Divergence Frontiers for Generative Models: Sample Complexity, Quantization Level, and Frontier Integral

Anonymous Author(s)
Affiliation
Address
email

Abstract

The spectacular success of deep generative models calls for quantitative tools to measure their statistical performance. Divergence frontiers have recently been proposed as an evaluation framework for generative models, due to their ability to measure the quality-diversity trade-off inherent to deep generative modeling. However, the statistical behavior of divergence frontiers estimated from data remains unknown to this day. In this paper, we establish non-asymptotic bounds on the sample complexity of the plug-in estimator of divergence frontiers. Along the way, we introduce a novel integral summary of divergence frontiers. We derive the corresponding non-asymptotic bounds and discuss the choice of the quantization level by balancing the two types of approximation errors arisen from its computation. We also augment the divergence frontier framework by investigating the statistical performance of smoothed distribution estimators such as the Good-Turing estimator. We illustrate the theoretical results with numerical examples from natural language processing and computer vision.

1 Introduction

Deep generative models have recently taken a giant leap forward in their ability to model complex, high-dimensional distributions. Recent advances are able to produce incredibly detailed and realistic images [26, 41, 24], strikingly consistent and coherent text [40, 57, 5], and music of near-human quality [11]. The advances in these models, particularly in the image domain, have been spurred by the development of quantitative evaluation tools which enable a large-scale comparison of models, as well as diagnosing of where and why a generative model fails [46, 30, 21, 45, 23].

Djolonga et al. [13] recently proposed *divergence frontiers* as a unified evaluation framework to quantify the quality-diversity trade-off inherent to generative modeling [45, 29, 50, 35]. In particular, a good generative model must produce high-quality samples that are likely under the target distribution but must also cover the target distribution with diverse samples. While the framework is mathematically elegant and has found empirical success, the statistical properties of divergence frontiers are not well understood. Their estimation from data for large generative models involves two approximations: (a) joint quantization of the model and target distributions into discrete distributions with level k, and (b) estimation of the frontier between the quantized distributions from samples.

Djolonga et al. [13] argue that the quantization always introduces a positive bias, making the distributions appear closer than they really are; while a small sample size can result in a pessimistic estimate of the divergence frontier. The latter effect is due to the *missing mass* of the sample—the two distributions might appear farther than they really are because the samples do not cover some parts of the distributions. The first consideration favors a large k, while the second favors a small k.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

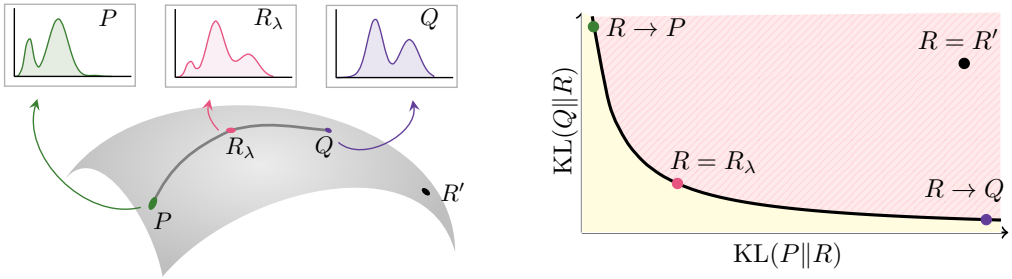

Figure 1: **Left**: Comparing two distributions P and Q. Here, $R_\lambda = \lambda P + (1-\lambda)Q$ is the interpolation between P and Q for $\lambda \in (0,1)$ and R' denotes some arbitrary distribution. **Right**: The corresponding divergence frontier (black curve) between P and Q. The interpolations R_λ for $\lambda \in (0,1)$ make up the frontier, while all other distributions such as R' must lie above the frontier.

In this paper, we are interested in answering the following questions: (a) Given two distributions, how many samples are needed to achieve a desired estimation accuracy, or in other words, what the sample complexity of the estimation procedure is; (b) Given a sample size budget, we ask how to choose the quantization level to balance the errors induced by the two approximations; (c) Can we have estimators better than the naïve plug-in estimator. The main contributions of the paper are:

(a) We establish non-asymptotic bounds for the estimation of divergence frontiers, providing upper bounds on its sample complexity.

(b) We propose a novel integral summary of divergence frontiers, derive corresponding non-asymptotic bounds and discuss the choice of the quantization level by balancing the errors induced by the two approximations.

(c) We show both theoretically and empirically that smoothed distribution estimators, such as the add-constant estimator and the Good-Turing estimator, improve the estimation accuracy.

(d) We demonstrate, through simulations on synthetic data as well as generative adversarial networks for images and transformer-based language models for text, that the bounds exhibit the correct dependence of the estimation error on the sample size n and the support size k.

As an added benefit, our analysis is transparent, and can be immediately generalized to a large class of f-divergences under some regularity assumptions. To bound the estimation error, we apply techniques from the missing mass literature to control the mass that never appears in the sample, and utilize the regularity properties of the f-divergence to bound the rest of them.

Related work. The most widely used metrics for generative models include Inception Score [46], Fréchet Inception Distance [21], and Kernel Inception Distance [3]. They summarize the performance by a single value and thus cannot distinguish different failure cases, i.e., low quality and low diversity. Motivated by this limitation, Sajjadi et al. [45] propose a metric to evaluate the quality of generative models using two separate components: precision and recall. This formulation is extended in [50] to arbitrary probability measures using a density ratio estimator, while alternative definitions based on non-parametric representations of the manifolds of the data were proposed in [29]. These notions are generalized by the divergence frontier framework of Djolonga et al. [13].

Another line of related work is the estimation of functionals of discrete distributions; see [53] for an overview. In particular, estimation of KL divergences has been studied by [7, 58, 6, 18] in the both fixed and large alphabet regimes. These results focus on the expected $\mathbf{L}_1$ and $\mathbf{L}_2$ risks and require additional assumptions on the two distributions such as boundedness of density ratio which is not needed in our results. On the practical side, there is a new line of successful work that uses deep neural networks to find data-dependent quantizations for the purpose of estimating information theoretic quantities from samples [44, 17].

