination of expectations, e.g., integral probability metrics (Müller, 1997; Sriperumbudur et al., 2012) such as the Wasserstein metric.

For nonlinear objective functional terms of the form $G(E_P[h(g)])$, such as appear in equation (1) and equation (8), we cannot compute the variance of the corresponding statistical estimator at finite n but we can use the delta method to obtain the asymptotic variance

$$\lim_{n\to\infty} n\mathbb{V}\mathrm{ar}\left[G(E_{P_n}[h(g)])\right] = (G'(E_P[h(g)])^2\mathrm{Var}_P[h(g)]. \tag{13}$$

Motivated by this, we propose the following variance-penalized KL and Rényi divergence variational formulas:

$$D_{\mathrm{KL}}^{\Gamma,\lambda}(Q\|P) \equiv \sup_{g\in\Gamma}\left\{E_Q[g] - \log E_P[e^g] - \lambda\left(\mathrm{Var}_Q[g] + \mathrm{Var}_P[e^g]/(E_P[e^g])^2\right)\right\}, \tag{14}$$

$$R_\alpha^{\Gamma,\lambda}(Q\|P) \equiv \sup_{g\in\Gamma}\left\{\frac{1}{\alpha-1}\log E_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_P\left[e^{\alpha g}\right] - \lambda\left(\frac{1}{(\alpha-1)^2}\frac{\mathrm{Var}_Q[e^{(\alpha-1)g}]}{(E_Q[e^{(\alpha-1)g}])^2} + \frac{1}{\alpha^2}\frac{\mathrm{Var}_P[e^{\alpha g}]}{(E_P[e^{\alpha g}])^2}\right)\right\}. \tag{15}$$

Remark 1. *Both equation (11) and equation (13) suggest that the statistical estimators for the above penalized divergences should use a variance penalty strength that decays with the sample size $\lambda = \lambda_0/n$, though other forms of n-dependence may be useful in practice.*

Though the variance penalty introduces bias, as $\lambda \to 0$ the penalized divergence converges to the corresponding non-penalized divergence, as made precise by the following theorem.

Theorem 2. *Let $\Gamma \subset \mathcal{M}_b(\Omega)$. We have the following convergence results:*

$$\lim_{\lambda \to 0^+} D_{KL}^{\Gamma,\lambda}(Q\|P) = D_{KL}^{\Gamma}(Q\|P)\,, \tag{16}$$

$$\lim_{\lambda \to 0^+} R_{\alpha}^{\Gamma,\lambda}(Q\|P) = R_{\alpha}^{\Gamma}(Q\|P)\,, \tag{17}$$

and if $f^*(y) < \infty$ for all $y \in \mathbb{R}$ then

$$\lim_{\lambda \to 0^+} D_{f}^{\Gamma,\lambda}(Q\|P) = D_{f}^{\Gamma}(Q\|P)\,. \tag{18}$$

Remark 3. *Note that the corresponding statistical estimators, $D_f^{\Gamma,\lambda}(Q_n\|P_n)$, etc., have additional bias due to the supremum over g. We present partial results on bias bounds in Appendix D.*

The proof of Theorem 2 is given in Appendix C. Moreover, there we also prove that

$$\lim_{\lambda \to \infty} D_{\mathrm{KL}}^{\Gamma,\lambda}(Q\|P) = \lim_{\lambda \to \infty} R_{\alpha}^{\Gamma,\lambda}(Q\|P) = \lim_{\lambda \to \infty} D_{f}^{\Gamma,\lambda}(Q\|P) = 0 \tag{19}$$

under appropriate assumptions (see Theorem 13). The same proof techniques can be applied to other divergences with a variational characterization.

Finally, for non-zero λ the penalized divergences (12), (14), (15) retain the divergence property and are therefore appropriate for quantifying the "distance" between probability distributions:

Theorem 4. *Under appropriate assumptions on f and Γ (see Appendix B for details) and letting $D^{\Gamma,\lambda}$ denote any of $D_f^{\Gamma,\lambda}$, $D_{KL}^{\Gamma,\lambda}$, or $R_\alpha^{\Gamma,\lambda}$ we have $D^{\Gamma,\lambda}(Q\|P) \geq 0$ and $D^{\Gamma,\lambda}(Q\|P) = 0$ if and only if $Q = P$.*

3.2 Variance-Reduced Divergence Estimation Algorithm

We now propose the following divergence neural estimation (DNE) methods with variance penalty, generalizing equations (9)-(10).

(DNE-VP$_\lambda$) $$\sup_{\eta}\{H[g_\eta; Q_n, P_n] - \lambda V[g_\eta; Q_n, P_n]\}\,. \tag{20}$$

We will compare the above method to the non-penalized estimator (i.e., with $\lambda = 0$)

(DNE) $$\sup_{\eta} H[g_\eta; Q_n, P_n]\,. \tag{21}$$

In the above, the test function space is a neural network $\Gamma = \{g_\eta, \eta \in E\}$ with parameters η and H denotes the objective functional of the divergence, e.g., for the Rényi divergences (1)

$$H_\alpha[g; Q, P] = \frac{1}{\alpha - 1}\log E_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_P\left[e^{\alpha g}\right]\,,\quad \alpha \neq 0, 1 \tag{22}$$

and for f divergences (2)

$$H_f[g; Q, P] = E_Q[g] - E_P[f^*(g)]\,. \tag{23}$$

Finally, V is the variance penalty corresponding to the chosen divergence (see Section 3.1), e.g., for Rényi divergences

$$V_\alpha[g; Q_n, P_n] = \frac{1}{(\alpha-1)^2}\frac{\mathrm{Var}_{Q_n}[e^{(\alpha-1)g}]}{(E_{Q_n}[e^{(\alpha-1)g}])^2} + \frac{1}{\alpha^2}\frac{\mathrm{Var}_{P_n}[e^{\alpha g}]}{(E_{P_n}[e^{\alpha g}])^2}\,,\quad \alpha \neq 0, 1 \tag{24}$$

and for f divergences

$$V_f[g; Q_n, P_n] = \mathrm{Var}_{Q_n}[g] + \mathrm{Var}_{P_n}[f^*(g)]\,. \tag{25}$$

We solve equation (20) using a stochastic gradient descent method. Pseudocode for DNE-VP is presented in Algorithm 1.

Algorithm 1 Divergence Neural Estimation with Variance Penalty (DNE-VP): Recall that $E_{Q_m}[g] = \frac{1}{m}\sum_{j=1}^{m} g(X_j)$ and similarly for P_m. We utilize the Adam optimizer (Kingma & Ba, 2014) with parameters α, β_1, β_2, but one can replace it with any other optimization method.

for K steps **do**
  Generate independent samples $X_1, ..., X_m \sim Q$, $Y_1, ..., Y_m \sim P$
  $\eta \leftarrow \text{Adam}(\nabla_\eta(-H[g_\eta; Q_m, P_m]), \eta, \alpha, \beta_1, \beta_2)$
end for
return $H[g_\eta; Q_N, P_N]$

4 Results on Synthetic Datasets

Figure 1 presents the statistical estimation of Rényi divergence between two one-dimensional Gaussians which both have zero mean but different variance values. The order of Rényi divergence, α, controls how much weight to put on the tails of the distributions, thus it can become very sensitive to the few samples from the tails. The same conclusion can be deduced from the variational formula (i.e., equation (1) where α multiplies the exponentials' argument). Therefore, a larger α value implies larger statistical variance. Indeed, high estimation variance is observed with DNE (upper leftmost panel of Figure 1) despite the fact that we applied truncation as proposed by Song & Ermon (2019) with truncation threshold set to 1. In contrast, the DNE-VP$_\lambda$ estimator with $\lambda = 0.1$ greatly reduces the statistical variance even when α is large (lower leftmost panel). For fairness, we imposed the same truncation operation in the output of DNE-VP$_\lambda$. We report a 80% reduction of variance for $\alpha = 2$ which becomes 99% for $\alpha = 10$.

