Proper Losses for Discrete Generative Models

Dhamma Kimpara [1] **Rafael Frongillo** [1 *] **Bo Waggoner** [1 *]

Abstract

We initiate the study of proper losses for evaluating generative models in the discrete setting. Unlike traditional proper losses, we treat both the generative model and the target distribution as black-boxes, only assuming ability to draw i.i.d. samples. We define a loss to be black-box proper if the generative distribution that minimizes expected loss is equal to the target distribution. Using techniques from statistical estimation theory, we give a general construction and characterization of black-box proper losses: they must take a polynomial form, and the number of draws from the model and target distribution must exceed the degree of the polynomial. The characterization rules out a loss whose expectation is the cross-entropy between the target distribution and the model. By extending the construction to arbitrary sampling schemes such as Poisson sampling, however, we show that one can construct such a loss.

1. Introduction

Generative models are widely used tools in machine learning and statistics. For example, Generative Adversarial Networks (GANs) have recently been successful particularly in natural language and image generation. However, the *evaluation* of generative models is still an open area of research, with many evaluation methods proposed (Borji, 2019; Theis et al., 2015). This paper investigates theoretical foundations for evaluating generative models using a *proper losses* approach.

Specifically, we consider evaluating generative models that aim to match some underlying "target" distribution. For example, a GAN's goal may be to produce sentences from the same distribution as a random sentence drawn from Wikipedia; or to produce an image of a human face drawn from the same distribution as all U.S. passport photos. In areas such as climate modeling or weather forecasting, the goal may be to produce possible future trajectories from the same distribution as the actual climate. We abstract away from how the model is trained and learned; our focus is only on methods of evaluating the model. We leave the issue of training for future investigation.

We take the approach of the *proper losses* and *proper scoring rule* literature (McCarthy, 1956; Savage, 1971; Gneiting & Raftery, 2007), using one or more observations drawn from the target distribution to evaluate the model. However, many generative models are essentially "black boxes". One typically cannot obtain a closed form expression for the probabilities a model assigns to different outputs. This rules out using traditional proper losses for evaluating distributions, such as ℓ_2 loss or log loss. As a theoretical foundation, we instead assume only that we can draw independent and identically-distributed (i.i.d.) observations from the model p and compare these to observations from the target distribution q. The question is whether, and/or how, one can design losses under these restrictions that are *proper*: the expected loss is minimized by setting the model's distribution equal to the target, i.e. setting $p = q$.

Our results. As the initial work taking this approach, we focus on distributions over discrete, usually finite, sample spaces. We discuss extensions to the continuous setting in Section 7. First, we consider an easier problem: If we had full access to the target distribution q, i.e. in closed form or as an oracle, can we design proper losses for evaluating the model p from samples? We call this the report-black-box (RBB) setting. We show that the naive approach of plugging the empirical distributions directly into a distance function such as ℓ_2 does not yield a proper loss. However, by using the samples to construct unbiased estimators of the error introduced, we can correct for them and produce losses that are in fact proper.

Extending the unbiased-estimator approach, we characterize RBB-proper losses as those whose expectation is a polynomial in the model distribution, e.g. expected loss $\|p - q\|_k^k$ for even integers k. For such polynomials, we explicitly construct RBB-proper losses using the classical theory of unbiased estimators. Furthermore, the minimum number of

[*]Equal contribution [1]University of Colorado Boulder. Correspondence to: Dhamma Kimpara <dhamma.kimpara@colorado.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

observations that must be drawn from the model is exactly the degree of the polynomial. On the other hand, the characterization implies impossibility results for many popular forms of distances, including the cross entropy (expected log loss).

Second, we consider the full problem: what if we only have sample access to the target distribution q as well as the model p? Leveraging the above results, we give a similar characterization and construction for black-box (BB) proper losses. Again, the degree of the polynomial in p (respectively, q) governs the the size of the necessary and sufficient sample that must be drawn.

Generalizing, we consider more general sampling schemes that do not draw a predetermined number of observations. In particular, using Poisson sampling, we are able to overcome the above impossibility result and construct the first black-box proper loss that in expectation equals the cross entropy between p and q.

Finally, we experimentally evaluate our losses as a proof of concept.

1.1. Related Work

Our approach is based on the axiomatic approach in the proper scoring rule and proper losses literature, e.g. (Gneiting & Raftery, 2007). Most similar to our work in this tradition is Haghtalab et al. (2019), which examined convergence rates of the log loss for distribution learning in a setting similar to our simplified setting. Our characterizations will cover these proper scores as a special case, along with multi-observation losses that elicit a distribution (Casalaina-Martin et al., 2017).

There are many losses used in evaluation and training of GANs and other Neural Network (NN) based generative models (Borji, 2019; Theis et al., 2015). In adversarial training, much attention is given to obtaining unbiased gradients. These training losses cannot be translated into a proper loss because the loss is in a variational form that is inherent to the adversarial training method (Goodfellow et al., 2014; Bińkowski et al., 2018). However, the energy distance, a special case of the Maximum Mean Discrepancy (MMD) metric, has been used in its closed form to directly train NN based generative models (Dziugaite et al., 2015; Li et al., 2015; Székely & Rizzo, 2005; Bińkowski et al., 2018). The MMD in general is typically only available in a variational form and thus is not proper in practice. However, the energy distance actually can be used to construct a loss satisfying our definition of black-box proper. So it can be viewed as a pre-existing proof of concept for the ideas formalized and generalized in this paper. See Appendix G for further discussion.

In distribution learning (Han et al., 2020) and classical machine learning (Nguyen et al., 2010; Györfi & Van der Meulen, 1987; Hall & Morton, 1993; Joe, 1989), there is a line of work devoted to estimating divergences between pairs of distributions. While these literatures provide convergence and consistency results, the estimators and distances generally do not result in proper losses.

2. Background

For this work $\mathbb{N} = \{0, 1, 2, 3, \dots\}$. We primarily work with distributions over a finite domain $\mathcal{X}$. The set of all probability distributions over $\mathcal{X}$ is denoted by $\Delta_{\mathcal{X}}$. We denote a distribution by a vector of probabilities $p \in \Delta_{\mathcal{X}} \subset \mathbb{R}^{\mathcal{X}}$, where p_x is the probability p places on $x \in \mathcal{X}$. We use $\delta_x \in \Delta_{\mathcal{X}}$ to denote an indicator vector, i.e., the distribution placing probability one on x. Norms without a subscript are 2-norms: $\|\cdot\| = \|\cdot\|_2$.

In our setting, there is *target* distribution $q \in \Delta_{\mathcal{X}}$. We will generally use Y to denote observations drawn from q. We aim to evaluate a model that we will represent as $p \in \Delta_{\mathcal{X}}$, also a distribution. We will generally use X to denote observations drawn from p. Uppercase letters generally refer to random variables while lowercase letters are realizations, e.g. $X = x$.

We will also use various unbiased estimators from classical statistical estimation theory (see Appendix A). A function f is an unbiased estimator for a parameter θ of a family of distributions $\{F_\theta\}$ if, for any θ and any random variable $Z \sim F_\theta$, we have $\mathbb{E}\, f(Z) = \theta$. Unless otherwise specified, we will always use the minimum variance unbiased estimator (MVUE, see Appendix A).

We next recall the classical approach to evaluating p, which assumes full access to p in closed form. Then we introduce our setting, where we cannot access p except by drawing samples.

2.1. The classical approach: proper losses

We proceed with our theory via the perspective of proper losses. This literature was developed to elicit and evaluate general statistical reports or predictions from an agent. Introduced in (Brier, 1950), a proper loss (also historically termed a *proper scoring rule*) is a function $r(p, y)$ that assigns a loss to a model or forecast p on an observation y, where y is drawn from the target q. As we will see, proper losses do not apply in our setting because they assume ability to query the value of p_x on any x. Nevertheless, they are a useful starting point.