Notation. Let $\mathcal{P}(\mathcal{X})$ be the space of probability distributions on some measurable space $\mathcal{X}$. For any $P, Q \in \mathcal{P}(\mathcal{X})$, let $\mathrm{KL}(P\|Q)$ be the Kullback-Leibler (KL) divergence between P and Q. For $\lambda \in (0,1)$, we define the *interpolated KL divergence* as $\mathrm{KL}_\lambda(P\|Q) := \mathrm{KL}(P\|\lambda P + (1-\lambda)Q)$. For a partition $\mathcal{S} := \{S_1, \ldots, S_k\}$ of $\mathcal{X}$, we define $P_\mathcal{S}$ the quantized version of P so that $P_\mathcal{S} \in \mathcal{P}(\mathcal{S})$ with $P_\mathcal{S}(S_i) = P(S_i)$ for any $i \in [k] := \{1, \ldots, k\}$.

2 Divergence frontiers

Divergence frontiers compare two distributions P and Q using a frontier of statistical divergences. Each point on the frontier compares the individual distributions against a mixture R_λ of the two. By sweeping through mixtures, the curve interpolates between measurements of precision and recall. Figure 1 illustrates divergence frontiers, which we formally introduce below.

Evaluating generative models via divergence frontiers. Consider a generative model $Q \in \mathcal{P}(\mathcal{X})$ which attempts to model the target distribution $P \in \mathcal{P}(\mathcal{X})$. It has been argued in [29, 45] that one must consider two types of errors to evaluate Q with respect to P: (a) a loss in precision, which is the mass of P that Q does not adequately capture, and, (b) a loss in recall, which is the mass of Q that has low or zero probability mass under P.

Suppose P, Q are uniform distributions on their supports, and R is uniform on the union of their supports. Then, the loss in precision is the mass of $\mathrm{Supp}(Q) \setminus \mathrm{Supp}(P)$, or equivalently, the mass of $\mathrm{Supp}(R) \setminus \mathrm{Supp}(P)$. We measure this using the surrogate $\mathrm{KL}(Q\|R)$, which is large if there exists a such that $Q(a)$ is large but $R(a)$ is small. Likewise, the loss in recall is measured by $\mathrm{KL}(P\|R)$. When P and Q are not constrained to be uniform, it is not clear what the measure R should be. Djolonga et al. [13] propose to vary R over all possible probability measures and consider the Pareto frontier of the multi-objective optimization $\min_R \big(\mathrm{KL}(P\|R), \mathrm{KL}(Q\|R)\big)$. This leads to a curve called the *divergence frontier*, and is reminiscent of the precision-recall curve in binary classification.

Formally, it can be shown that the divergence frontier $\mathcal{F}(P, Q)$ of probability measures P, Q is carved out by mixtures $R_\lambda = \lambda P + (1-\lambda)Q$ for $\lambda \in (0, 1)$ (cf. Figure 1). It admits the closed-form

$$\mathcal{F}(P, Q) = \Big\{ \big(\mathrm{KL}_\lambda(P\|Q),\, \mathrm{KL}_{1-\lambda}(Q\|P)\big) \,:\, \lambda \in (0,1) \Big\},$$

where $\mathrm{KL}_\lambda(P\|Q) = \mathrm{KL}(P\|\lambda P + (1-\lambda)Q)$ is the interpolated KL divergence.

Practical computation of divergence frontiers. In practical applications, P is a complex, high-dimensional distribution which could either be discrete, as in natural language processing, or continuous, as in computer vision. Likewise, Q is often a deep generative model such as GPT-3 for text and GANs for images. It is infeasible to compute the divergence frontier $\mathcal{F}(P, Q)$ directly because we only have samples from P and the integrals or sums over Q are intractable.

Therefore, the recipe used by practitioners [45, 13] has been to (a) jointly quantize P and Q over a partition $\mathcal{S} = \{S_1, \cdots, S_k\}$ of $\mathcal{X}$ to obtain discrete distributions $P_\mathcal{S} = (P(S_1), \cdots, P(S_k))$ and $Q_\mathcal{S} = (Q(S_1), \cdots, Q(S_k))$, (b) estimate the quantized distributions from samples to get $\hat{P}_\mathcal{S}$, $\hat{Q}_\mathcal{S}$, and (c) compute $\mathcal{F}(\hat{P}_\mathcal{S}, \hat{Q}_\mathcal{S})$. In practice, the best quantization schemes are data-dependent transformations such as k-means clustering or lattice-type quantization of dense representations of images or text [44].

Scalar summary of divergence frontiers. Since it can be cumbersome to compare divergence frontiers $\mathcal{F}(P, Q_j)$ for a number of models $Q_1, Q_2, \cdots$, it is beneficial to summarize the frontier with a scalar. It can also be used to quantify the convergence of the divergence frontier over all λ.

We define a novel integral summary, called the *frontier integral* $\mathrm{FI}(P, Q)$ of two distributions P, Q as

$$\mathrm{FI}(P, Q) := 2 \int_0^1 \big(\lambda\, \mathrm{KL}_\lambda(P\|Q) + (1-\lambda)\mathrm{KL}_{1-\lambda}(Q\|P)\big)\, \mathrm{d}\lambda\,. \tag{1}$$

We can interpret the frontier integral as the average value of the scalarized objective which defines each point on the divergence frontier.

While the length of the divergence frontier could be unbounded (e.g., when $\mathrm{KL}(P\|Q)$ is unbounded), the frontier integral is always bounded in $[0, 1]$. Moreover, it is a symmetric divergence with $\mathrm{FI}(P, Q) = 0$ if and only if $P = Q$. In practice, the frontier integral can be estimated using the same recipe as the divergence frontier.