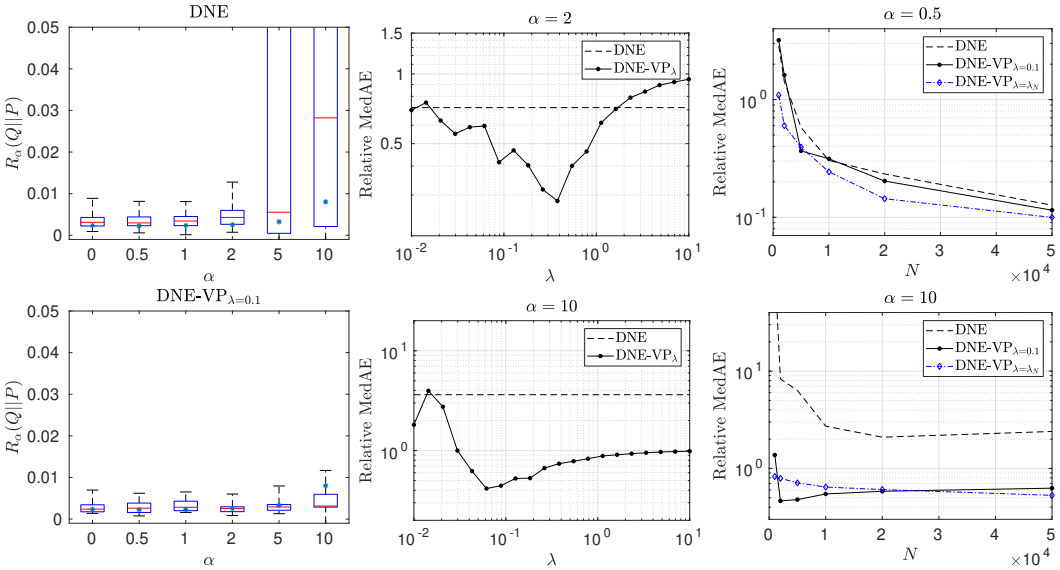

Figure 1: Comparison between the estimator without VP (DNE) and with VP (DNE-VP$_\lambda$) for Rényi divergence between two one-dimensional Gaussians with $Q = \mathcal{N}(0, 1.1)$ and $P = \mathcal{N}(0, 1)$. We use $N = 5K$ sample size, 512 as batch size and results are averaged over 100 i.i.d. runs. **Left column:** DNE and DNE-VP$_\lambda$ estimators for increasing values of α. The variance of DNE becomes uncontrollably high for $\alpha > 3$. **Middle column:** Relative MedAE (the lower, the better) for varying penalty coefficient λ and two values of α. The relative MedAE for large values of λ is close to one which implies that the estimated value of DNE-VP$_\lambda$ approaches zero. **Right column:** Relative MedAE for increasing sample size N. We additionally present a penalty coefficient that varies with sample size, shown in blue ($\lambda_N = \frac{500}{N}$ and $\lambda_N = \frac{2000}{N}$ for $\alpha = 0.5$ and $\alpha = 10$, respectively).

The proposed approach introduces an additional hyper-parameter, λ, which controls the strength of the VP. Our theory suggests that λ should depend on the sample size (and perhaps also on the other parameters), therefore we perform two sets of experiments. In the first experiment, we explore the

range of optimal values for λ in terms of MedAE[1]. As is evident from the middle panels of Figure 1, λ-values in the vicinity of 0.1 are a reasonable compromise between variance and bias. In the second experiment, we demonstrate the performance in terms of MedAE as a function of the sample size, N. As suggested in Remark 1, monotone performance is obtained when λ is inversely proportional to N (blue dashed line in rightmost upper panel of Figure 1).

Our second synthetic example constitutes the estimation of MI using various approaches with and without VP. Here, we let Q be a zero-mean multivariate correlated Gaussian random vector of dimension d. We impose element-wise correlation, i.e., $corr(x_i, x_{\frac{d}{2}+j}) = \delta_{i,j}\rho$ to the samples $x \sim Q$ where $i, j = 1, \ldots, \frac{d}{2}$ and $\delta_{i,j}$ is Kronecker's delta. With P we denote the product of the marginals, which in this case is simply a zero-mean standardized multivariate Gaussian. Figure 2 presents the estimated MI per training step. We consider the Renyi-based MI with $\alpha = 0.5$ as well as the standard MI using the DV variational formula. Notice that these two variants result in different true values (black lines in Figure 2). The plotted results demonstrate the successful reduction of variance when VP is added to the objective functional. Interestingly, the extension of VP to InfoNCE and CLUB estimators (second row of panels in Figure 2) implies that our approach can be applied to any MI estimators, thus offering a general variance reduction framework. Bias, variance and MSE plots as well as several more experiments can be found in Appendix F.

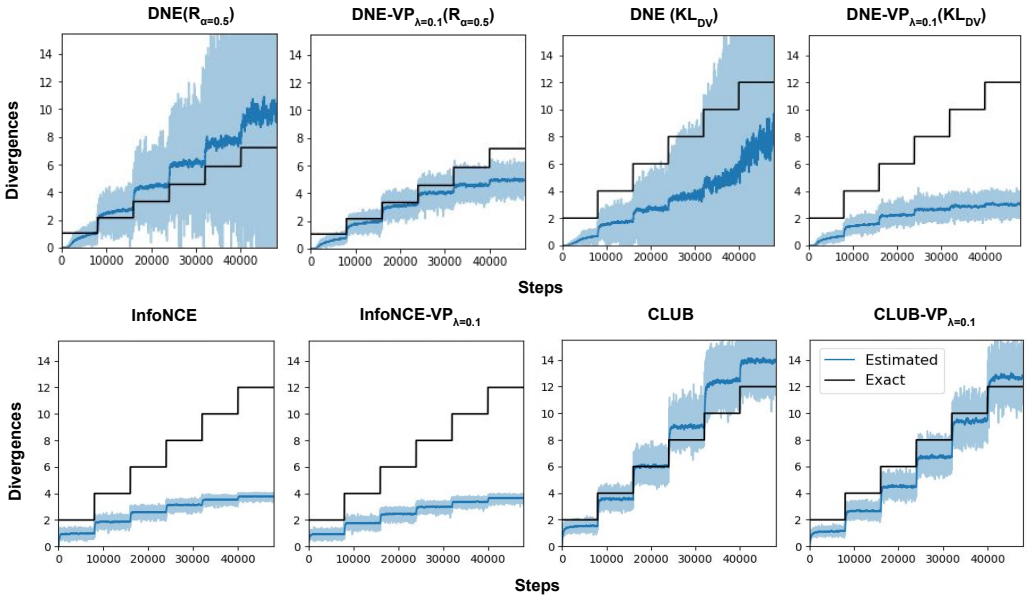

Figure 2: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. The number of samples is set to $512K$ and batch size to 64. Panels with $R_{\alpha=0.5}$ in their titles present the Rényi-based MI with $\alpha = 0.5$ whereas the rest of the methods estimate the standard MI (i.e., the KL divergence). In each panel, the true values are shown as a step function (black line). The correlation coefficient of the Gaussian, ρ, for each step is: 0.3084, 0.4257, 0.5091, 0.5741, 0.6273 and 0.6717. The running estimates per minibatch are displayed as shadow blue curves. The dark blue curves shows the moving average of the estimated MI, with a bandwidth equal to 200 steps.

5 Real Data Applications

5.1 Detecting Rare Biological Sub-Populations

Using the dataset from Levine et al. (2015), we test the efficacy of DNE-VP$_\lambda$ in discriminating cell populations which are contaminated with a rare sub-population with distinguishable statistical

[1]Recall that MedAE stands for median absolute error and it is a more robust-to-outliers metric.

properties. Specifically, we consider single-cell mass cytometry measurements on 16 bone marrow protein markers[2] (i.e., $d = 16$) coming from healthy and diseased individuals with acute myeloid leukemia. For each run we created three subsets of healthy samples with sample size $N = 20K$ which we denote by P and one dataset as a mixture of 99% healthy and 1% diseased samples which is denoted by Q. Notice that the actual number of diseased samples is only 200 thus it is considered as a rare sub-population.

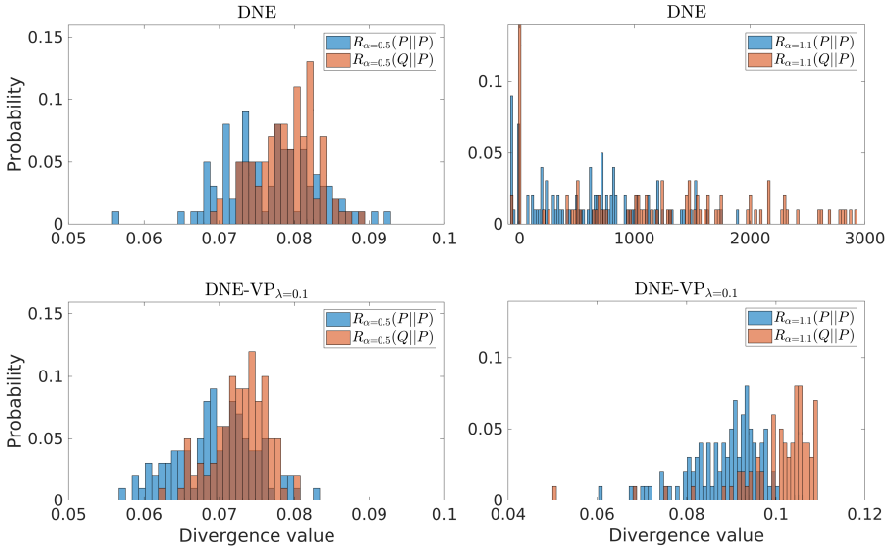

Figure 3: Comparison of DNE and DNE-VP$_\lambda$ estimators for Rényi divergence on biological data. The histograms of the estimated divergence value are constructed from 100 i.i.d. runs between datasets of $N = 20K$ samples each. Healthy dataset's distribution is denoted by P whereas healthy + 1% diseased dataset's by Q. **Left column:** Rényi divergence with $\alpha = 0.5$. Neither DNE nor DNE-VP$_\lambda$ are able to discriminate between the healthy and the 1% contaminated dataset. **Right column:** Rényi divergence with $\alpha = 1.1$. For this α value, VP is compulsory for a stable estimation of Rényi divergence. Furthermore, we are able to discriminate between healthy and 1% contaminated distributions with high accuracy (87.5%).