Definition 2.1. A loss function $r : \Delta_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$ is *proper* if for all $p, q \in \Delta_{\mathcal{X}}$, $\mathop{\mathbb{E}}_{y\sim q} r(q, y) \leq \mathop{\mathbb{E}}_{y\sim q} r(p, y)$. A loss is *strictly proper* if the above inequality is strict for all $p \neq q$.

In other words, for any fixed target distribution q, the op-

timal model p (i.e. the one that minimizes expected loss) is $p = q$. A classic result fully characterizes all proper losses via *Bregman divergences*, which can be used as measures of "distance" between two distributions. For reference, we define Bregman divergences and recall the scoring rule characterization in Appendix D.

The two most common proper losses are the *squared loss* $r(p, y) = \|p - \delta_y\|_2^2$, where δ_y is the indicator vector on y; and the *log loss* $r(p, y) = \log p_y$, whose expectation is the cross-entropy $\ell(p, q) = -\sum_{\mathcal{X}} q_x \ln p_x$.

3. Report Black Box Proper

To develop our results, in this section we consider a simplified setting where we have full access to the target distribution q. We aim to evaluate the model p based only on i.i.d. observations drawn from it. In later sections, we will assume only sample access to q as well.

3.1. Basic definitions

To evaluate p, we will draw i.i.d. observations from p. Formally, we draw a sample $(X_1, \ldots, X_n)$ where the X_i are independent random variables taking values in $\mathcal{X}$, each distributed according to p. It will be convenient to represent the sample as a *histogram* $H \in \mathbb{N}^{\mathcal{X}}$, where $H_x = |\{i : X_i = x\}|$. It is without loss of generality to consider loss functions that take H as input rather than the individual samples.[1]

We use $\mathcal{H}_n$ to denote the set of histograms arising from n samples, i.e. $\mathcal{H}_n = \{h \in \mathbb{N}^{\mathcal{X}} : \|h\|_1 = n\}$. We write $H \sim p^n$ to denote that the random histogram $H \in \mathcal{H}_n$ is distributed according to p^n, the distribution over all samples of size n drawn i.i.d. from p. Given a histogram $h \in \mathcal{H}_n$, the empirical distribution is $\hat{p} = \frac{1}{n}h$.

Definition 3.1. A *report-black-box (RBB) loss* is a function $L : \mathcal{H}_n \times \Delta_{\mathcal{X}} \to \mathbb{R}$. Here $L(h, q)$ is the loss assigned to a histogram h of n samples drawn from the model when the target distribution is q.

Definition 3.2. For a RBB loss $L : \mathcal{H}_n \times \Delta_{\mathcal{X}} \to \mathbb{R}$, the associated *expected loss* is $\bar{L}(p, q) = \mathop{\mathbb{E}}_{H \sim p^n} L(H, q)$.

The key property we want our loss functions to satisfy is *properness*, i.e., that expected loss is minimized by setting the model p equal to the target q. Therefore, the following definition becomes useful:

Definition 3.3. A function $\ell : \Delta_{\mathcal{X}} \times \Delta_{\mathcal{X}} \to \mathbb{R}$ is called a *proper divergence* if for all fixed q,

$$\ell(q, q) \leq \ell(p, q) \quad (\forall p).$$

It is called a *strictly proper divergence* if the above inequality is strict for all $p \neq q$.

Examples of proper divergences are the squared distance $\|p - q\|^2$ and the cross-entropy $\mathop{\mathbb{E}}_{Y \sim q} \log p_Y$. A proper divergence ℓ represents our goal: we would like to use such a divergence to evaluate p. In general, we cannot use ℓ directly, because evaluating the divergence requires access to the closed form of p, and we can only draw observations from p. However, we can *implement* a divergence ℓ if we can construct a RBB loss L whose expectation is ℓ. As such, the following captures what it means for L to be "proper" in our setting.

Definition 3.4. A report-black-box loss function L is *report-black-box proper (RBB proper)* if $\bar{L}(p, q)$ is a proper divergence. If ℓ is some proper divergence and there exists L such that $\bar{L} = \ell$, we will say that L *implements* ℓ and that ℓ is *implementable*.

3.2. Proof of concept: squared loss

Is there any proper divergence that is implementable? A priori, it might seem that given a loss $L : \mathcal{H}_n \times \Delta_{\mathcal{X}} \to \mathbb{R}$, there is always a way to tweak a misreport p to put higher weight on certain points and improve the expected loss.

Let us begin by investigating the ℓ_2 divergence $\ell(p, q) = \|p - q\|_2^2$. In the traditional proper loss (or proper scoring rule) setting, this yields a proper loss $r(p, y) = \|p - \delta_y\|_2^2$. Can we utilize squared loss as a RBB proper loss function by simply replacing p with $\hat{p}$? In fact, no:

Claim 1. *The loss $L(h, q) = \|\hat{p} - q\|_2^2$, where $\hat{p} = \frac{1}{n}h$ is the empirical distribution, is not RBB proper for any sample size n.*

Sketch. A straightforward calculation, using $p = \mathbb{E}\,\hat{p}$, shows that $\mathbb{E}\,\|\hat{p} - q\|^2 = \|p - q\|^2 + \sum_{x \in \mathcal{X}} \mathrm{Var}(\hat{p}_x)$. In general, this is not minimized by $p = q$; for example, with a 0.1-weighted coin, the optimal model p is always a coin with weight strictly less than 0.1 (notice this decreases the variance of $\hat{p}$). □

In summary, the expected loss of this naive approach is the proper divergence $\|p - q\|_2^2$ plus an extra term. However, the key insight is that the extra term can be estimated unbiasedly from a finite number of observations. Let $n \geq 2$ and let $s_n^2(\alpha) = \frac{1}{n-1}\left[\alpha(1-\alpha)^2 + (1-\alpha)\alpha^2\right]$. Then (Claim A.2.1) s_n^2 is an unbaised estimator for $\mathrm{Var}(\hat{p}_x)$, that is, $\mathop{\mathbb{E}}_{\hat{p} \sim p^n} s_n^2(\hat{p}_x) = \mathrm{Var}(\hat{p}_x)$. This proves the following.

[1] By exchangeability of i.i.d. samples, any function $f(S)$ of the sample $S = (X_1, \ldots, X_n)$ can be simulated by a function $g(H)$ of the histogram, where g simply arranges the samples that make up H in a uniformly random order to obtain S' and applies $f(S')$. Then $g(H)$ has the same distribution as $f(S)$, because S' has the same distribution as S.

Claim 2. *The ℓ_2 divergence $\ell(p,q) = \|p-q\|^2$ is implementable. In particular, for any $n \geq 2$, the following loss is RBB proper and satisfies $\bar{L}(p,q) = \|p-q\|^2$ (here $\hat{p} = \frac{1}{n}h$):*

$$L_n(h,q) = \|\hat{p} - q\|^2 - \sum_{\mathcal{X}} s_n^2(\hat{p}_x),$$

We discuss the reason underlying the variance term and generalize this construction to other divergences in Appendix D. A similar proof of concept can arise from considering the *energy distance* in continuous space, as discussed in Section 1.1 and Appendix G.

3.3. Minimum number of draws required

Now that we know it is possible to implement at least some proper divergences, a natural question is how many observations one needs to draw from p in order to do so. In cases where a generative model is expensive to sample, we might prefer to use RBB proper losses that can utilize a smaller sample size. To do so, we define the notion of a tight lower bound on the observations needed to implement a proper divergence.