Error decomposition. In Section 3, we decompose the error in estimating the frontier integral into two components: the statistical error of estimating the quantized distribution and the quantization error. Our goal is to derive the rate of convergence for the overall estimation error. To control the statistical error, we use a different treatment for the masses that appear in the sample and the ones that never appear (i.e., the missing mass). We obtain a high probability bound as well as a bound for its

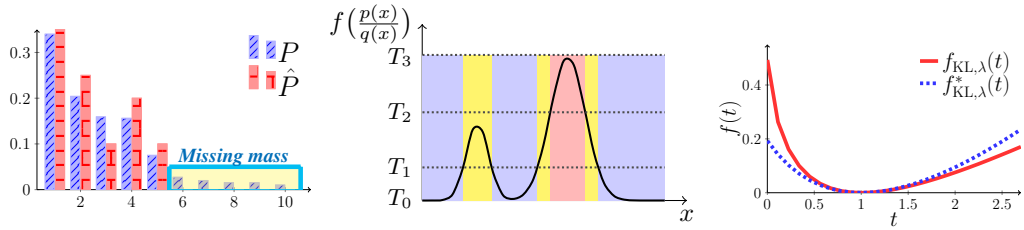

Figure 2: **Left**: The missing mass in a sample. **Middle**: Oracle quantization into 3 bins: blue, yellow and red. Bin i is given by the set $\{x : f(p(x)/q(x)) \in [T_{i-1}, T_i)\}$. **Right**: The generator and conjugate generator of KL_λ at $\lambda = 1/2$.

expectation, leading to upper bounds for its sample complexity and rate of convergence. These results carry over to the divergence frontiers as well. As for the quantization error, we construct a distribution-dependent quantization scheme whose error is at most $O(k^{-1})$, where k is the quantization level. A combination of these two bounds sheds light on the optimal choice of the quantization level. In Section 5, we verify empirically the tightness of the rates on synthetic and real data.

3 Main results

Let $P, Q \in \mathcal{P}(\mathcal{X})$. Let $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ be independent i.i.d. samples from P and Q, respectively. Denote by $\hat{P}_n$ and $\hat{Q}_n$ the respective empirical measures of P and Q. Note that the two samples are assumed to have the same size for simplicity. In this section, we summarize our main theoretical results. We denote by C an absolute constant which can vary from line to line. The precise statements and proofs can be found in the Appendix.

Sample complexity for the frontier integral. We are interested in deriving a non-asymptotic bound for the absolute error of the plug-in estimator, i.e., $\left|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)\right|$. When both P and Q are supported on a finite alphabet with k items, a natural strategy is to exploit the smoothness properties of FI, giving a naïve upper bound $O(L\sqrt{k/n})$ on the absolute error, where $L = \log 1/p_*$ with $p_* = \min_{a \in \mathrm{Supp}(P)} P(a)$ reflecting the smoothness of FI. The dependency on p_* requires P to have a finite support and a short tail. However, in many real-world applications, the distributions can either be supported on a countable set or have long tails [9, 55]. By considering the *missing mass* in the sample, we are able to obtain a high probability bound that is independent of p_*.

Theorem 1. *Assume that P and Q are discrete and let $k = \max\{|\mathrm{Supp}(P)|, |\mathrm{Supp}(Q)|\} \in \mathbb{N} \cup \{\infty\}$. For any $\delta \in (0,1)$, it holds that, with probability at least $1-\delta$,*

$$\left|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)\right| \leq C \left[\sqrt{\frac{\log 1/\delta}{n}} + (\alpha_n(P) + \alpha_n(Q)) \log n + \beta_n(P) + \beta_n(Q)\right], \quad (2)$$

where $\alpha_n(P) = \sum_{a \in \mathcal{X}} \sqrt{n^{-1} P(a)}$ and $\beta_n(P) = \mathbb{E}\left[\sum_{a : \hat{P}_n(a) = 0} P(a) \max\{1, \log(1/P(a))\}\right]$. Furthermore, if the support size $k < \infty$, then $\alpha_n(P) \leq \sqrt{k/n}$ and $\beta_n(P) \leq k \log n / n$. In particular, with probability at least $1-\delta$,

$$\left|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)\right| \leq C \left[\sqrt{\frac{\log 1/\delta}{n}} + \sqrt{\frac{k}{n}} + \frac{k}{n}\right] \log n\,. \quad (3)$$

Before we discuss the bounds in Theorem 1, let us introduce the missing mass problem. This problem was first studied by Good and Turing [15], where the eponymous Good-Turing estimator was proposed to estimate the probability that a new observation drawn from a fixed distribution has never appeared before, in other words, is missing in the current sample; see Figure 2 (left) for an illustration. The Good-Turing estimator has been widely used in language modeling [25, 10, 9] and studied in theory [33, 39, 38]. An inspiring result coming from this line of work is that the missing mass in a sample of size n concentrates around its expectation [32], which itself decays as $O(k/n)$ when the distribution is supported on k items [2].

There are several merits to Theorem 1. First, (2) holds for any distributions with a countable support. Second, it does not depend on p_* and is adapted to the tail behavior of P and Q. For instance, if P is

defined as $P(a) \propto a^{-2}$ for $a \in [k]$, then $\alpha_n(P) \propto (\log k)/\sqrt{n}$, which is much better than $\sqrt{k/n}$ in (3) in terms of the dependency on k. This phenomenon is also demonstrated empirically in Section 5. Third, it captures a parametric rate of convergence, i.e., $O(n^{-1/2})$, up to a logarithmic factor. In fact, this rate is not improvable in a related problem of estimating $\mathrm{KL}(P\|Q)$, even with the assumption that P/Q is bounded [6]. The bound in (3) is a distribution-free bound, assuming k is finite. Note that it also gives an upper bound on the sample complexity by setting the right hand side of (4) to be ϵ and solve for n, this is roughly $O((\sqrt{\log 1/\delta} + \sqrt{k})^2/\epsilon^2)$.

The proof of Theorem 1 relies on two new results: (a) a concentration bound around $\mathbb{E}[\mathrm{FI}(\hat{P}_n, \hat{Q}_n)]$, which can be obtained by McDiarmid's inequality, and (b) an upper bound for the statistical error, i.e., $\mathbb{E}\left|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)\right|$, which is upper bounded by

$$O\left([\alpha_n(P) + \alpha_n(Q)] \log n + \beta_n(P) + \beta_n(Q)\right) \leq O\left(\sqrt{k/n} \log n\right). \tag{4}$$

The concentration bound gives the term $\sqrt{n^{-1} \log 1/\delta}$. The statistical error bound is achieved by splitting the masses of P and Q into two parts: one that appears in the sample and one that never appears. The first part can be controlled by a Lipschitz-like property of the frontier integral, leading to the term $\alpha_n(P) + \alpha_n(Q)$, and the second part, $\beta_n(P) + \beta_n(Q)$, falls into the missing mass framework. In addition, the rate k/n for β_n we show here matches the rate for the missing mass.