For Rényi divergence with $\alpha = 0.5$ (left panels in Figure 3), both DNE and DNE-VP$_\lambda$ are stable. Despite the improvement in the separation of the two histograms, the observed variance reduction of DNE-VP$_\lambda$ is minimal and not enough to discriminate between the healthy and the contaminated with 1% diseased samples distributions. When considering Rényi divergence with $\alpha = 1.1$, we observe that DNE fails to produce stable estimates. In contrast, DNE-VP$_\lambda$ always computes stable estimates. Additionally, the two histograms are satisfactorily separated, implying that larger values of α are crucial, provided there is a way to handle the statistical variance. For completeness, Table 1 reports the first and second order statistics of the histograms shown in Figure 3.

Table 1: Mean values and standard deviation for the histograms shown in Figure 3.

Divergence	DNE		DNE-VP$_{\lambda=0.1}$	
	mean	std	mean	std
$R_{\alpha=0.5}(P\|\|P)$	0.0765	0.0066	0.0695	0.0053
$R_{\alpha=0.5}(Q\|\|P)$	0.0789	0.0039	0.0720	0.0036
$R_{\alpha=1.1}(P\|\|P)$	676	515	0.0890	0.0089
$R_{\alpha=1.1}(Q\|\|P)$	1445	1165	0.1000	0.0120

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

5.2 DISENTANGLED REPRESENTATION LEARNING IN SPEECH SYNTHESIS

An important application of MI is disentangled representation learning. In the context of representation disentanglement, the extraction of meaningful latent features for high-dimensional data is challenging, especially when explicit knowledge needs to be distilled into interpretable representations. One popular approach to enforce representation disentanglement is via MI minimization. Moreover, a superior disentanglement will allow a greater degree of interpretability and controllability, especially for generative models maintaining high production capacity. In this section, we employ the proposed DNE-VP$_\lambda$ estimator for MI estimation in order to learn disentangled representation, and, particularly, in the context of speech synthesis and analysis.

A universal text-to-speech synthesizer can generate speech from text with speaker factor and speaking style similar to a reference signal. Previous works aimed to encode the information from reference speech into a fixed-length style and speaker embedding using trainable encoders (Wang et al., 2018; Tjandra et al., 2020; Chien et al., 2021; Tan et al., 2021). The major challenges for such speech synthesizers are controllability and generalisability, especially when trying to generalize the models with multiple speakers and multiple styles. During training, content information is leaked into the style embeddings ("content leakage") and speaker information into style embeddings ("style leakage"). Thus at inference, when the reference speech has different content from the input text, the decoder expects the content from the style vector ignoring some part of the content text. Moreover, speaker information could be expected from the style encoder leading to completely different speaker attribute. To alleviate that, Paul et al. (2021) suggested replacing the KL-based MI with Rényi-based MI and minimizing the Rényi divergence between the joint distribution and the product of marginals for the content-style and style-speaker pairs. However, reliable estimation of Rényi divergence was problematic due to high statistical variance. Taking advantage of the proposed variance reduction technique, we employ a VP term in the loss function which is denoted as DNE-VP$_\lambda$ (R_α). By doing so, content, style, and speaker spaces become representative and (ideally) independent of each other. We introduce two variations of this framework: sum of three Rényi divergences DNE($R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1}$) (i.e., sum of the corresponding objective functionals) and DNE($R_{\alpha=0.5}$). We tested several different λ values, aiming to reduce the statistical variance of the adversarial component. Notice that larger λ values were helpful in this application.

Table 2: Objective evaluation tests. Lower scores indicate better performance.

Methods	No Shuffle			Shuffle		
	RMSE-F0	MCD	WER(%)	RMSE-F0	MCD	WER(%)
DNE ($R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1}$)	28.59	5.35	21.6	45.75	6.39	28.7
DNE ($R_{\alpha=0.5}$)	28.59	5.27	18.3	47.26	6.60	26.6
DNE-VP$_{\lambda=5}$ ($R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1}$)	30.29	**5.23**	21.2	48.15	**6.39**	27.3
DNE-VP$_{\lambda=10}$ ($R_{\alpha=0} + R_{\alpha=0.5} + R_{\alpha=1}$)	**27.76**	5.36	**18.1**	47.62	6.48	28.7
DNE-VP$_{\lambda=5}$ ($R_{\alpha=0.5}$)	28.69	5.87	**17.3**	46.53	6.72	**25.4**
DNE-VP$_{\lambda=10}$ ($R_{\alpha=0.5}$)	29.71	5.33	22.8	**45.47**	6.54	26.2

We evaluate the performance of disentanglement strategies using three performance scores from 100 random samples shown in Table 2. Mel-cepstral distortion (MCD) measures the spectral distance between the synthesized and reference mel-spectrum features. Root mean squared error (RMSE) evaluates the similarity in F0 modeling between reference and synthesized speech. Lastly, the content preservation criterion is evaluated by word error rate (WER). During inference, we evaluate the performance on two conditions: 'no shuffle' and 'shuffle'. During inference, 'no shuffle' feeds the same reference speech into style and speaker encoders and its corresponding text to predict the speech features, whereas 'shuffle' feeds random speech. We observe that the proposed DNE-VP$_\lambda$ variants outperform baseline approaches without VP in terms of all evaluation metrics. Our proposed systems greatly reduced content leakage by improving the word error rate by approximately 5-18% relative to the baseline systems. Furthermore, RMSE-F0 and MCD scores show that the disentanglement module during training assists the TTS to achieve more accurate rendering of prosodic patterns as well as synthesizing proper speech content to its corresponding text without any significant leakage issues.

REFERENCES

David Barber and Felix V. Agakov. The IM algorithm: A variational approach to information maximization. In *NIPS*, pp. 201–208, 2003. URL https://proceedings.neurips.cc/paper/2003/file/a6ea8471c120fe8cc35a2954c9b9c595-Paper.pdf.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 531–540, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/belghazi18a.html.

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9781118341919. URL https://books.google.com/books?id=a3gavZbxyJcC.

P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118625965. URL https://books.google.com/books?id=6ItqtwaWZZQC.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. *arXiv e-prints*, art. arXiv:2011.05953, November 2020a.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational Representations and Neural Network Estimation of Rényi Divergences. *arXiv e-prints*, art. arXiv:2007.03814, July 2020b.

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

Michel Broniatowski and Amor Keziou. Parametric estimation and tests through divergences and the duality technique. *Journal of Multivariate Analysis*, 100(1):16–36, 2009. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2008.03.011. URL https://www.sciencedirect.com/science/article/pii/S0047259X08001036.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2180–2188, 2016.

Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137:43–90, 2002.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *ICML 2020*, July 2020. URL https://www.microsoft.com/en-us/research/publication/club-a-contrastive-log-ratio-upper-bound-of-mutual-information/.

Chung Ming Chien, Jheng Hao Lin, Chien Yu Huang, Po Chun Hsu, and Hung Yi Lee. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. *arXiv preprint arXiv:2103.04088*, 2021.

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

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

P. Dupuis and R.S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. Wiley series in probability and statistics. John Wiley & Sons, New York, 1997. ISBN 0-471-07672-4. URL http://opac.inria.fr/record=b1092351. A Wiley-Interscience Publication.

Paul Dupuis and Yixiang Mao. Formulation and properties of a divergence used to compare probability measures without absolute continuity. *arXiv e-prints*, art. arXiv:1911.07422, November 2019.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67:4986–5004, 2021.

G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. ISBN 9781118626399. URL https://books.google.com/books?id=wI4fAwAAQBAJ.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Applications of mathematics : stochastic modelling and applied probability. Springer, 2004. ISBN 9780387004518. URL https://books.google.gr/books?id=e9GWUsQkPNMC.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative Adversarial Nets. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, pp. 2672–2680, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved Training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Qing Guo, Junya Chen, Dong Wang, Yuewei Yang, Xinwei Deng, Lawrence Carin, Fan Li, and Chenyang Tao. Tight mutual information estimation with contrastive fenchel-legendre optimization. *CoRR*, abs/2107.01131, 2021.

Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR 2019*. ICLR, April 2019. URL https://openreview.net/pdf?id=Bklr3j0cKX.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020. ISSN 2169-3536. doi: 10.1109/access.2020.3031549. URL http://dx.doi.org/10.1109/ACCESS.2020.3031549.

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. ISSN 10974172. doi: 10.1016/j.cell.2015.05.047. URL http://dx.doi.org/10.1016/j.cell.2015.05.047.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/mcallester20a.html.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. URL http://www.jstor.org/stable/1428011.

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

S. Nowozin, B. Cseke, and R. Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 271–279, 2016.

Dipjyoti Paul, Sankar Mukherjee, Yannis Pantazis, and Yannis Stylianou. A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning Based on Rényi Divergence Minimization. In *Proc. Interspeech 2021*, pp. 3625–3629, 2021. doi: 10.21437/Interspeech.2021-660.

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 (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Christian P Robert and George Casella. *Monte Carlo statistical methods; 2nd ed.* Springer texts in statistics. Springer, Berlin, 2005. URL https://cds.cern.ch/record/1187871.

Jiaming Song and Stefano Ermon. Understanding the Limitations of Variational Mutual Information Estimators. *arXiv e-prints*, art. arXiv:1910.06222, October 2019.

P. Aditya Sreekar, Ujjwal Tiwari, and Anoop M. Namboodiri. Reducing the variance of variational estimates of mutual information by limiting the critic's hypothesis space to RKHS. In *ICPR*, pp. 10666–10674. IEEE, 2020.