Definition 3.5. Let $n \in \mathbb{N}$. A proper divergence, ℓ, is *n-minimally-implementable* if for all $n' \geq n$, there exists a RBB loss $L : \mathcal{H}_{n'} \times \Delta_{\mathcal{X}} \to \mathbb{R}$ that implements ℓ and, for all $k < n$, there does not exist a RBB loss $L : \mathcal{H}_k \times \Delta_{\mathcal{X}} \to \mathbb{R}$ that implements ℓ.

3.4. Characterization of Discrete Losses

We have seen that naively applying a proper divergence as a loss function introduces an extra penalty term, which can be corrected if we can unbiasedly estimate the penalty from samples. To make this approach fully general, we turn to the theory of U-estimation, which defines unbiased estimators. The key idea is that histogram $H \sim p^n$ has a multinomial distribution. There are classical results (Lemma A.3.1) describing which functions of multinomials have unbiased estimators. We utilize these results to characterize the proper divergences that are n-implementable. Also, we characterize the minimal-implementability of every such implementable divergence. We first recall the definition of a polynomial function of a vector.

Definition 3.6. A function $f : \Delta_{\mathcal{X}} \to \mathbb{R}$ is a *polynomial* if it is of the form

$$f(p) = \sum_{k \in K} a_k \prod_{x \in \mathcal{X}} p_x^{j_{k,x}},$$

where the sum is over a finite index set K, where each $\mathbf{j}_k \in \mathbb{N}^{\mathcal{X}}$ is unique, and where each a_k is a nonzero real number. In this case, the *degree* of f is $\max_{k \in K} \|\mathbf{j}_k\|_1$, i.e. the largest sum of exponents of any monomial. We say a function is *a polynomial in its jth argument of degree n* if, for all fixed values of the other arguments, the induced function of the jth argument alone is a polynomial, and there exists a maximum degree n of any such induced polynomial.

Theorem 1. *Let $\ell(p,q)$ be a proper divergence. Then ℓ is implementable if and only if it is a polynomial in its first argument. Furthermore, if ℓ is implementable, then ℓ is n-minimally implementable where n is the degree of the polynomial.*

Given a sample-size budget of n, Theorem 1 tells us which proper divergences can be implemented in evaluating a black-box model. Furthermore, the proof will actually construct a loss that minimally-implements the proper divergence.

Proof. Let ℓ be a proper divergence that is a polynomial in its first argument, in particular, of degree n. We show ℓ is implementable using sample size n. Write ℓ in the form of Definition 3.6, i.e. for each fixed q,

$$\ell(p,q) = \sum_{k \in K^{(q)}} a_k^{(q)} \prod_{x \in \mathcal{X}} p_x^{j_{k,x}^{(q)}},$$

where $K^{(q)}$ is finite, each $a_k^{(q)}$ is a nonzero constant, and each $\|\mathbf{j}_k^{(q)}\|_1 \leq n$. By classical results (Lemma A.3.1), any given monomial in p of degree at most n has an unbiased estimator using n samples from p. In particular, the minimum-variance unbiased estimator (MVUE) of the monomial $\prod_{\mathcal{X}} p_x^{j_{k,x}^{(q)}}$ is:

and satisfies $\mathop{\mathbb{E}}_{H \sim p^n}\left[t_{n,\mathbf{j}_k^{(q)}}(H)\right] = \prod_{\mathcal{X}} p_x^{j_{k,x}^{(q)}}$ (Lemma A.3.1). Therefore, the loss $L(h,q) = \sum_k a_k^{(q)} t_{n,\mathbf{j}_k^{(q)}}(h)$ satisfies $\bar{L} = \ell$, and it implements ℓ.

Now suppose ℓ is not a polynomial of degree at most n in its first argument. That is, there exists q such that $\ell(p,q)$ either has higher degree or is not a polynomial at all. The characterization of the U-estimable functions under the multinomial distribution, Lemma A.3.1, directly implies there does not exist an unbiased estimator for $\ell(\cdot,q)$ using sample size n, i.e. there does not exist $L : \mathcal{H}_n \times \Delta_{\mathcal{X}} \to \mathbb{R}$ such that $\mathop{\mathbb{E}}_{H \sim p^n} L(H,q) = \ell(p,q)$. This shows that non-polynomials are not implementable; and that polynomials of degree $n' > n$ are not implementable with only n observations. For the other part of minimally-implementable, our construction above implies that for all $n \geq deg(\ell)$ there exists a loss $L : \mathcal{H}_n \times \Delta_{\mathcal{X}} \to \mathbb{R}$ that implements ℓ. □

Corollary 1. *Let ℓ be a polynomial divergence as defined in definition 3.6. If L is constructed as according to Theorem 1 to implement ℓ, then L can be computed in $O(\sum_{k \in K} \|\mathbf{j}_k\|_1) = O(|K| deg(\ell))$ time.*

We immediately obtain some positive examples, such as:

Corollary 2. *For any even integer $k \geq 2$, the proper divergence $\|p-q\|_k^k = \sum_x (p_x - q_x)^k$ is implementable, and in particular, is k-minimally implementable.*

However, we also obtain impossibility results:

Corollary 3. *The cross-entropy $\sum_x q_x \log p_x$ is not implementable for any finite sample size.*

In Section 5, we will return to this example and show that cross-entropy actually can be "implemented" with a more creative approach to sampling.

3.5. Linearly Decomposable Losses

We now examine a special case of the previous characterization that includes many popular distance metrics. In our motivating example we found a report-black-box proper loss that in expectation is the the squared loss. It turns out to be the sum over all $\mathcal{X}$ of a coordinate-wise loss. We now leverage this and construct losses that implement certain distances that are linearly decomposable. We extend these losses to handle the case when $\mathcal{X}$ is countably infinite in Appendix C.

Corollary 4. *A linearly decomposable proper divergence, $\ell(p,q) = \sum_{\mathcal{X}} d(p_x, q_x)$, is n-minimally-implementable if and only if $d(\cdot,\cdot)$ is a polynomial with degree equal to n in the first argument.*

4. Black Box Properness

The RBB setting, while an important step, is not the most common in evaluating generative models. In this section, the fully black-box setting, we must evaluate only with samples from both the candidate model and the target distribution. We extend our definitions to encompass this setting. The RBB setting will be a special case of this more general setting.

Definition 4.1. A *black-box (BB) loss* is a function $L : \mathcal{H}_n \times \mathcal{H}_m \to \mathbb{R}$ where $L(h^p, h^q)$ is the loss assigned to histogram h^p of n samples drawn from the model on histogram h^q of m samples drawn from the target distribution.

Definition 4.2. For a black-box loss $L : \mathcal{H}_n \times \mathcal{H}_m \to \mathbb{R}$, the associated *expected loss* is $\bar{L}(p,q) = \mathop{\mathbb{E}}_{\substack{H^p \sim p^n \\ H^q \sim q^m}} L(H^p, H^q)$.

Definition 4.3. A black-box loss function L is *black-box proper (BB proper)* if $\bar{L}$ is a proper divergence ℓ. If ℓ is some proper divergence and there exists L such that $\bar{L} = \ell$, we will say that L *implements* ℓ and that ℓ is *implementable*.

We again define the notion of minimal-implementability. In cases where the target distribution is difficult to sample, we might prefer to use BB proper losses that can utilize a smaller target sample size. For example, generative models for forecasting e.g. climate may only have access to one observation from q, i.e. the weather that actually occurs on a given day. On the other hand, other settings may present other tradeoffs between model and target sample size.

Definition 4.4. A proper divergence, ℓ, is (n', m')-*minimally-implementable* if for all $n \geq n'$ and $m \geq m'$ there exists a BB loss $L : \mathcal{H}_n \times \mathcal{H}_m \to \mathbb{R}$ that implements ℓ and for all (k, j) where $k < n'$ or $j < m'$, there does not exist a loss $L : \mathcal{H}_k \times \mathcal{H}_j \to \mathbb{R}$ that implements ℓ.