Statistical consistency of the divergence frontiers. While Theorem 1 establishes the consistency of the frontier integral, it is also of great interest to know whether the divergence frontier itself can be consistently estimated. In fact, a similar bound holds for $\mathrm{KL}_\lambda(\hat{P}_n\|\hat{Q}_n)$, with the rate $O(\lambda^{-1}\sqrt{k/n}\,\log n)$ for each $\lambda \in (0, 1)$. Furthermore, if we truncate the frontier within the interval $[\lambda_n, 1 - \lambda_n]$ for some $\lambda_n \to 0$, then the truncated empirical frontier converges uniformly to its population counterpart at rate $O(\lambda_n^{-1}\sqrt{k/n}\,\log n)$. The truncation is necessary without imposing additional assumptions, since KL_λ is close to KL for small λ and it is known that the minimax quadratic risk of estimating the KL divergence over all distributions with k bins is always infinity [6].

Upper bound for the quantization error. Recall from Section 2 that computing the divergence frontiers in practice usually involves a quantization step. Since every quantization will inherently introduce a bias in the estimation procedure, it is desirable to control the error, which we call the quantization error, induced by this step. We show that there exists a quantization scheme with error proportional to the inverse of its size. We implement this scheme and empirically verify this rate in section 5; certain regimes appear to show even faster convergence.

Let $\mathcal{X}$ be an *arbitrary* measurable space and $\mathcal{S}$ be a partition of $\mathcal{X}$. The quantization error of $\mathcal{S}$ is the difference $|\mathrm{FI}(P_\mathcal{S}, Q_\mathcal{S}) - \mathrm{FI}(P, Q)|$. It can be shown that there exists a distribution-dependent partition $\mathcal{S}_k$ with level $|\mathcal{S}_k| = k$ whose quantization error is no larger than the inverse of its level, i.e.,

$$|\mathrm{FI}(P, Q) - \mathrm{FI}(P_{\mathcal{S}_k}, Q_{\mathcal{S}_k})| \leq Ck^{-1}. \tag{5}$$

The key idea behind the construction of this partition is visualized in Figure 2 (middle). Combining this bound with the bounds in (4) leads to the following bound for the overall estimation error.

Theorem 2. *Assume that $\mathcal{S}_k$ is a partition of $\mathcal{X}$ such that $|\mathcal{S}_k| = k \geq 2$ and its quantization error satisfies the bound in* (5)*. Then*

$$\begin{aligned}\mathbb{E}\left|\mathrm{FI}(\hat{P}_{\mathcal{S}_k,n}, \hat{Q}_{\mathcal{S}_k,n}) - \mathrm{FI}(P, Q)\right| &\leq C\left[(\alpha_n(P) + \alpha_n(Q)) \log n + \beta_n(P) + \beta_n(Q) + k^{-1}\right] \\ &\leq C\left[\left(\sqrt{\frac{k}{n}} + \frac{k}{n}\right) \log n + \frac{1}{k}\right].\end{aligned} \tag{6}$$

Based on the bound in (6), a good choice of k is $\Theta(n^{1/3})$ which balances between the two types of errors. We illustrate in Section 5 that our experimental results approximately conform to the scaling observed from Theorem 2 by investigating the two errors separately. This balancing is enabled by the existence of a good quantizer with a distribution-free bound in (5). In practice, this suggests a data-dependent quantizer using nonparametric density estimators. However, directions such as kernel density estimators [42, 20, 52] and nearest-neighbor methods [1] have not met with empirical success, as they suffer from the curse of dimensionality common in nonparametric estimators. In particular, [54, 48, 49] propose quantized divergence estimators but only prove asymptotic consistency, and

little progress has been made since then. On the other hand, modern data-dependent quantization techniques based on deep neural networks can successfully estimate properties of the density from high dimensional data [44, 17]. Theoretical results for those techniques could complement our analysis. We leverage these powerful methods to scale our approach on real data in Section 5.2.

4 Towards better estimators and interpolated f-divergences

Smoothed distribution estimators. When the support size k is large, the empirical measure performs poorly in practice. To overcome this challenge, practitioners often use more sophisticated distribution estimators such as the Good-Turing estimator [15, 38] and add-constant estimators [27, 4]. We focus on the add-constant estimator defined below and state here its estimation error when it is applied to estimate the frontier integral from data. We investigate and compare the performance of other distribution estimators in Section 5.

For notational simplicity, we assume that P and Q are supported on a common finite alphabet with size $k < \infty$. Note that this is true for the quantized distributions $P_{\mathcal{S}}$ and $Q_{\mathcal{S}}$. For any constant $b > 0$, the add-constant estimator of P is defined as $\hat{P}_{n,b}(a) = (N_a + b)/(n + kb)$ for each $a \in \mathrm{Supp}(P)$, where $N_a = |\{i : X_i = a\}|$ is the number of times a appears in the sample. The add-constant estimator of Q is defined analogously.

Thanks to the smoothing, there is no mass missing in the add-constant estimator. As a result, we can directly utilize the smoothness properties of the frontier integral to get the following bound.

Proposition 3. *Under the same assumptions as in Theorem 2, we have*

$$\begin{aligned}
&\mathbb{E}\left|\mathrm{FI}(\hat{P}_{\mathcal{S}_k,n,b}, \hat{Q}_{\mathcal{S}_k,n,b}) - \mathrm{FI}(P,Q)\right| \\
&\leq C\left[\left(\frac{n(\alpha_n(P)+\alpha_n(Q))}{n+bk} + \gamma_{n,k}(P) + \gamma_{n,k}(Q)\right)\log\left(n/b+k\right) + \frac{1}{k}\right] \\
&\leq C\left[\frac{\sqrt{nk}+bk}{n+bk}\log\left(n/b+k\right) + \frac{1}{k}\right], \qquad (7)
\end{aligned}$$

where $\gamma_{n,k}(P) = (n+bk)^{-1}bk\sum_{a\in\mathcal{X}}|P(a) - 1/k|$.