R. Srinivasan. *Importance Sampling: Applications in Communications and Detection*. Springer Berlin Heidelberg, 2013. ISBN 9783662050521. URL https://books.google.com/books?id=SgmrCAAAQBAJ.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electron. J. Statist.*, 6:1550–1599, 2012. doi: 10.1214/12-EJS722. URL https://doi.org/10.1214/12-EJS722.

Daxin Tan, Hingpang Huang, Guangyan Zhang, and Tan Lee. CUHK-EE voice cloning system for ICASSP 2021 M2VoC challenge. *arXiv preprint arXiv:2103.04699*, 2021.

Andros Tjandra, Ruoming Pang, Yu Zhang, and Shigeki Karita. Unsupervised learning of disentangled speech content and style representation. *arXiv preprint arXiv:2010.12973*, 2020.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pp. 5180–5189. PMLR, 2018.

Liangjian Wen, Yiji Zhou, Lirong He, Mingyuan Zhou, and Zenglin Xu. Mutual information gradient estimation for representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ByxaUgrFvH.

Liangjian Wen, Haoli Bai, Lirong He, Yiji Zhou, Mingyuan Zhou, and Zenglin Xu. Gradient estimation of information measures in deep learning. *Knowledge-Based Systems*, 224:107046, 2021. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2021.107046. URL https://www.sciencedirect.com/science/article/pii/S0950705121003099.

A NOTATION

We denote the extended reals by $\overline{\mathbb{R}} \equiv \mathbb{R} \cup \{-\infty, \infty\}$. We let $(\Omega, \mathcal{M})$ be a measurable space and $\mathcal{P}(\Omega)$ be the set of probability measures on $(\Omega, \mathcal{M})$. For $P \in \mathcal{P}(\Omega)$ we let E_P denote the expectation with respect to P and Var_P denote the variance with respect to P. We let Q_n, P_n be n-sample empirical measures, constructed using i.i.d. samples from Q and P respectively. A set $\Psi \subset \mathcal{P}(\Omega)$ will be called $\mathcal{P}(\Omega)$**-determining** if whenever $Q, P \in \mathcal{P}(\Omega)$ satisfy $E_Q[g] = E_P[g]$ for all $g \in \Psi$ we have $Q = P$. A map $D : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to [0, \infty]$ will be said to have the **divergence property** if $D(Q\|P) = 0$ if and only if $Q = P$.

B DIVERGENCE PROPERTY

In this appendix we prove the divergence property for the variance-penalized divergences introduced in Section 3.1 (the definition of the divergence property is found in Appendix A). The techniques we use are based on the proof of Theorem C.3 in Birrell et al. (2020a). Though we focus on the Rényi, KL, and f-divergences the same techniques can be adapted to other divergences with a variational formulation, such as integral probability metrics.

Theorem 5 (Divergence Property for $R_\alpha^{\Gamma,\lambda}$)**.** *Let $\alpha > 0$, $\alpha \neq 1$, $\Gamma \subset \mathcal{M}_b(\Omega)$, $\lambda \geq 0$, and suppose there exists a nonempty set $\Psi \subset \mathcal{M}_b(\Omega)$ such that*

1. *Ψ is $\mathcal{P}(\Omega)$-determining,*
2. *for all $\psi \in \Psi$ there exists $\epsilon_0 > 0$ with $\epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.*

Then $R_\alpha^{\Gamma,\lambda}$ has the divergence property.

Remark 6. *The definition of $\mathcal{P}(\Omega)$-determining can be found in Appendix A.*

Proof. Let $Q, P \in \mathcal{P}(\Omega)$. The second assumption on Ψ implies $0 \in \Gamma$ so we can bound equation (15) below by the value at $g = 0$, which implies $R_\alpha^{\Gamma,\lambda}(Q\|P) \geq 0$. The variance penalty is non-negative and so we have

$$0 \leq R_\alpha^{\Gamma,\lambda}(Q\|P) \leq R_\alpha^{\Gamma}(Q\|P) \leq R_\alpha(Q\|P)\,, \tag{26}$$

where we used equation (1). R_α is known to satisfy the divergence property and so if $Q = P$ then $R_\alpha(Q\|P) = 0$, hence $R_\alpha^{\Gamma,\lambda}(Q\|P) = 0$ as well.

Finally, suppose $R_\alpha^{\Gamma,\lambda}(Q\|P) = 0$. Given $\psi \in \Psi$ let $g_\epsilon = \epsilon\psi \in \Gamma$ for $|\epsilon| < \epsilon_0$ (here we are using the second assumption on Ψ). Then we can bound equation (15) below as follows

$$\begin{aligned} 0 =& R_\alpha^{\Gamma,\lambda}(Q\|P) \\ \geq& \frac{1}{\alpha-1}\log E_Q[e^{(\alpha-1)\epsilon\psi}] - \frac{1}{\alpha}\log E_P\left[e^{\alpha\epsilon\psi}\right] \\ &- \lambda\left(\frac{1}{(\alpha-1)^2}\frac{\mathrm{Var}_Q[e^{(\alpha-1)\epsilon\psi}]}{(E_Q[e^{(\alpha-1)\epsilon\psi}])^2} + \frac{1}{\alpha^2}\frac{\mathrm{Var}_P[e^{\alpha\epsilon\psi}]}{(E_P[e^{\alpha\epsilon\psi}])^2}\right) \equiv h(\epsilon)\,. \end{aligned} \tag{27}$$

It is straightforward to calculate $h(0) = 0$ and $h'(0) = E_Q[\psi] - E_P[\psi]$ (in particular, the first derivative of the variance penalty vanishes at $\epsilon = 0$). These, together with the fact that $h(\epsilon)$ is C^1 and $h(\epsilon) \leq 0$ for all $|\epsilon| < \epsilon_0$ imply

$$0 = h'(0) = E_Q[\psi] - E_P[\psi]\,. \tag{28}$$

This holds for all $\psi \in \Psi$ and Ψ was assumed to be $\mathcal{P}(\Omega)$-determining, hence we conclude that $Q = P$. This completes the proof of the divergence property for $R_\alpha^{\Gamma,\lambda}$. □

Theorem 7 (Divergence Property for $D_{\mathrm{KL}}^{\Gamma,\lambda}$)**.** *Let $\Gamma \subset \mathcal{M}_b(\Omega)$, $\lambda \geq 0$, and suppose there exists a nonempty set $\Psi \subset \mathcal{M}_b(\Omega)$ such that*

1. *Ψ is $\mathcal{P}(\Omega)$-determining,*
2. *for all $\psi \in \Psi$ there exists $\epsilon_0 > 0$ with $\epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.*

Then $D_{KL}^{\Gamma,\lambda}$ has the divergence property.

The proof of Theorem 7 follows the same template as the proof of Theorem 5; we omit the details. The proof for $D_f^{\Gamma,\lambda}$ divergences is slightly more nuanced than the Rényi or KL cases, and so we provide a proof below.

Theorem 8 (Divergence property for $D_f^{\Gamma,\lambda}$)**.** *Let $\lambda > 0$, $\Gamma \subset \mathcal{M}_b(\Omega)$, and $f : \mathbb{R} \to (-\infty, \infty]$ be lower semicontinuous and convex, with $f(1) = 0$. Suppose f and Γ also satisfy the following:*

1. *There exist a nonempty set $\Psi \subset \Gamma$ with the following properties:*
   (a) *Ψ is $\mathcal{P}(\Omega)$-determining.*
   (b) *For all $\psi \in \Psi$ there exists $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ such that $c_0 + \epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.*
2. *f is finite and strictly convex on a neighborhood of 1.*
3. *f^* is finite and C^1 on a neighborhood of $\nu_0 \equiv f'_+(1)$.*

Then $D_f^{\Gamma,\lambda}$ has the divergence property.

Remark 9. *f'_+ denotes the right derivative.*

Proof. The assumptions on f imply that $f^*(\nu_0) = \nu_0$ and $(f^*)'(\nu_0) = 1$ (see Lemma A.9 in Birrell et al. (2020a)). Assumption 1 implies there exists $c_0 \in \Gamma \cap \mathbb{R}$, hence we can bound equation (12) below by its value at $g = c_0$, $\nu = c_0 - \nu_0$ to find

$$D_f^{\Gamma,\lambda}(Q\|P) \geq E_Q[\nu_0] - E_P[f^*(\nu_0)] = 0 \tag{29}$$

(note that the variance penalty vanishes here). We clearly have $D_f^{\Gamma,\lambda} \leq D_f^{\Gamma} \leq D_f$. $D_f(Q\|P) = 0$ when $Q = P$, hence $D_f^{\Gamma,\lambda}(Q\|P) = 0$ as well.