4.1. Proof of concept: squared loss

We provide an illustrative example for the ℓ_2 proper divergence by extending the techniques we developed in Theorem 1. Again, the key idea is an unbiased estimator, namely $\delta^{Bin}_{j,k}(t) = \frac{t(t-1)\cdots(t-k+1)}{j(j-1)\cdots(j-k+1)}$. By Lemma A.4.1, if $T \sim \text{Binom}(j, \alpha)$ and $j \geq k$, then $\mathbb{E}\left[\delta^{Bin}_{j,k}(T)\right] = \alpha^k$. The point is that, for any x, the number of observations H^p_x is distributed Binomially, as is H^q_x, and they are independent.

Claim 3. *For distributions over a finite domain $\mathcal{X}$, the squared loss $\|q-p\|^2$ is implementable. In particular, for any $n \geq 2$ and $m \geq 2$, it is implemented by*

$$L_{n,m}(H^p, H^q) = \sum_{\mathcal{X}} \left[\frac{H^p_x(H^p_x - 1)}{n(n-1)} - \frac{2H^p_x H^q_x}{nm} + \frac{H^q_x(H^q_x - 1)}{m(m-1)} \right].$$

We observe that, although $L_{n,m}$ contains a sum over all $\mathcal{X}$, only at most $n+m$ terms will be nonzero, so $L_{n,m}$ is efficient to implement regardless of the size of the domain $\mathcal{X}$

Proof. Observe that $L_{n,m}(H^p, H^q) = \sum_{\mathcal{X}} \left[\delta^{Bin}_{n,2}(H^p_x) - 2\delta^{Bin}_{n,1}(H^p_x)\delta^{Bin}_{m,1}(H^q_x) + \delta^{Bin}_{m,2}(H^q_x)\right]$. Using that $\delta^{Bin}_{n,i}(H^p_x)$ is an unbiased estimator for p^i_x, and symmetrically for $\delta^{Bin}_{m,i}(H^q_x)$, along with independence of H^p and H^q, we immediately get $\mathbb{E}\, L(H^p, H^q) = \sum \left(p^2_x - 2p_x q_x + q^2_x\right) = \|p-q\|^2$. □

The fact that there exists any proper loss with only $n = 2$ observations from p and $m = 2$ observations from q is somewhat remarkable: however large the sample space $\mathcal{X}$, for example all sentences up to a fixed length or all images of a certain number of pixels, merely 4 total observations suffice to incent the learner to exactly set the model p to match the target q. In fact, slightly better is possible: the Brier score, i.e. the proper divergence $\sum_{\mathcal{X}} p^2_x - 2p_x q_x$, is $(2,1)$-minimally-implementable, as our next result will imply.

4.2. Characterization of Discrete Losses

As in the RBB setting, we utilize the theory of U-estimation to characterize the proper divergences that are implementable by BB losses. The proof follows similarly because H^p and H^q are independent random variables, so the RBB analysis above can essentially apply to each separately. The proof appears in Appendix B.

Theorem 2. *Let $\mathcal{T}$ be the set of all proper divergences. Let $\mathcal{F}_n$ be the set of all polynomials in the first argument with degree $\leq n$ and $\mathcal{F}_m$ be the set of all polynomials in the second argument with degree $\leq m$. The set of all (n, m)-implementable proper divergences is*

$$BB_{n,m} = \mathcal{T} \cap \mathcal{F}_n \cap \mathcal{F}_m.$$

Furthermore, a proper divergence ℓ is (j, k)-minimally-implementable if and only if it has degree in the first argument equal to j and degree in the second argument equal to k.

4.3. Consequences and connections to proper scoring rules

There are a number of consequences and special cases of note. One class of special cases is $m = \infty$, which we use to denote the case where we have full access to the target q in closed form. Then we obtain $BB_{n,\infty}$, which reduces to the report-black-box (RBB) setting. Similarly, $n = \infty$ denotes the case where we have full access to the model p in closed form, which reduces to the traditional proper loss setting. In particular, $BB_{\infty,1}$ is the set of proper losses. Furthermore, by the same reasoning as in Theorem 2, we know that any $\ell \in BB_{\infty,1}$ must be linear in the second argument and must also be a proper divergence. Hence one could follow this reasoning as an alternative approach to characterizing all proper scoring rules (Theorem D.0.1).

Corollary 5. *Let $\phi(BB_{\infty,1})$ be all the proper scoring rules (in the form of losses). Then*

$$BB_{n,1} = \bigcup_{L \in \phi(BB_{\infty,1})} \{\bar{L} : \bar{L} \text{ is a polynomial in the first argument with degree } \leq n\}.$$

Corollary 5 is relevant to fields using generative models to forecast events, such as in weather or climate forecasting. In these cases, we may be able to draw n i.i.d. observations from the learner's model p, but only $m = 1$ observation from nature, i.e. the weather that actually occurs. In such cases, Corollary 5 implies that we can construct a BB proper loss from any proper scoring rule that is a polynomial in p.

Corollary 6. *For $n, m \in \mathbb{N}$, $BB_{n,m} \subseteq BB_{n,\infty} \cap BB_{\infty,m}$. In other words, if a divergence is (n, m)-BB implementable then it is n-RBB implementable and implementable via a multi-observation proper loss with m observations.*

Corollary 7. *For $n, m \in \mathbb{N}, BB_{n,m} \subseteq BB_{n+1,m} \cap BB_{n,m+1} \subseteq BB_{n+1,m+1}$.*

Corollary 8. *If $\mathcal{T}$ in Theorem 2 is the set of all strictly proper divergences, then $BB_{1,m} = \emptyset$.*

5. Poisson Sampling

So far our results show that proper divergences must all be polynomial in the distributions in order to be implementable. As such, cross entropy cannot be (n, m)-implemented for any finite $n, m \in \mathbb{N}$. We now show that cross entropy can be implemented if we generalize to other sampling schemes. A sampling scheme is a (possibly randomized) stopping rule determining the number of samples to draw from a generative model. In Appendix E, we formally define and fully characterize implementable proper divergences under arbitrary sampling schemes. Here, we focus on the example of Poisson sampling specifically for the cross-entropy divergence. Other sampling schemes admit a multitude of other distinct classes of U-estimable functions.

We will determine the implementable proper divergences under Poisson sampling schemes. Poisson sampling gives us much more powerful estimators than in the scheme where we draw a deterministic sample size. The Poisson distribution is a discrete probability distribution over $\mathbb{N}$ with parameter $\theta > 0$ and probability mass function $f(j; \theta) = \Pr[T = j] = \frac{\theta^j e^{-\theta}}{j!}$.

The sampling scheme is as follows. Let $\alpha, \beta > 0$. First randomly draw the sample sizes $N \sim Poi(\alpha)$ and $M \sim Poi(\beta)$. Then draw N observations from p and M observations from q. Poisson sampling gives us two powerful properties. First, the counts of each outcome, h_x^p (resp. h_x^q), are independent and distributed according to $Poi(\alpha p_x)$ (resp. $Poi(\beta q_x)$). Second, we are able to unbiasedly estimate θ^k for any $k \in \mathbb{N}$ and thus can unbiasedly estimate any power series involving θ. This estimation is achieved by the Poisson estimator:

$$\delta_k^{Poi}(t) = t(t-1)\cdots(t-k+1).$$

By Lemma A.5.1, if $T \sim Poi(\theta)$ then $\mathbb{E}\,\delta_k^{Poi}(T) = \theta^k$ for any $k \in \mathbb{N}$. We will use this estimator extensively in this section. The first result immediately follows from this estimator. The second follows from the characterization of U-estimable functions in Lemma A.5.1.