Let us compare the bounds in Proposition 3 with the ones in Theorem 2. For the distribution-dependent bound, the term $\alpha_n(P)\log n$ in (6) is improved by a factor $n/(n+bk)$ in (7). The missing mass term $\beta_n(P)$ is replaced by the total variation distance between P and the uniform distribution on $[k]$ with a factor $bk/(n+bk)$. The improvements in both two terms are most significant when k/n is large. As for the distribution-free bound, when k/n is small, the bound in (7) scales the same as the one in (6); when k/n is large (i.e., bounded away from 0 or diverging), it scales as $O(\log n + \log(k/n) + k^{-1})$ while the one in (6) scales as $O(k\log n/n + k^{-1})$. Given the improvement, it would be an interesting venue for future work to consider adaptive estimators in the spirit of [14].

Generalization to f-divergences. Estimation of the χ^2 divergence is useful for variational inference [12] and GAN training [31, 51]. More generally, estimating f-divergences from samples is a fundamental problem in machine learning and statistics [36, 22, 8, 43]. We extend our previous results to estimating general f-divergences (which satisfy some regularity assumptions) using the same two-step procedure of quantization and estimation of multinomial distributions from samples.

We start by reviewing the definition of f-divergences; see Appendix A for more details. Let $f : (0,\infty) \to \mathbb{R}$ be a nonnegative and convex function with $f(1) = 0$. Let $P, Q \in \mathcal{P}(\mathcal{X})$ be dominated by some measure $\mu \in \mathcal{P}(\mathcal{X})$ with densities p and q, respectively. The f-divergence generated by f is defined as

$$D_f(P\|Q) = \int_{\mathcal{X}} q(x) f\left(\frac{p(x)}{q(x)}\right)\mathrm{d}\mu(x)\,,$$

with the convention that $f(0) = f(0^+)$ and $0f(p/0) = pf^*(0)$, where $f^*(0) = f^*(0^+) \in [0,\infty]$ for $f^*(t) = tf(1/t)$. We call f^* the conjugate generator to f. Note that the conjugacy here is unrelated to the convex conjugacy but is based on the *perspective transform*. The function f^* also generates an f-divergence, which is referred to as the *conjugate divergence* to D_f since $D_{f^*}(P\|Q) = D_f(Q\|P)$.

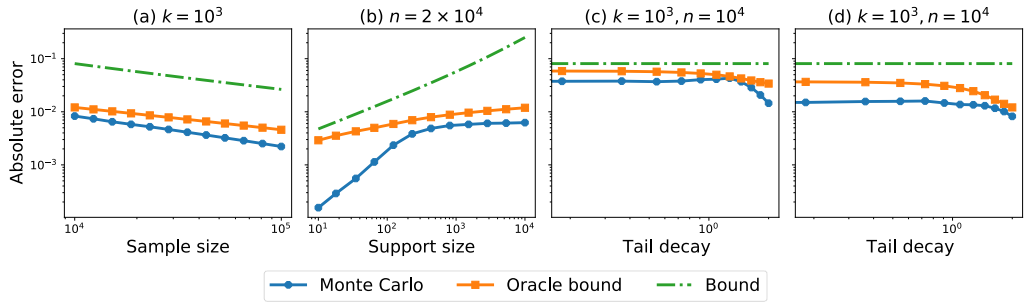

Figure 3: Statistical error bounds (4) and the Monte Carlo estimate (log-log scale). **(a)**: $\mathrm{Zipf}(2)$ and $\mathrm{Zipf}(2)$ with $k = 10^3$; **(b)**: $\mathrm{Zipf}(2)$ and $\mathrm{Zipf}(2)$ with $n = 2 \times 10^4$; **(c)**: $\mathrm{Dir}(\mathbf{1})$ and $\mathrm{Zipf}(r)$ with $k = 10^3$ and $n = 10^4$; **(d)**: $\mathrm{Zipf}(2)$ and $\mathrm{Zipf}(r)$ with $k = 10^3$ and $n = 10^4$. The bounds are scaled by 100.

The quantization error bound (5) holds for all f-divergences which are *bounded*, i.e., $f(0) + f^*(0) < \infty$. The high probability bounds in Theorem 1 also hold for f-divergences, under some regularity assumptions. The interpolated χ^2 divergence, defined analogously as the interpolated KL divergence, satisfies these conditions. In addition to the boundedness assumption, we assume $|f'(t)| \propto \log t^{-1}$ and $|(f^*)'(t)| \propto \log t^{-1}$ for small t. This guarantees that f is approximately Lipschitz and cannot vary too fast. We also require $tf''(t)$ and $t(f^*)''(t)$ to be bounded. This is a technical assumption that helps control the variation of f around zero.

Both the interpolated KL divergence and the frontier integral are f-divergences satisfying the regularity conditions. An illustration of the generator to $\mathrm{KL}_{1/2}$ can be found in Figure 2 (right). Hence, Theorem 1 and Theorem 2 are special cases of the corresponding results for f-divergences.

5 Experiments

We investigate the empirical behavior of the frontier integral on both synthetic and real data. Our main findings are: 1) the statistical error bound is tight—it approximately reveals the rate of convergence of the plug-in estimator. 2) The smoothed distribution estimators improve the estimation accuracy. We summarize our experimental settings and findings here. More details and additional results are deferred to Appendix G.

5.1 Synthetic data

Settings. We focus on the case when $k = |\mathcal{X}| < \infty$. Following the experimental settings in [38], we consider three types of distributions: 1) the $\mathrm{Zipf}(r)$ distribution with $r \in \{0, 1, 2\}$ where $P(i) \propto i^{-r}$. Note that $\mathrm{Zipf}(r)$ is regularly varying with index $-r$; see, e.g., [47, Appendix B]. 2) the Step distribution where $P(i) = 1/2$ for the first half bins and $P(i) = 3/2$ for the second half bins. 3) the Dirichlet distribution $\mathrm{Dir}(\alpha)$ with $\alpha \in \{\mathbf{1}/2, \mathbf{1}\}$. In total, there are 6 different distributions. Since the frontier integral is symmetric, there are 21 different pairs of (P, Q). For each pair (P, Q), we generate i.i.d. samples of size n from each of them, and then compute the absolute error $|\mathrm{FI}(\hat{P}_n, \hat{Q}_n) - \mathrm{FI}(P, Q)|$. We repeat the process 100 times and report its mean and standard error, which is referred to as the Monte Carlo estimate of the expected absolute error.