Finally, suppose $D_f^{\Gamma,\lambda}(Q\|P) = 0$. From assumption 1.b, given $\psi \in \Psi$ there exists $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ such that $g_\epsilon \equiv c_0 + \epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$. Therefore

$$\begin{aligned} 0 = D_f^{\Gamma,\lambda}(Q\|P) \geq & E_Q[g_\epsilon - (c_0 - \nu_0)] - E_P[f^*(g_\epsilon - (c_0 - \nu_0))] \\ & - \lambda(\mathrm{Var}_Q[g_\epsilon - (c_0 - \nu_0)] + \mathrm{Var}_P[f^*(g_\epsilon - (c_0 - \nu_0))]) \\ = & \nu_0 + \epsilon E_Q[\psi] - E_P[f^*(\nu_0 + \epsilon\psi)] - \lambda(\mathrm{Var}_Q[\nu_0 + \epsilon\psi] + \mathrm{Var}_P[f^*(\nu_0 + \epsilon\psi)]) \\ \equiv & h(\epsilon)\,. \end{aligned} \tag{30}$$

As computed in equation (29), we have $h(0) = 0$. Together with the fact that $h(\epsilon)$ is C^1 and $h(\epsilon) \leq 0$ for all $|\epsilon| < \epsilon_0$ we can conclude that

$$0 = h'(0) = E_Q[\psi] - E_P[(f^*)'(\nu_0)\psi] = E_Q[\psi] - E_P[\psi] \tag{31}$$

(again, the first derivative of the variance penalty vanishes at $\epsilon = 0$). Equation (31) holds for all $\psi \in \Psi$, a $\mathcal{P}(\Omega)$-determining set. Hence $Q = P$. This completes the proof of the divergence property for $D_f^{\Gamma,\lambda}$. □

We end this appendix by presenting several examples of $\mathcal{P}(\Omega)$-determining sets that can be used to construct Γ's satisfying the assumptions of the above theorems.

1. Exponentials, $e^{c\cdot x}$, $c \in \mathbb{R}^n$, i.e., the moment generating function; see Section 30 in Billingsley (2012).

2. The set of bounded continuous functions on a metric space.
3. The set of continuous functions on a metric space with $\|g\|_\infty \leq 1$ ($\|\cdot\|_\infty$ denotes the supremum norm).
4. The set of bounded 1-Lipschitz functions on a metric space.
5. The set of bounded 1-Lipschitz functions, g, on a metric space with $\|g\|_\infty \leq 1$ (items 2-5 follow from Theorem 2.1 in Billingsley (2013)).
6. The unit ball in a reproducing kernel Hilbert space, under appropriate assumptions (see Sriperumbudur et al. (2011)).
7. The set of ReLU neural networks. This follows from the universal approximation theorem (Cybenko, 1989) and also applies to other activation functions, e.g., sigmoid.
8. The set of ReLU neural networks with spectral normalization (Miyato et al., 2018).

C CONVERGENCE PROOFS

In this appendix we consider the limit of the variance-penalized divergences as the penalty strength approaches 0 or ∞. First we consider the easier case where the penalty strength approaches zero. For this we will use the following general convergence result.

Lemma 10. *Let $H[g] \in \mathbb{R}$, $V[g] \in [0, \infty)$ where $g \in \Gamma$ is an arbitrary index set. Define $D = \sup_{g\in\Gamma} H[g]$ and for $\lambda > 0$ define $D^\lambda = \sup_{g\in\Gamma}\{H[g] - \lambda V[g]\}$. Then*

$$\lim_{\lambda\to 0^+} D^\lambda = D. \tag{32}$$

Proof. The $V[g]$ are non-negative, hence D^λ is non-increasing in λ. Therefore

$$\lim_{\lambda\to 0^+} D^\lambda = \sup_{\lambda>0} D^\lambda = \sup_{\lambda>0}\sup_{g\in\Gamma}\{H[g] - \lambda V[g]\} = \sup_{g\in\Gamma}\sup_{\lambda>0}\{H[g] - \lambda V[g]\} = D\,. \tag{33}$$

□

The objective functionals of $D_{\mathrm{KL}}^{\Gamma,\lambda}$ and $R_\alpha^{\Gamma,\lambda}$ satisfy the assumptions of Lemma 10, as does $D_f^{\Gamma,\lambda}$'s if $f^*(y) < \infty$ for all $y \in \mathbb{R}$. Therefore Theorem 2 follows as a corollary to Lemma 10.

Next we investigate the limit as $\lambda \to \infty$. The proofs in this case are more involved than the $\lambda \to 0^+$ limit. We start with the following lemma.

Lemma 11. *Let $H[g] \in \mathbb{R}$, $V[g] \in [0, \infty)$ where $g \in \Gamma$ is an arbitrary index set. Define $D = \sup_{g\in\Gamma} H[g]$ and for $\lambda > 0$ define $D^\lambda = \sup_{g\in\Gamma}\{H[g] - \lambda V[g]\}$. Suppose $D^\lambda \geq 0$ and $D < \infty$. Then for $n \in \mathbb{Z}^+$ there exists $g_n \in \Gamma$ with*

$$0 \leq D^n \leq H[g_n] + 1/n \;\; \textit{and} \;\; \lim_{n\to\infty} V[g_n] = 0\,. \tag{34}$$

Proof. We have $0 \leq D^\lambda \leq D < \infty$ and so the D^λ are finite. The definition of D^λ then implies that there exists g_n with

$$D^n - 1/n \leq H[g_n] - nV[g_n] \tag{35}$$

for all n. The assumptions $V[g_n] \geq 0$ and $D^n \geq 0$ then imply $0 \leq D^n \leq H[g_n] + 1/n$. The inequality (35) also implies

$$0 \leq V[g_n] \leq H[g_n]/n - D^n/n + 1/n^2 \leq H[g_n]/n + 1/n^2 \leq D/n + 1/n^2 \to 0 \tag{36}$$

as $n \to \infty$. Therefore $\lim_{n\to\infty} V[g_n] = 0$ as claimed. □

Next we consider the implications of $V[g_n] \to 0$ when V takes the form of one of the variance penalties from Section 3.1.

Lemma 12. *Let $P \in \mathcal{P}(\Omega)$ and $h_n \in \mathcal{M}_b(\Omega)$.*

1. *If*

$$\lim_{n\to\infty} \mathrm{Var}_P[h_n] = 0 \tag{37}$$

then there exists a subsequence h_{n_j} *such that*

$$h_{n_j} - E_P[h_{n_j}] \to 0 \;\; P\text{-a.s.} \tag{38}$$

2. *If* $c \in \mathbb{R} \setminus \{0\}$ *and*

$$\lim_{n\to\infty} \mathrm{Var}_P[e^{ch_n}]/(E_P[e^{ch_n}])^2 = 0 \tag{39}$$

then there exists a subsequence h_{n_j} *such that*

$$h_{n_j} - \frac{1}{c}\log E_P[e^{ch_{n_j}}] \to 0 \;\; P\text{-a.s.} \tag{40}$$

Proof. Equation (38) is a direct consequence of Corollary 2.32 in Folland (2013). As for the second claim, we can write

$$\mathrm{Var}_P[e^{ch_n}]/(E_P[e^{ch_n}])^2 = \mathrm{Var}_P[e^{ch_n}/E_P[e^{ch_n}]] \tag{41}$$

and so the first item implies there exists a subsequence with

$$e^{ch_{n_j}}/E_P[e^{ch_{n_j}}] - E_P[e^{ch_{n_j}}/E_P[e^{ch_{n_j}}]] \to 0 \;\; P\text{-a.s.}, \tag{42}$$

i.e.,

$$e^{ch_{n_j}}/E_P[e^{ch_{n_j}}] - 1 \to 0 \;\; P\text{-a.s.} \tag{43}$$

Adding 1 to both sides, taking the logarithm, and then dividing by c gives the claimed result.

□

We are now ready to prove the following limit results as $\lambda \to \infty$.

Theorem 13. *Let* $Q, P \in \mathcal{P}(\Omega)$ *with* $Q \ll P$ *and* $\Gamma \subset \mathcal{M}_b(\Omega)$.

1. *If* $0 \in \Gamma$ *and* $D_{KL}^{\Gamma}(Q\|P) < \infty$ *then* $\lim_{\lambda\to\infty} D_{KL}^{\Gamma,\lambda}(Q\|P) = 0$.
2. *Let* $\alpha > 0$, $\alpha \neq 1$. *If* $0 \in \Gamma$ *and* $R_\alpha^{\Gamma}(Q\|P) < \infty$ *then* $\lim_{\lambda\to\infty} R_\alpha^{\Gamma,\lambda}(Q\|P) = 0$.
3. *Let* $f : \mathbb{R} \to (-\infty, \infty]$ *be lower semicontinuous, convex, finite on a neighborhood of* 1 *with* $f(1) = 0$, *and satisfy* $f^*(y) < \infty$ *for all* $y \in \mathbb{R}$. *If there exists* $c_0 \in \Gamma \cap \mathbb{R}$ *and* $D_f^{\Gamma}(Q\|P) < \infty$ *then* $\lim_{\lambda\to\infty} D_f^{\Gamma,\lambda}(Q\|P) = 0$.

Remark 14. *Note that if* $D_{KL}(Q\|P) < \infty$ *then* $D_{KL}^{\Gamma}(Q\|P) < \infty$ *and similarly for the other divergences.*

Proof.