Corollary 9. *For any $n, m \in \mathbb{N}$ and $\alpha, \beta > 0$, $BB_{n,m} \subset BB_{Poi(\alpha),Poi(\beta)}$, the set of all implementable functions with Poisson sampling from p and q.*

Corollary 10. *A proper divergence is $(Poi(\alpha), Poi(\beta))$-implementable for any $\alpha, \beta > 0$ if and only if it has an equivalent power series expression in the first and second arguments with non-negative integer powers and the power series satisfies 1) every coefficient of the first and second*

arguments is finite and 2) if the series diverges for any argument, the proper divergence also diverges in the same direction (goes to $+\infty$ or $-\infty$).

The proof of Corollary 10 appears in Appendix B. Crucially for implementing the cross entropy, many functions in C^∞ have an equivalent (Taylor) series that satisfy the conditions of corollary 10. We will use the Taylor series for $\ln(x)$ in the next section.

5.1. Cross Entropy

As a consequence of Corollary 10, we can construct a generic-black-box proper loss that implements the cross entropy. By similar methods we can also implement the Shannon entropy and the KL divergence. Note that with deterministic sampling, we cannot construct such a loss.

Lemma 1. *Let $h^p_{-x} = \sum_{y \neq x} h^p_y$ be number of occurences of all the outcomes except x. Then for any $\alpha, \beta > 0$ the loss,*

$$L(h^p, h^q) = \sum_{\mathcal{X}} \frac{1}{\beta} \delta_1^{Poi}(h^q_x) \sum_{k=1}^{\infty} \frac{1}{k} \frac{1}{\alpha^k} \delta_k^{Poi}(h^p_{-x}),$$

$(Poi(\alpha), Poi(\beta))$-implements the cross entropy, $\ell(p, q) = -\sum_{\mathcal{X}} q_x \ln(p_x)$.

We observe that this loss is always finite for a finite sample because $\delta_k^{Poi}(t) = 0$ when $t < k$. In fact, the loss is efficient to evaluate, i.e. polynomial time in terms of the number of samples drawn, as that number governs the number of nonzero terms. Memoization of the infinite sums and the summands in the infinite sum provides the most efficient way to compute this loss. Furthermore, the computation is highly parallelizable.

Corollary 11. *Let L be the loss that implements cross entropy in Lemma 1, $N \sim Poi(\alpha), M \sim Poi(\beta)$, and c be the number of unique non-zero integers in $\{h^p_{-x}\}_{x \in \mathcal{X}}$. Then L can be computed in $O(|\mathcal{X}| + cN) = O(|\mathcal{X}| + N^{1.5})$ time. If the histogram counts are stored in a dictionary-like data structure, and an element only has an entry if it was observed, then the amortized runtime is $O(\min(|\mathcal{X}|, \beta) + \alpha^{1.5})$.*

In correspondence with the cross entropy, this loss can equal infinity in expectation for certain p, q, although it is finite for every h^p, h^q. We note that the cross entropy can also be $(Poi(\alpha), m)$-implemented, with any $m \geq 1$, with the loss $L(h^p, h^q) = \sum_{\mathcal{X}} \hat{q}_x \sum_{k=1}^{\infty} \frac{1}{k} \frac{1}{\alpha^k} \delta_k^{Poi}(h^p_{-x})$, where $\hat{q} = \frac{1}{m} h^q$.

Proof. We will use the Taylor expansion for $\ln(t)$. For $t \in [0, 1], \ln(t) = -\sum_{k=1}^{\infty} \frac{1}{k}(1-t)^k$. Note that the series diverges to $-\infty$ at $t = 0$ but also $\lim_{t \to 0} \ln(t) = -\infty$. Next, we will use that H^p and H^q are independent; that H^q_x is distributed $Poi(\beta q_x)$; and that H^p_{-x} is distributed $Poi(\alpha(1 - p_x))$.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{H^p \sim p^n \\ H^q \sim q^m}} L(H^p, H^q) = \\
&= \mathop{\mathbb{E}}_{\substack{H^p \sim p^n \\ H^q \sim q^m}} \sum_{\mathcal{X}} \frac{1}{\beta} \delta_1^{Poi}(H^q_x) \sum_{k=1}^{\infty} \frac{1}{k} \frac{1}{\alpha^k} \delta_k^{Poi}(H^p_{-x}) \\
&= \sum_{\mathcal{X}} q_x \sum_{k=1}^{\infty} \frac{1}{k} (1 - p_x)^k \\
&= -\sum_{\mathcal{X}} q_x \ln(p_x),
\end{aligned}$$

including the case where both the proper divergence and the expected value of the BB loss equals ∞ (i.e. there exists x with $q_x > 0$ and $p_x = 0$). □

We note that the KL-divergence, $\ell(p, q) = \sum_{\mathcal{X}} q_x \ln \frac{q_x}{p_x}$, can be implemented as well. In fact, it equals the cross-entropy plus the Shannon entropy of q. Shannon entropy can be estimated unbiasedly with Poisson sampling because H^q_x and H^q_{-x} are distributed as *independent* Poissons, so

$$\mathbb{E}\, H^q_x \sum_{k=1}^{\infty} \frac{1}{k} \frac{1}{\beta^k} \delta_k^{Poi}(H^q_{-x}) = q_x \sum_{k=1}^{\infty} \frac{1}{k} (1 - q_x)^k = q_x \ln(q_x).$$

6. Experiments

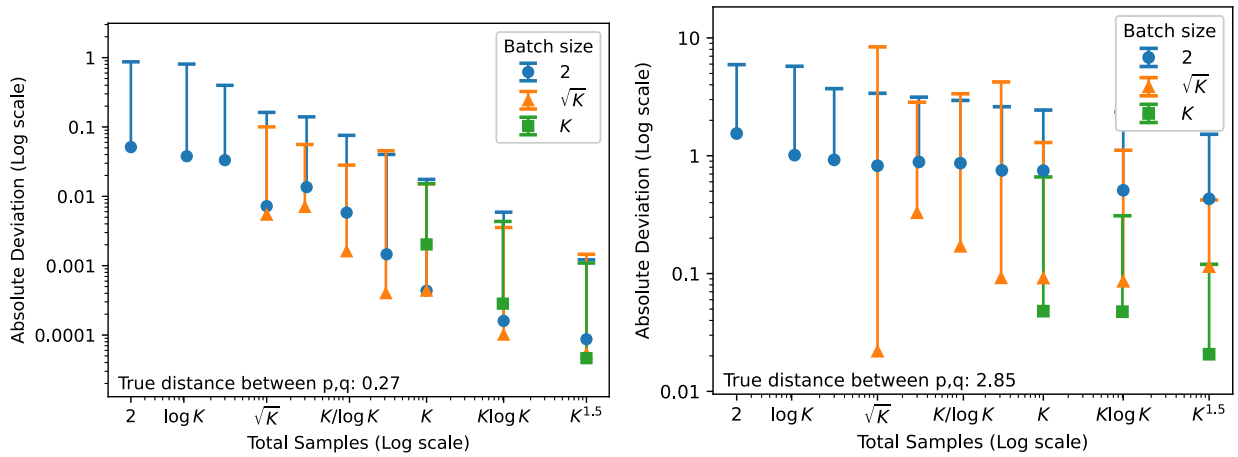
For a proof of concept, we performed numerical experiments to evaluate our loss functions on a variety of pairs of distributions. We focused on the black-box setting, since this evaluation setting is more difficult than the report black-box setting. For this section we define $K := |\mathcal{X}|$ and we call divergences distances.