Statistical error. To study the tightness of the statistical error bounds (4), we compare both the distribution-free bound ("bound") and the distribution-dependent bound ("oracle bound") with the Monte Carlo estimate. We consider three different experiments. First, we fix the support size $k = 10^3$ and increase the sample size n from 10^3 to 10^4. Second, we fix $n = 2 \times 10^4$ and increase k from 10 to 10^4. Third, we fix $k = 10^3$ and $n = 10^4$, and set Q to be the $\mathrm{Zipf}(r)$ with r ranging from 0 to 2. We report four plots in Figure 3; see Appendix G for more results. Note that the two bounds are divided by the same constant 100 for the sake of comparison.

As shown in Figure 3(a), the two bounds and the Monte Carlo estimate exhibit a near-identical rate in n. In some cases, the Monte Carlo estimate demonstrates a faster rate of convergence than the bounds. This suggests that the bounds capture the worst-case behavior of the plug-in estimator. In the second experiment, we observe in Figure 3(b) that the oracle bound exhibits a slower rate than

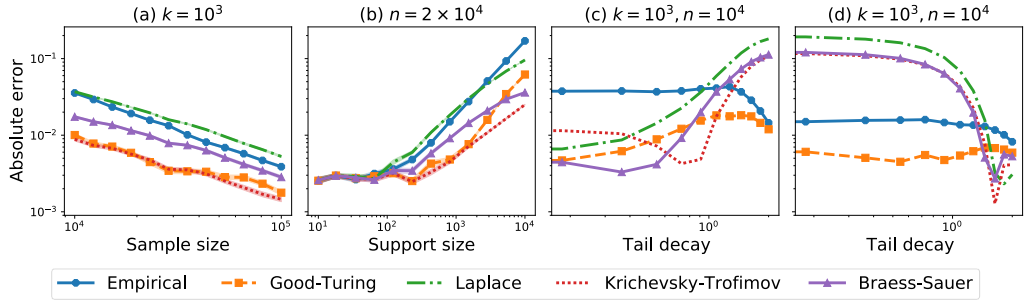

Figure 4: Absolute error for different distribution estimators (log-log scale). **(a)**: $\mathrm{Zipf}(0)$ and $\mathrm{Dir}(\mathbf{1}/2)$ with $k = 10^3$; **(b)**: $\mathrm{Zipf}(0)$ and $\mathrm{Dir}(\mathbf{1}/2)$ with $n = 2 \times 10^4$; **(c)**: $\mathrm{Dir}(\mathbf{1})$ and $\mathrm{Zipf}(r)$ with $k = 10^3$ and $n = 10^4$; **(d)**: $\mathrm{Zipf}(2)$ and $\mathrm{Zipf}(r)$ with $k = 10^3$ and $n = 10^4$.

the bound in k. The Monte Carlo estimate has a near identical rate as the bound for small k and a near-identical rate as the oracle bound for large k. Figures 3(c) and 3(d) show that, while the bound remains the same for different tails of Q, the oracle bound is adapted to the decaying index of Q and captures well the behavior of the Monte Carlo estimate.

Distribution estimators. We then compare 4 distribution estimators with the empirical measures ("Empirical") as discussed in [38]—the *(modified) Good-Turing* estimator and three add-constant estimators, namely, the *Laplace* estimator, the *Krichevsky-Trofimov* estimator and the *Braess-Sauer* estimator. We consider the same three experiments as for the statistical error. As shown in Figures 4(a) and 4(b), the rates in n and k are similar across all estimators, and the Krichevsky-Trofimov estimator and the Good-Turing estimator outperform others. Furthermore, we observe from additional results in Appendix G that when two distributions are the same, all estimators performs similarly with the Good-Turing estimator being the worst. Next, we observe in Figures 4(c) and 4(d) that the add-constant estimators perform worse than the empirical one for some tail-decaying regimes; while the Good-Turing estimator is relatively robust to the tail decay in Q.

Quantization error. We consider three quantization strategies: 1) the uniform quantization which partitions the distributions into equally spaced bins; 2) the greedy quantization which sorts the bins according to the ratios $\{P(a)/Q(a)\}_{a \in \mathcal{X}}$ and then add one bin at a time so that the frontier integral is maximized; 3) the oracle quantization we used to prove (5); see Figure 2. Due to space limit, we defer the results to Appendix G. We observe that when both P and Q have slow-decaying tails, the absolute error of the oracle quantization decays faster than $O(k^{-1})$; when one of them has fast-decaying tail, its absolute error decays slower than $O(k^{-1})$ in the beginning and then faster than $O(k^{-1})$. In our experiments, the oracle quantization always outperforms the greedy one. When either P or Q is not ordered, the uniform quantization performs the worst. When both P and Q are ordered, its absolute error is quite small in the beginning and then becomes larger.

5.2 Real Data

Tasks and Datasets. We consider two domains: images and text. For the image domain, we train a generative model for the CIFAR-10 dataset [28] based on StyleGAN2 [24]. We use the publicly available code[1] with their default hyperparameters. To evaluate the frontier integral, we use the test set of 10k images as the target distribution P and we sample 10k images from the generative model as the model distribution Q. For the text domain, we fine-tune a pretrained GPT-2 [40] model with 124M parameters (i.e., GPT-2 small) on the Wikitext-103 dataset [34]. We use the open-source HuggingFace Transformers library [56] for training, and generate 10k 500-token completions using top-p sampling and 100-token prefixes.