1. Let $D = D_{\mathrm{KL}}^{\Gamma}(Q\|P)$ and $D^\lambda = D_{\mathrm{KL}}^{\Gamma,\lambda}(Q\|P)$. The assumption $0 \in \Gamma$ allows us to bound equation (14) below by the value at $g = 0$, resulting in the bound $D^\lambda \geq 0$. This, together with equation (14) and the assumption that $D < \infty$, implies that D and D^λ are of the form required by Lemma 11, hence we conclude that there exists $g_n \in \Gamma$ with

$$0 \leq D^n \leq E_Q[g_n] - \log E_P[e^{g_n}] + 1/n \tag{44}$$

and

$$\lim_{n\to\infty} \left(\mathrm{Var}_Q[g_n] + \mathrm{Var}_P[e^{g_n}]/(E_P[e^{g_n}])^2\right) = 0\,. \tag{45}$$

Note that both terms in (45) are non-negative, hence they each converge to zero individually. Combining parts 1 and 2 of Lemma 12 we therefore obtain a subsequence g_{n_j} with

$$g_{n_j} - E_Q[g_{n_j}] \to 0 \;\; Q\text{-a.s.} \tag{46}$$

and

$$g_{n_j} - \log E_P[e^{g_{n_j}}] \to 0 \;\; P\text{-a.s.} \tag{47}$$

The assumption $Q \ll P$ implies that $g_{n_j} - \log E_P[e^{g_{n_j}}] \to 0$ Q-a.s. as well. Therefore there exists $\omega \in \Omega$ with $g_{n_j}(\omega) - E_Q[g_{n_j}] \to 0$ and $g_{n_j}(\omega) - \log E_P[e^{g_{n_j}}] \to 0$. Combining these facts with equation (44) we therefore obtain

$$0 \leq D^{n_j} \leq E_Q[g_{n_j}] - g_{n_j}(\omega) + g_{n_j}(\omega) - \log E_P[e^{g_{n_j}}] + 1/n_j \to 0 \tag{48}$$

as $j \to \infty$. Therefore $\lim_{j\to\infty} D^{n_j} = 0$. D^λ is non-increasing in λ, hence we can conclude that $\lim_{\lambda\to\infty} D^\lambda = 0$ as claimed.

2. The proof in this case is very similar to the proof of claim 1; we omit the details.

3. If we let $\nu_0 = f'_+(1)$ (right derivative) then $f^*(\nu_0) = \nu_0$ (see Lemma A.9 in Birrell et al. (2020a)). Bounding equation (12) below by the value at $g = c_0$, $\nu = c_0 - \nu_0$ we see that $D_f^{\Gamma,\lambda}(Q\|P) \geq 0$. Similarly to the proof of part 1, using Lemmas 11 and 12 one can show that there exists $\omega \in \Omega$ and subsequences $g_{n_j} \in \Gamma$, $\nu_{n_j} \in \mathbb{R}$ such that

$$\begin{aligned} &g_{n_j}(\omega) - E_Q[g_{n_j}] \to 0\,, \\ &f^*(g_{n_j}(\omega) - \nu_{n_j}) - E_P[f^*(g_{n_j} - \nu_{n_j})] \to 0\,, \\ &0 \leq D_f^{\Gamma,n_j}(Q\|P) \leq \frac{1}{n_j} + E_Q[g_{n_j} - \nu_{n_j}] - E_P[f^*(g_{n_j} - \nu_{n_j})]\,. \end{aligned} \tag{49}$$

Combining these with the inequality

$$f^*(y) = \sup_{x\in\mathbb{R}}\{yx - f(x)\} \geq y - f(1) = y \tag{50}$$

we can conclude that $\lim_{j\to\infty} D_f^{\Gamma,n_j}(Q\|P) = 0$. $D_f^{\Gamma,\lambda}(Q\|P)$ is non-increasing in λ, hence this implies $\lim_{\lambda\to\infty} D_f^{\Gamma,\lambda}(Q\|P) = 0$ as claimed.

□

D BIAS BOUNDS

In this appendix we derive bounds on the bias of Rényi and f-divergence variational formula estimators. The key lemma is the following simple results regarding the expectation of a supremum or infimum.

Lemma 15. *Given an objective functional, $H : \mathcal{M}_b(\Omega) \times \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \overline{\mathbb{R}}$, and a test function space $\Gamma \subset \mathcal{M}_b(\Omega)$ we have*

$$\begin{aligned} &\mathbb{E}[\sup_{g\in\Gamma} H[g; Q_n, P_n] \geq \sup_{g\in\Gamma} \mathbb{E}[H[g; Q_n, P_n]]\,, \\ &\mathbb{E}[\inf_{g\in\Gamma} H[g; Q_n, P_n] \leq \inf_{g\in\Gamma} \mathbb{E}[H[g; Q_n, P_n]]\,. \end{aligned} \tag{51}$$

The next lemma provides a bound on the bias of statistical estimators of Λ_f^P from equation (4).

Lemma 16. *Let f be convex with $f(1) = 0$, $P \in \mathcal{P}(\Omega)$, and P_n be n-sample empirical measures from Q and P respectively. Then for all $g \in \mathcal{M}_b(\Omega)$ the generalized cumulant generating function satisfies*

$$\mathbb{E}[\Lambda_f^{P_n}[g]] \leq \Lambda_f^P[g]\,. \tag{52}$$

Proof. Using equation (4) we can compute

$$\begin{aligned}\mathbb{E}\left[\Lambda_f^{P_n}[g]\right] =& \mathbb{E}\left[\inf_{\nu\in\mathbb{R}}\{\nu + E_{P_n}[f^*(g-\nu)]\}\right] \\ \leq& \inf_{\nu\in\mathbb{R}} \mathbb{E}\left[\nu + E_{P_n}[f^*(g-\nu)]\right] \\ =& \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\} = \Lambda_f^P[g]\,.\end{aligned} \tag{53}$$

□

Lemmas 16 and 15 allow us to bound the bias of both f-divergences and Rényi divergences.

Corollary 17 (Rényi Divergence Bias Bound)**.** *For $\alpha \in (0,1)$ and $g \in \mathcal{M}_b(\Omega)$ we have*

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\alpha-1}\log E_{Q_n}[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_{P_n}[e^{\alpha g}]\right] \\ \geq& \frac{1}{\alpha-1}\log E_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_P[e^{\alpha g}]\end{aligned} \tag{54}$$

and

$$\mathbb{E}[R_\alpha^\Gamma(Q_n\|P_n)] \geq R_\alpha^\Gamma(Q\|P)\,. \tag{55}$$

Proof. To prove equation (54) we compute

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\alpha-1}\log E_{Q_n}[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_{P_n}[e^{\alpha g}]\right] \\ =& -\frac{1}{1-\alpha}\mathbb{E}\left[\Lambda^{Q_n}[(\alpha-1)g]\right] - \frac{1}{\alpha}\mathbb{E}\left[\Lambda^{P_n}[\alpha g]\right] \\ \geq& -\frac{1}{1-\alpha}\Lambda^Q[(\alpha-1)g] - \frac{1}{\alpha}\Lambda^P[\alpha g] \\ =& \frac{1}{\alpha-1}\log E_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_P[e^{\alpha g}]\,.\end{aligned} \tag{56}$$

Equation (55) then follows from Lemma 15. □

Remark 18. *When $\alpha > 1$ the biases of the two terms in equation (56) compete and so we can not obtain a bias bound via the above method.*

Similarly, we have:

Corollary 19 (f-Divergence Bias Bound)**.**

$$\mathbb{E}[E_{Q_n}[g] - \Lambda_f^{P_n}[g]] \geq E_Q[g] - \Lambda_f^P[g]$$

for all $g \in \mathcal{M}_b(\Omega)$ and

$$\mathbb{E}[D_f^\Gamma(Q_n\|P_n)] \geq D_f^\Gamma(Q\|P)\,. \tag{57}$$

E EXPERIMENTAL DETAILS

Here, we describe the experimental setup and architectural details for the experiments presented in the paper. All experiments were performed using a fully-connected, feed-forward neural network of three hidden layers. We chose the number of units per layer to be 16, 16 and 8 for the one-dimensional Gaussians synthetic example as well as for the biological data example and $\tanh(\cdot)$ as activation function for hidden and output layers. In addition, we use gradient penalty with $\lambda_{\text{gp}} = 0.1$ during training using the biological data sets, for improved stability. We apply Adam optimizer as the training algorithm, with learning rate $\lambda_{\text{lr}} = 0.0005$ for the one-dimensional experiments, while the number of iterations is $N_{\text{it}} = 20000$ and batch size is $m = 124$. Due to slower convergence, the respective values for the biological data set are $\lambda_{\text{lr}} = 0.005$ and $N_{\text{it}} = 200000$. Moreover, we set a large value to the batch size $m = 1024$, so that samples from the tails are included in the statistical average with high probability at each step. For 40-dimensional Gaussians, we used three hidden layers with 32, 32 and 16 hidden units respectively for DNE methods. The learning rate is set to 0.0001. For other methods, we chose the parameters as provided in the repository [3]. We follow the same architectures for the application of speech synthesis as mentioned in (Paul et al., 2021). Implementation is carried out using TensorFlow and code will be made available upon acceptance.