We consider the task of distinguishing different power law distributions, which often arise in connection with natural language data. Results for other pairs and types of distributions appear in Appendix H.

For each pair of distributions p and q, at each number of total samples, we measured the absolute deviation between the loss value and the true distance between the distributions. We drew up to $K^{1.5}$ total samples. We repeated this experiment for various batch sizes, where at each iteration, we drew the same batch size from p and q.

Of course, our losses work even with different batch sizes. For simplicity we kept the batch sizes the same.

We can discern from our experiments that given a budget of samples, the black-box loss is generally more accurate when all the samples are used in the computation of a single black-box loss value. This is opposed to splitting the sample



(a) Squared Distance

(b) Cross Entropy

Figure 1. The y-axis is the absolute deviation between the black-box loss value and the true distances (Squared distance and Cross Entropy). $K = 10,000$, p and q are Zipfians with parameters 1 and 2, respectively. 30 trials for each parameter setting was recorded (including batch size). The horizontal bars represent the maximum absolute deviation of any of the 30 trials. The solid markers represent the average of the trials. Squared distance was estimated using the loss from claim 3. Cross entropy was estimated with a poisson sampling loss. Batch sizes were the same between draws from p and q.

into smaller batches and computing the average of the BB loss over all the batches. This suggests a theoretical result that it is always better to use the largest possible batch size. Note however that in the squared distance case, varying the batch size did not change the accuracy much. We also note that the figures in appendix H show that the Poisson estimator consistently under-estimates the true loss value.

We also exhibit excellent convergence. In both cases, at and past $K \log K$ total samples, the average value of our losses over all the trials are within ≈10% of the true distance between the distributions. The normalized plots showing multiplicative error are included in the appendix.

7. Discussion

Larger batches are better. As we saw in the experiments, if one has a budget of samples, it is best to use all those samples in a single instance of a BB proper loss function, rather than split those samples up into smaller batches and taking the average of the loss over all the batches. In general, a direct extension would be to analyze the variance and convergence rate of these BB losses with regards to the batch size.

Losses for continuous domains. We have focused on the discrete case in this work, leaving the continuous case to further investigation. However, we illustrate an initial result in the continuous setting. Let $F_p(\cdot)$ be the CDF of distribution p and $F_S(x) = \frac{|\{i:X_i \leq x\}|}{n}$ be the empirical CDF based on sample $S = (X_1, \ldots, X_n)$ where each X_i is drawn i.i.d. from p.

Theorem 3. *Let $\mathcal{X} = \mathbb{R}$ and $\alpha_i \in \mathbb{R}$ for all i. Let a proper divergence be of the form $\ell(p,q) = \int_{\mathbb{R}} g(\{F_p(x + \alpha_i)\}_{i=1}^m, \cdot)dx$. If $g(\cdot,\cdot)$ is a polynomial in the first argument with powers $\mathbf{j}_k^{(q)} \in \mathbb{Z}_+^{|\mathcal{X}|}$ such that $\|\mathbf{j}_k^{(q)}\|_1 = n$, the number of samples, then g is n-minimally-implementable.*

The proof appears in Appendix B. As a corollary, we are able to implement the Cramér distance which we exhibit in Appendix F. These types of distances can easily be extended to a high dimensional distance by picking a direction at random and defining the empirical CDFs based on the hyperplane defined by that random direction. We illustrate this via a high dimensional version of the Cramér distance in appendix F.

Future work. A direction of future work is that of constructing black-box proper losses for continuous settings, which is the most common use-case for GANs. Another important study would be to investigate the properness of existing losses used in evaluation. Finally, it would be interesting to investigate the use of BB proper losses in evaluating implicit distributions of black-boxes for desired properties. For example evaluating a dice for uniformity or evaluating prepared quantum states.

Broader Impacts

The evaluation of generative models, such as GANs, is a very open question with important societal impacts in domains such as climate forecasting. We provide an initial theoretical foundation for this question. Instead of direct applications, we anticipate this work leading to further theo-

retical investigation. It may inform practitioners' choices of which losses they use for evaluating generative models. Of course, such evaluation can be used for ethical or unethical purposes. We do not know of particular risks or negative impacts of this work beyond risks of generative models in general.

Acknowledgements

We thank Claire Monteleoni and Amit Rege for helpful discussions and references. This material is based upon work supported by the US National Science Foundation under Grant No. IIS-2045347.

References

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Borji, A. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.

Brier, G. W. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Casalaina-Martin, S., Frongillo, R., Morgan, T., and Waggoner, B. Multi-observation elicitation. In *Conference on Learning Theory*, pp. 449–464. PMLR, 2017.

Cramér, H. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

Glasser, G. J. Minimum variance unbiased estimators for poisson probabilities. *Technometrics*, 4(3):409–418, 1962.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Györfi, L. and Van der Meulen, E. C. Density-free convergence properties of various estimators of entropy. *Computational Statistics & Data Analysis*, 5(4):425–436, 1987.

Haghtalab, N., Musco, C., and Waggoner, B. Toward a characterization of loss functions for distribution learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Hall, P. and Morton, S. C. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45(1): 69–88, 1993.

Han, Y., Jiao, J., and Weissman, T. Minimax estimation of divergences between discrete distributions. *IEEE Journal on Selected Areas in Information Theory*, 1(3):814–823, 2020.

Hoeffding, W. Range preserving unbiased estimators in the multinomial case. In *The Collected Works of Wassily Hoeffding*, pp. 613–615. Springer, 1994.

Joe, H. Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41(4):683–697, 1989.

Kolmogorov, A. N. Unbiased estimates. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, 14(4):303–326, 1950.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015.

McCarthy, J. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956.

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

Rockafellar, R. T. *Convex Analysis*, volume 36. Princeton University Press, 1970.

Savage, L. J. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

Székely, G. J. and Rizzo, M. L. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

Theis, L., Oord, A. v. d., and Bethge, M. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

Appendix

A. Results from Statistical Estimation Theory

We have extensively utilized results from Unbiased Estimation (U-Estimation) theory. These estimators are fundamental to our construction of proper losses for generative models.

A.1. Definitions

Since there are possibly infinitely many U-estimators for many quantities, the literature provides a criteria for the 'best' estimator:

Definition A.1.1. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. from some member of a family of densities, $p_\theta, \theta \in \Omega$. An estimator δ is a Minimum Variance Unbiased Estimator (MVUE) for $g(\theta)$ if for some $n \in \mathbb{N}$, for all $\theta \in \Omega$,

1. δ is an unbiased estimator for $g(\theta)$, $\mathbb{E}\,\delta(Y_1, Y_2, \ldots, Y_n) = g(\theta)$,
2. $\text{Var}(\delta(Y_1, Y_2, \ldots, Y_n)) \leq \text{Var}(\tilde{\delta}(Y_1, Y_2, \ldots, Y_n))$ for any other unbiased estimator $\tilde{\delta}$.

A.2. MVUE for Variance

Fact A.2.1. *(Canonical MVUE for Variance) Let $(y_i)_{i=1}^n$, $n \geq 2$, be i.i.d. realizations of a random variable Y. Then the MVUE for variance is*

$$s_n^2((y_i)_{i=1}^n) := \frac{1}{n-1}\sum_{i=1}^n (y_i - \frac{1}{n}\sum y_i)^2.$$

Claim A.2.1. *If $Y_i \overset{iid}{\sim} Ber(\alpha)$ and $Z = Y_1 + Y_2 + \cdots + Y_m$ then $Z \sim Bin(m, \alpha)$. Let $\bar{Y} = Z/m = \frac{1}{m}\sum_i Y_i$. Then for $m \geq 2$,*

$$s_m^2(\bar{Y}) = \frac{1}{m-1}[\bar{Y}(1-\bar{Y})^2 + (1-\bar{Y})(\bar{Y})^2]$$

is a MVUE for variance. Note that Var$(\bar{Y}) = \frac{\alpha(1-\alpha)}{m}$.