Settings. We take the following steps to compute the frontier integral. First, we represent each image/text by its features [21, 45, 29]. Second, we learn a low-dimensional feature embedding which maintains the neighborhood structure of the data while encouraging the features to be uniformly distributed on the unit sphere [44]. Third, we quantize these embeddings on a uniform lattice with k bins. For each support size k, this gives us quantized distributions $P_{\mathcal{S}_k}, Q_{\mathcal{S}_k}$. Finally, we sample n

[1] `https://github.com/NVlabs/stylegan2-ada-pytorch`

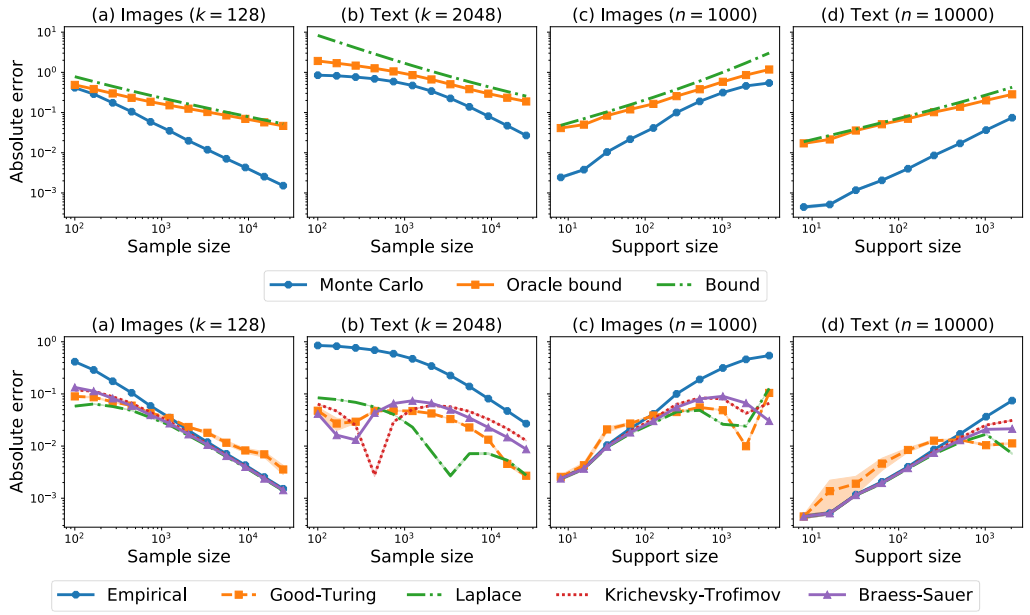

Figure 5: Absolute error for the statistical error bounds (**top**) and different distribution estimators (**bottom**) (log-log scale). **(a)**: Image data (CIFAR-10) with $k = 128$; **(b)**: Text data (WikiText-103) with $k = 2048$; **(c)**: Image data (CIFAR-10) with $n = 1000$; **(d)**: Text data (WikiText-103) with $n = 10000$. The bounds are scaled by 30.

i.i.d observations each from these distributions and consider the empirical distributions $\hat{P}_{\mathcal{S}_k,n}, \hat{Q}_{\mathcal{S}_k,n}$ as well as the add-constant and Good-Turing estimators computed from these samples.

Statistical error. Following the corresponding part on synthetic data, we consider two experiments. First, we fix the support size k and vary the sample size n from 100 to 25000. Second, we fix the sample size n and vary the support size k from 8 to 2048 in powers of 2. The findings are similar—the bounds capture the worst-case behavior in n of real image and text data, and, in k, the Monte Carlo estimate can grows either at a similar rate or faster than the bounds depending on the value of n.

Distribution Estimators. As in the previous section, we compare the empirical estimator with four distribution estimators. From Figure 5 (bottom), we observe similar rates (i.e., similar slopes) for all estimators with respect to the sample size n when $n > k$. The absolute error of the Good-Turing estimator is the worst among all estimators considered for $k = 128$ and n large. However, for $k = 2048$, the empirical estimator is the worst. The various add-b estimators work the best in the regime of $n < k$, where each add-b estimator attains the smallest error at a different n. In particular, the Laplace estimator is the best or close to the best in all each of the settings considered.

6 Conclusion

In this paper, we study the statistical behavior of the divergence frontiers and the proposed integral summary estimated from data. We decompose the estimation error into two components, the statistical error and the quantization error, to conform with the approximation procedure commonly used in practice. We establish non-asymptotic bounds on both of the two errors. Our bounds shed light on the optimal choice of the quantization level k—they suggests that the two errors can be balanced at $k = \Theta(n^{1/3})$. We also derive a new concentration inequality for the frontier integral, which provides the sample complexity of achieving a small error in high probability. Finally, we demonstrate both theoretically and empirically that the use of smoothed distribution estimators can improve the estimation accuracy. All the results can be generalized to a large class of interpolation-based f-divergences. Provided theoretical results on modern data-dependent quantization schemes using deep neural networks, it would be an interesting venue to specialize our bounds to such quantization schemes. Extending our results to β-divergences could also be interesting.

References

[1] M. Alamgir, G. Lugosi, and U. von Luxburg. Density-preserving quantization with application to graph downsampling. In *Conference on Learning Theory*, pages 543–559. PMLR, 2014.

[2] D. Berend and A. Kontorovich. The missing mass problem. *Statistics & Probability Letters*, 82 (6):1102–1110, 2012.

[3] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

[4] D. Braess and T. Sauer. Bernstein polynomials and learning theory. *J. Approx. Theory*, 128(2): 187–206, 2004.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

[6] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli. Estimation of KL divergence: Optimal minimax rate. *IEEE Trans. Inf. Theory*, 64(4):2648–2674, 2018.

[7] H. Cai, S. R. Kulkarni, and S. Verdú. Universal divergence estimation for finite-alphabet sources. *IEEE Trans. Inf. Theory*, 52(8):3456–3475, 2006.

[8] L. Chen, C. Tao, R. Zhang, R. Henao, and L. Carin. Variational Inference and Model Selection with Generalized Evidence Bounds. In *ICML*, volume 80, pages 892–901, 2018.

[9] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.*, 13(4):359–393, 1999.

[10] K. W. Church and W. A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5:19–54, 1991.

[11] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A Generative Model for Music. *arXiv Preprint*, 2020.

[12] A. B. Dieng, D. Tran, R. Ranganath, J. W. Paisley, and D. M. Blei. Variational Inference via χ Upper Bound Minimization. In *NeurIPS 2017*, pages 2732–2741, 2017.

[13] J. Djolonga, M. Lucic, M. Cuturi, O. Bachem, O. Bousquet, and S. Gelly. Precision-Recall Curves Using Information Divergence Frontiers. In *AISTATS*, pages 2550–2559, 2020.

[14] A. Goldenshluger and O. Lepski. Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities. *Probability Theory and Related Fields*, 143:41–71, 2009.

[15] I. J. Good. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, 40(3-4):237–264, 1953.