F FURTHER RESULTS ON SYNTHETIC DATA

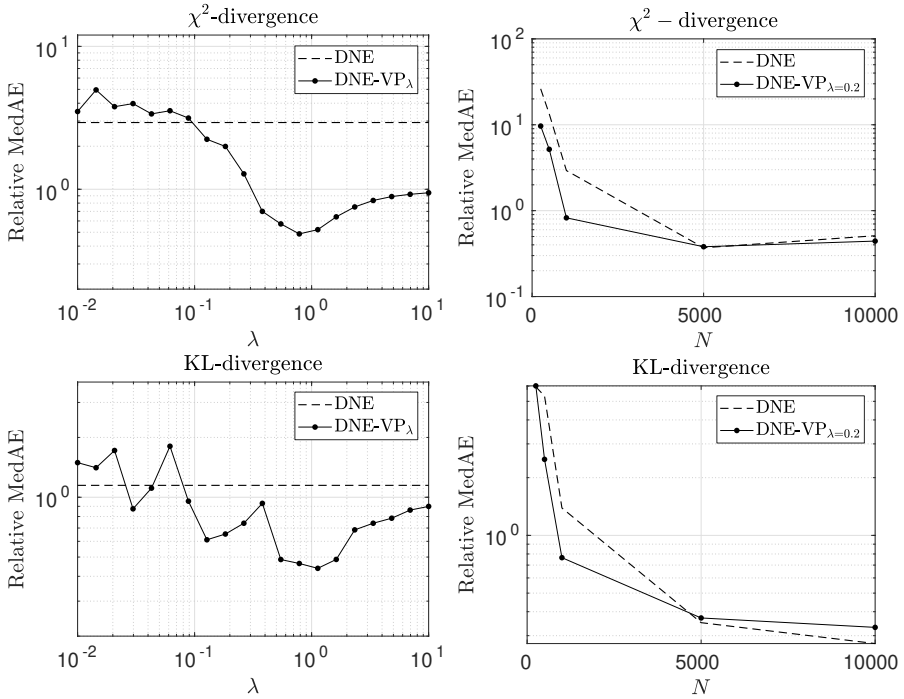

Figure 4: Comparison between the estimator without VP (DNE) and with VP (DNE-VP$_\lambda$) for f-divergence special cases of χ^2-divergence and KL-divergence, between two one-dimensional Gaussians with $Q = \mathcal{N}(0, 1.1)$ and $P = \mathcal{N}(0, 1)$. **Left column:** We compare the relative MedAE for varying penalty coefficient λ for sample size $N = 1000$, batch size 512, averaged over 50 i.i.d. runs. **Right column:** Relative MedAE for increasing sample size N similar to Figure 1.

[3] https://github.com/Linear95/CLUB

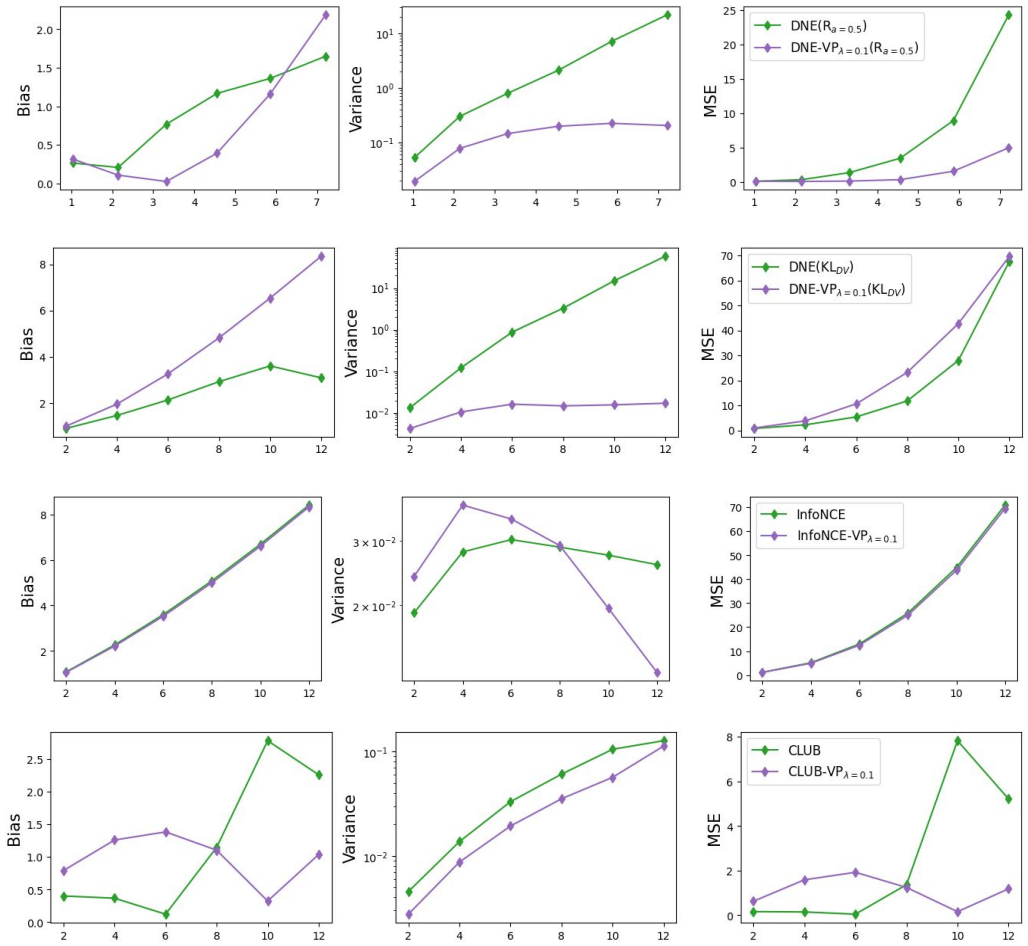

Figure 5: Estimation quality comparison of MI estimators considering bias, variance and mean square estimation error (MSE=bias^2+variance) for DNE, InfoNCE and CLUB variants. The horizontal axis shows the exact divergence values. The results are averaged over 20 i.i.d. runs. The results show improvements in the variance reduction front when VP is employed leading to lower MSE errors.

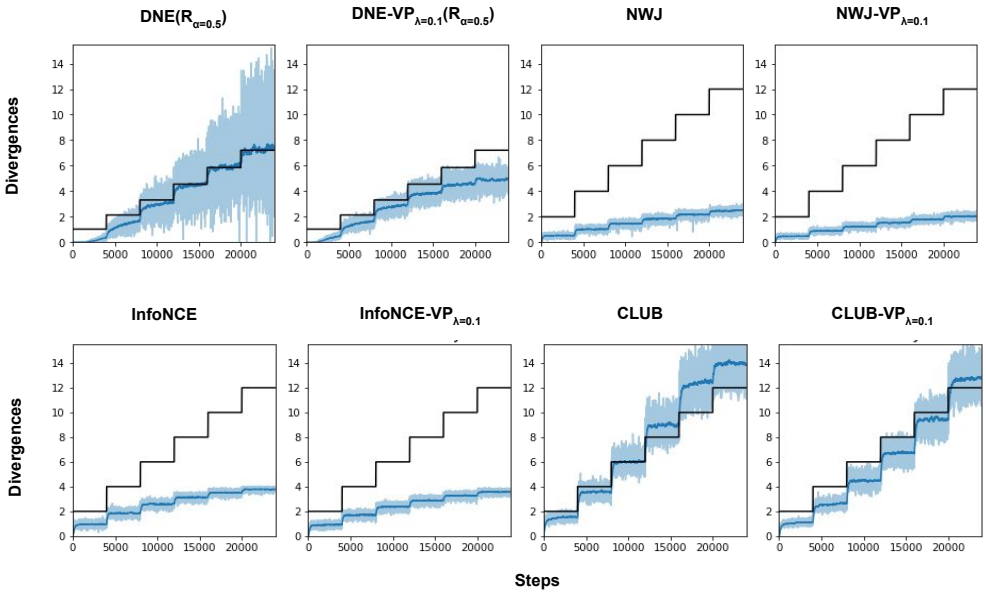

Figure 6: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Panels with $R_{\alpha=0.5}$ in their titles present the Rényi-based MI with $\alpha = 0.5$ whereas the rest of the methods estimate the standard MI (i.e., the KL divergence). Batch size is 64 and number of sample size set to 256K.

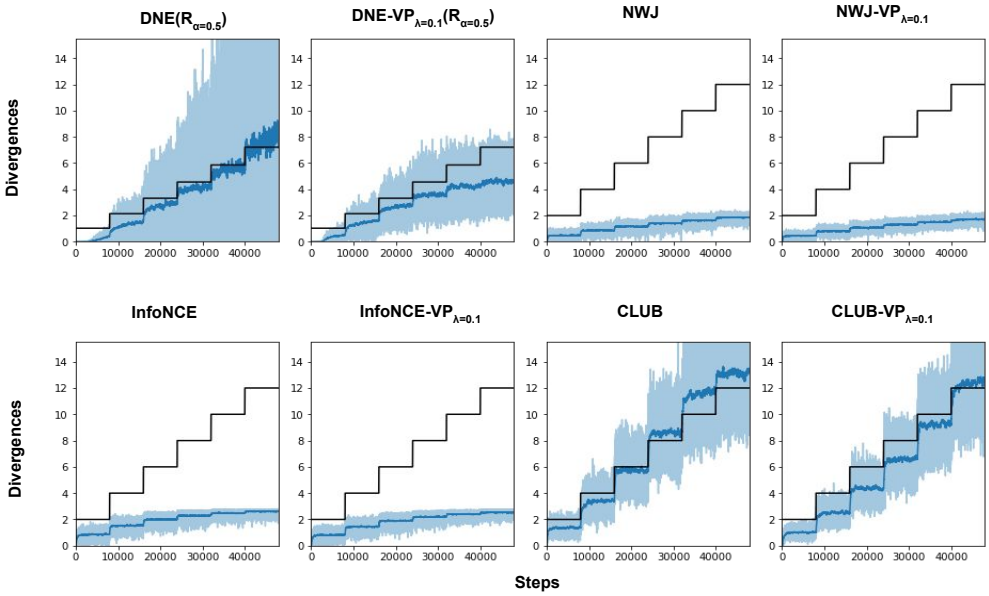

Figure 7: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 16 and number of sample size set to 512K.