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{Z=m\bar{Y}} s_m^2(\bar{Y}) &= Var(\bar{Y}) \\
&= Var(\frac{1}{m}\sum_{i=1}^m Y_i) \\
&= \frac{1}{m^2} Var(\sum_{i=1}^m Y_i) \\
&= \frac{1}{m^2}[\sum_{i=1}^m Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j)] \\
&= \frac{1}{m^2}[mVar(Y_1) + 0] \\
&= \frac{1}{m} Var(Y_1) \\
&= \frac{1}{m}\,\mathbb{E}\, s_m^2(Y_1) \\
&= \frac{1}{m}\,\mathbb{E}\,\frac{1}{m-1}\sum_{i=1}^m (Y_i - \bar{Y})^2 \\
&= \frac{1}{m}\,\mathbb{E}\,\frac{1}{m-1}\sum_{i=1}^m (\mathbb{1}_{Y_i=1} - \bar{Y})^2
\end{aligned}$$

$$= \frac{1}{m}\,\mathbb{E}\,\frac{1}{m-1}[m\bar{Y}(1-\bar{Y})^2 + m(1-\bar{Y})(0-\bar{Y})^2]$$
$$= \mathbb{E}_{\bar{Y}}\,\frac{1}{m-1}[\bar{Y}(1-\bar{Y})^2 + (1-\bar{Y})(\bar{Y})^2].$$

□

A.3. Multinomial Estimator

Lemma A.3.1. *(Kolmogorov, 1950) Let $Y \sim M(m,p)$ be a multinomial random variable. A real-valued function $f(p)$ has an unbiased estimator on the basis of an observation from Y if and only if f is a polynomial of degree at most m. The unique MVUE of such a polynomial is constructed using the following estimators (Hoeffding, 1994),*

$$t_{m,\mathbf{j}_k}(y) = \frac{\prod_{\mathcal{X}} y_x(y_x-1)\cdots(y_x-j_{k,x}+1)}{m(m-1)\cdots(m-\|\mathbf{j}_k\|_1+1)}.$$

Where y_x is the number of observations of element $x \in \mathcal{X}$ in the sample and $\mathbb{E}\,t_{m,\mathbf{j}_k}(Y) = \prod_{\mathcal{X}} p_x^{j_{k,x}}$. The binomial distribution is a special case of the multinomial distribution.

A.4. Binomial Estimator

Lemma A.4.1. *(Lehmann & Casella, 2006). Let $T \sim Bin(m,\alpha)$. Then $f(\alpha)$ is unbiasedly estimable if and only if f is a polynomial with degree $\leq m$. The MVUE estimator for $f(\alpha) = \alpha^k$, $k \leq m$, is*

$$\delta^{Bin}_{m,k}(t) = \frac{t(t-1)\cdots(t-k+1)}{m(m-1)\cdots(m-k+1)}.$$

Hence $\mathbb{E}\,\delta^{Bin}_{m,k}(T) = \alpha^k$.

A.5. Poisson Estimator

Lemma A.5.1. *(Glasser, 1962) A function of the Poisson parameter θ has an unbiased estimator if and only if the function can be expressed as a series in integer non-negative powers of θ. Let $T \sim Poi(\theta)$. Then for all $k \in \mathbb{N}$ the MVUE estimator of θ^k is*

$$\delta^{Poi}_k(t) = t(t-1)\cdots(t-k+1).$$

Note that if $t < k$ then $\delta^{Poi}_k(t) = 0$. Hence $\mathbb{E}\,\delta^{Poi}_k(T) = \theta^k$.

The MVUE for any estimable $F(\theta)$ can be constructed by writing the function as a power series and replacing all the θ^t with its unbiased estimator given in Lemma A.5.1 (Glasser, 1962). Suppose our random variable for the count of x is $H_x \sim Poi(\alpha p_x)$. We can unbiasedly estimate p_x^k:

$$\frac{1}{\alpha^k}\,\mathbb{E}\,\delta^{Poi}_k(H_x) = \frac{1}{\alpha^k}(\alpha p_x)^k = p_x^k.$$

B. Omitted Proofs

B.1. Omitted Proofs from Section 4

Proof of Theorem 2. Let $\ell \in BB_{n,m}$. We first show that ℓ is (n,m)-implementable. ℓ is a proper divergence by definition. ℓ is a also a polynomial in both arguments with bounded degree, so let us write it in the following form:

$$\ell(p,q) = \sum_{k\in K} a_k \prod_{\mathcal{X}} p_x^{i_{k,x}} \prod_{\mathcal{X}} q_x^{j_{k,x}},$$

where K is finite; for all $k \in K$, a_k is a nonzero constant; $\mathbf{i}_k, \mathbf{j}_k \in \mathbb{N}^{\mathcal{X}}$ with the pair unique for each k; $\|\mathbf{i}_k\|_1 \leq n$ and $\|\mathbf{j}_k\|_1 \leq m$. The construction of the implementing loss is similar to that in the proof of Theorem 1. Again, we have that

$H^p \sim Multinomial(n, p)$ and also $H^q \sim Multinomial(m, q)$. Hence we use the estimators from classical results, $t_{\mathbf{i}_k}$ and $t_{\mathbf{j}_k}$, to estimate each summand in ℓ:

$$\mathop{\mathbb{E}}_{\substack{H^p \sim p^n \\ H^q \sim q^m}} L(H^p, H^q) = \mathop{\mathbb{E}}_{\substack{H^p \sim p^n \\ H^q \sim q^m}} \sum_{k \in K} a_k t_{\mathbf{i}_k}(H^p) t_{\mathbf{j}_k}(H^q) = \ell(p, q).$$

Where the last equality follows by independence of H^p and H^q, and Lemma A.3.1. Thus ℓ is (n, m)-implementable, and a proper divergence by definition.

For the converse, if ℓ is a proper divergence but $\ell \notin BB_{n,m}$, then by the characterization (Lemma A.3.1) of the U-estimable functions under a multinomial distribution, for all k, there does not exist $t_{\mathbf{i}_k}$ or $t_{\mathbf{j}_k}$ such that $\mathop{\mathbb{E}}_{H \sim p^n} t_{\mathbf{i}_k}(H^p) = \prod_{\mathcal{X}} p_x^{i_{k,x}}$ and similarly for $\prod_{\mathcal{X}} q_x^{j_{k,x}}$. Thus ℓ is not (n, m)-implementable. Minimal-implementability also follows from this characterization, Lemma A.3.1. □

Proof of Corollary 8. By Lemma A.3.1, the losses in $BB_{1,m}$ that are implementable are $\{g : g(p, q) = \sum_{\mathcal{X}} f_x(q) p_x\}$. Where the degree of each $f_x(q)$ is $\leq m$. For a generic g we now find the report that minimizes the expected loss.

$$\begin{aligned} \frac{d}{dp} g(p, q) &= \frac{d}{dp} \sum_{\mathcal{X}} f_x(q) p_x \\ &= \sum_{\mathcal{X}} f_x(q) \end{aligned}$$

Thus any report minimizes the expected loss of any function that is $(1, m)$-implementable hence none of these expected losses are strictly proper divergences. In other words, all $(1, m)$-implementable divergences are constant for a fixed q. □

B.2. Omitted Proofs from section 5

Proof of Corollary 10. By characterization in Lemma A.5.1 of functions estimable under a Poisson distribution, $\mathcal{F}_{Poi(\alpha)} = \{\ell(\cdot, \cdot) : \ell \text{ is a power series in the first argument with non-negative integer powers}\}$. $\mathcal{F}_{Poi(\beta)}$ is similarly defined in terms of the second argument. The corollary follows by applying Theorem E.0.1. □