[16] L. Györfi and T. Nemetz. f-dissimilarity: A generalization of the affinity of several distributions. *Annals of the Institute of Statistical Mathematics*, 30:105–113, 1978.

[17] P. Hämäläinen, T. Saloheimo, and A. Solin. Deep residual mixture models. *arXiv preprint arXiv:2006.12063*, 2020.

[18] Y. Han, J. Jiao, and T. Weissman. Minimax Estimation of Divergences Between Discrete Distributions. *IEEE J. Sel. Areas Inf. Theory*, 1(3):814–823, 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*, pages 770–778, 2016.

[20] A. Hegde, D. Erdogmus, T. Lehn-Schioler, Y. N. Rao, and J. C. Principe. Vector-quantization by density matching in the minimum kullback-leibler divergence sense. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 1, pages 105–109. IEEE, 2004.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proc. of NeurIPS*, page 6629–6640, 2017.

[22] D. J. Im, H. Ma, G. W. Taylor, and K. Branson. Quantitatively Evaluating GANs With Divergences Proposed for Training. In *ICLR*, 2018.

[23] T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, pages 4401–4410, 2019.

[24] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training Generative Adversarial Networks with Limited Data. In *NeurIPS*, 2020.

[25] S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust. Speech Signal Process.*, 35(3):400–401, 1987.

[26] D. P. Kingma and P. Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *NeurIPS*, pages 10236–10245, 2018.

[27] R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Inf. Theory*, 27(2):199–206, 1981.

[28] A. Krizhevsky and G. Hinton. Learning Multiple Layers of Features from Tiny Images. *Technical Report*, 2009.

[29] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019.

[30] D. Lopez-Paz and M. Oquab. Revisiting Classifier Two-Sample Tests. In *ICLR*, 2017.

[31] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley. Least Squares Generative Adversarial Networks. In *IEEE ICCV*, pages 2813–2821, 2017.

[32] D. Mcallester and L. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4(Oct):895–911, 2003.

[33] D. A. McAllester and R. E. Schapire. On the convergence rate of good-turing estimators. In N. Cesa-Bianchi and S. A. Goldman, editors, *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory (COLT 2000), June 28 - July 1, 2000, Palo Alto, California, USA*, pages 1–6. Morgan Kaufmann, 2000.

[34] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. In *Proc. of ICLR*, 2017.

[35] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo. Reliable fidelity and diversity metrics for generative models. In *ICML*, volume 119, pages 7176–7185, 2020.

[36] X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Trans. Inf. Theory*, 56(11):5847–5861, 2010.

[37] F. Nielsen and R. Bhatia. *Matrix information geometry*. Springer, 2013.

[38] A. Orlitsky and A. T. Suresh. Competitive Distribution Estimation: Why is Good-Turing Good. In *NeurIPS*, pages 2143–2151, 2015.

[39] A. Orlitsky, N. P. Santhanam, and J. Zhang. Always good turing: Asymptotically optimal probability estimation. In *44th Symposium on Foundations of Computer Science (FOCS 2003), 11-14 October 2003, Cambridge, MA, USA, Proceedings*, pages 179–188. IEEE Computer Society, 2003.

[40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[41] A. Razavi, A. van den Oord, and O. Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *NeurIPS*, pages 14837–14847, 2019.

[42] P. M. H. Ritter. Quantizing density estimators. In *NeurIPS*, volume 2, page 825. MIT Press, 2002.

[43] P. K. Rubenstein, O. Bousquet, J. Djolonga, C. Riquelme, and I. O. Tolstikhin. Practical and Consistent Estimation of f-Divergences. In *NeurIPS*, pages 4072–4082, 2019.

[44] A. Sablayrolles, M. Douze, C. Schmid, and H. Jégou. Spreading vectors for similarity search. In *Proc. of ICLR*, 2019.

[45] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018.

[46] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *NeurIPS*, pages 2226–2234, 2016.

[47] G. R. Shorack. *Probability for Statisticians*. Springer, 2000.

[48] J. Silva and S. Narayanan. Universal consistency of data-driven partitions for divergence estimation. In *2007 IEEE International Symposium on Information Theory*, pages 2021–2025. IEEE, 2007.

[49] J. Silva and S. S. Narayanan. Information divergence estimation based on data-dependent partitions. *Journal of Statistical Planning and Inference*, 140(11):3180–3198, 2010.

[50] L. Simon, R. Webster, and J. Rabin. Revisiting precision and recall definition for generative model evaluation. *CoRR*, abs/1905.05441, 2019.

[51] C. Tao, L. Chen, R. Henao, J. Feng, and L. Carin. Chi-square Generative Adversarial Network. In *ICML*, volume 80, pages 4894–4903, 2018.

[52] M. M. Van Hulle. Faithful representations with topographic maps. *Neural Networks*, 12(6): 803–823, 1999.

[53] S. Verdú. Empirical Estimation of Information Measures: A Literature Guide. *Entropy*, 21(8): 720, 2019.

[54] Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51(9):3064–3074, 2005.

[55] Y. Wang, D. Ramanan, and M. Hebert. Learning to Model the Tail. In *NeurIPS*, pages 7029–7039, 2017.

[56] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proc. of EMNLP*, pages 38–45, 10 2020.

[57] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending Against Neural Fake News. In *NeurIPS*, 2019.

[58] Z. Zhang and M. Grabchak. Nonparametric estimation of Küllback-Leibler divergence. *Neural Comput.*, 26(11):2570–2593, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Paper contributions are clearly and accurately stated in the abstract and introduction. Theoretical results can be found in Section 3 and Section 4 with details in Appendix. Experimental results can be found in Section 5 with details in Appendix G.
   (b) Did you describe the limitations of your work? [Yes] We describe the limitations of our work in a few places—below Theorem 2, below Proposition 3 and in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We believe there are no potential negative societal impacts for the results discussed in the paper since they are mainly theoretical results.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are discussed in the main text and precisely stated in Appendix.
   (b) Did you include complete proofs of all theoretical results? [Yes] Proofs are included in Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included them in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provided all the training details in Appendix G.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We reported error bars using shaded regions in all the plots in Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The details are provided in Appendix G.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We provided links and cited relevant references in Section 5 and Appendix G.
   (b) Did you mention the license of the assets? [No] We are not releasing any new assets. The licenses of the existing assets can be found in the cited sources.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]