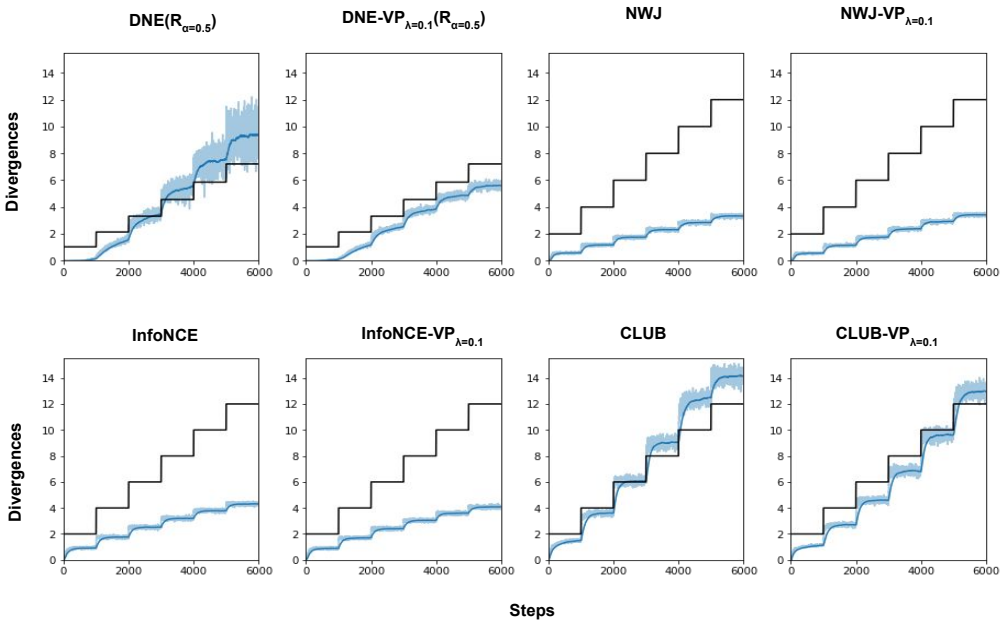

Figure 8: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 512 and number of sample size set to 256K.

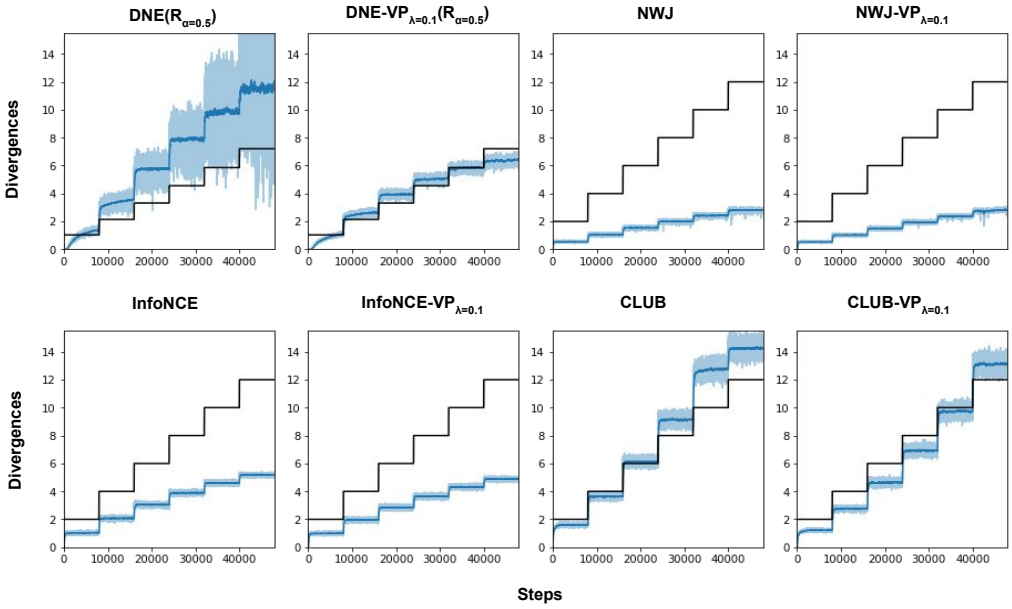

Figure 9: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 512 and number of sample size set to 4M.

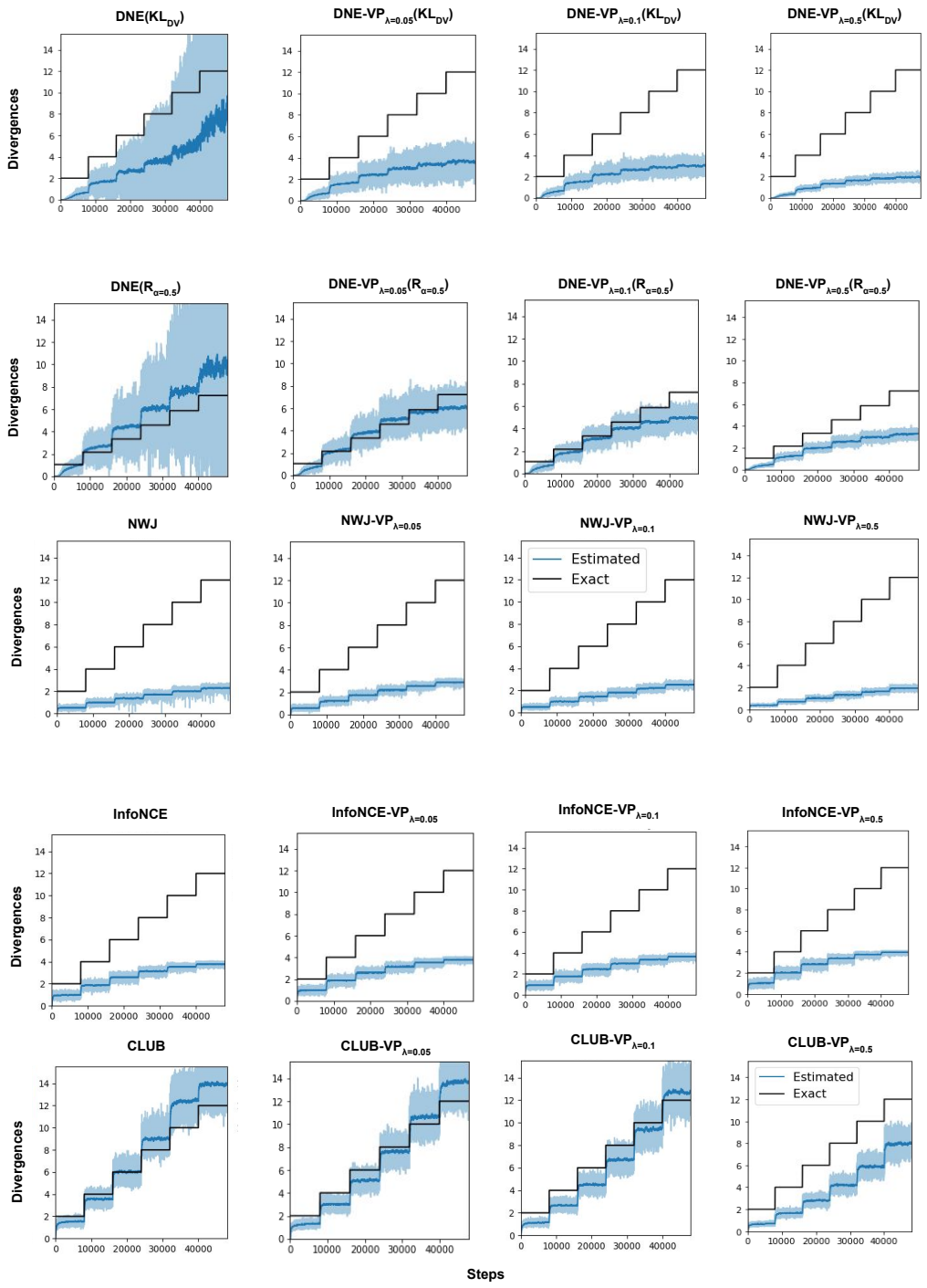

Figure 10: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector with different λ values. Batch size is 64 and number of sample size set to 512K.

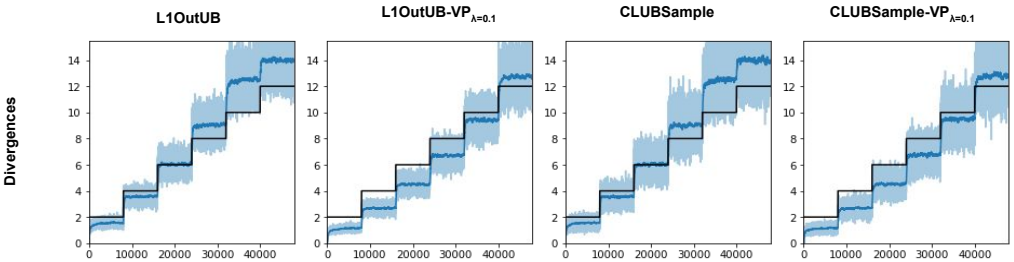

Figure 11: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 64 and number of sample size set to 512K.