B.3. Proof of Lemma 1

Proof. We will use the Taylor expansion for $\ln(x)$. For $x \in [0, 1], \ln(x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}(x-1)^k$. Note that the series diverges to $-\infty$ at $x = 0$ but also $\lim_{x \to 0} \ln(x) = -\infty$. Shannon entropy is implemented by the fact that H^p_{-x} is a Poisson random variable distributed according to $Poi(\alpha p_{-x})$ that is independent from H^p_x.

$$\begin{aligned} \mathop{\mathbb{E}}_{\substack{H^p \sim p^n \\ H^q \sim q^m}} L(H^p, H^q) &= -\mathop{\mathbb{E}}_{\substack{H^p \sim p^n \\ H^q \sim q^m}} \sum_{\mathcal{X}} \frac{1}{\beta} \delta_1^{Poi}(H^q_x) \sum_{k=1}^{\infty} \frac{-1}{k} \frac{1}{\alpha^k} \delta_k^{Poi}(H^p_{-x}) \\ &= -\sum_{\mathcal{X}} \frac{1}{\beta} \mathop{\mathbb{E}}_{H^q \sim q^m} \left[\delta_1^{Poi}(H^q_x)\right] \sum_{k=1}^{\infty} \frac{-1}{k} \frac{1}{\alpha^k} \mathop{\mathbb{E}}_{H \sim p^n} \left[\delta_k^{Poi}(H^p_{-x})\right] \\ &= -\sum_{\mathcal{X}} q_x \sum_{k=1}^{\infty} \frac{-1}{k} p_{-x}^k \\ &= -\sum_{\mathcal{X}} q_x \sum_{k=1}^{\infty} \frac{-1}{k} (1 - p_x)^k \\ &= -\sum_{\mathcal{X}} q_x \sum_{k=1}^{\infty} \frac{-1}{k} (-1)^k (p_x - 1)^k \\ &= -\sum_{\mathcal{X}} q_x \ln(p_x). \end{aligned}$$

□

B.4. Proofs from section 7

Proof of Theorem 3. We only need to show that we can unbiasedly estimate each term in g. The result then follows by linearity of expectation. To do this we will show that the vector valued random variable $T = \{F_S(x+\alpha_i)\}_{i=1}^k$ is a function of a multinomial random variable. Hence the unbiased estimators and result follows from Lemma A.3.1.

Without loss of generality let $\alpha_1 \leq \alpha_2 \cdots \leq \alpha_m$ and define $\alpha_0 := -\infty$. Now define $Z \sim Multinomial(n, (F_p(x+\alpha_1), F_p(x+\alpha_2) - F_p(x+\alpha_1), \ldots, F_p(x+\alpha_m) - F_p(x+\alpha_{m-1}))$. Z is a vector valued random variable where the count Z_i, $i \in \{1, 2, \ldots m\}$, corresponds to how many samples fall in the interval $[x+\alpha_i, x+\alpha_{i-1}]$. Hence we can rewrite the random variable $F_S(x+\alpha_i)$ as

$$F_S(x+\alpha_i) = \frac{1}{n}(Z_1 + Z_2 + \cdots Z_i) = \frac{1}{n}\sum_{j=1}^{i} Z_j.$$

Now if g is a polynomial in the first argument, then

$$g(\{F_p(x+\alpha_i)\}_{i=1}^m, \cdot) = \sum_{\mathbf{j}_k^{(q)}} a_{\mathbf{j}_k^{(q)}} \prod_{i=1}^m F_p(x+\alpha_i)^{j_{k,i}^{(q)}}.$$

Where $a_{\mathbf{j}_k^{(q)}}$ subsumes the second argument. Now we show that the product, $\prod_{i=1}^m F_p(x+\alpha_i)^{j_{k,i}^{(q)}}$, is unbiasedly estimable. The result follows by linearity of expectation. Now by the condition of the theorem, $\|\mathbf{j}_k^{(q)}\|_1 \leq n$ so we only have at most n distinct α_i in each product hence we can ignore all the other α_i that have a power of 0. Thus now we define a multinomial random variable as before except now only with the α_i that are involved. Let's reindex $\mathbf{j}_k^{(q)}$ and α so that all the entries where $j_{k,i}^{(q)} = 0$ are above index B. Again let $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_B$, then the corresponding random variable is $Z \sim Multinomial(n, (F_p(x+\alpha_1), F_p(x+\alpha_2) - F_p(x+\alpha_1), \ldots))$. Thus by the multinomial characterization, we can estimate polynomials of the parameters of this distribution (we know n):

$$\begin{aligned}\prod_{i=1}^B F_p(x+\alpha_i)^{j_{k,i}^{(q)}} &= \prod_{i=1}^B \left(F_p(x+\alpha_1) + \sum_{\gamma=2}^{i} [F_p(x+\alpha_\gamma) - F_p(x+\alpha_{\gamma-1})] \right)^{j_{k,i}^{(q)}} \\ &= \prod_{i=1}^B \left(\mathbb{E}\,\frac{Z_1}{n} + \sum_{\gamma=2}^{i} \mathbb{E}\,\frac{Z_\gamma}{n} \right)^{j_{k,i}^{(q)}}.\end{aligned}$$

Since $\|\mathbf{j}_k^{(q)}\|_1 \leq n$ for all k, this term will have degree at most n in the parameters of the multinomial distribution and so is unbiasedly estimable with the multinomial estimator. Each parameter is exactly $\frac{1}{n}\,\mathbb{E}\,Z_i$. While this means there are different multinomial distributions for each product term in the polynomial, each term effectively 'sees' these different distributions. Thus by linearity of expectation this is a valid way to construct the unbiased estimator. □

C. Countably Infinite Domains

Lemma C.0.1. *Let $\mathcal{X}$ be countably infinite. Let $\mathcal{X}_k \subset \mathcal{X}$ be a finite subset for all $k \in \mathbb{N}$. Let a proper divergence be of the form*

$$\ell(p, q) = \sum_{k=1}^{\infty} a_k d_{\mathcal{X}_k}(p_x, q_x),$$

where $\sum_{k=1}^{\infty} a_k$ converges and $d_{\mathcal{X}_k}$ is a (r,t)-implementable divergence on the empirical distribution restricted to $\mathcal{X}_k$, and bounded for all $\mathcal{X}_k, p$, and q. Then ℓ is (r,t)-implementable.

For example, $\mathcal{X}$ could be a set of all english word sentences and $\mathcal{X}_k$ could be the set of all length k sentences.

D. The Variance Bias Term and Connection to Proper Scoring Rules

One explanation of why the variance appears in the calculations for naive squared error has to do with Bregman Divergences and the Jensen Gap. For this section the empirical distribution is defined as $\hat{p} = \frac{1}{|H|}H$

We recall the definition of a Bregman divergence:

Definition D.0.1. Let G be a convex, real valued function. Then the Bregman divergence of G is

$$D_G(q, p) = G(q) - [G(p) + \langle \partial G(p), q - p \rangle]$$

where $dG(p)$ is a subgradient of G at p (Rockafellar, 1970).

One reason Bregman divergences are important is because they are known to characterize traditional proper losses:

Theorem D.0.1 (McCarthy (1956); Savage (1971); Gneiting & Raftery (2007))**.** *A loss r is proper if and only if there exists a convex, real valued function G such that*

$$\mathop{\mathbb{E}}_{y \sim q} r(p, y) = \mathop{\mathbb{E}}_{y \sim q} [D_G(\delta_y, p) - G(\delta_y)].$$

Where $\delta_y := e_y$ is the indicator vector for coordinate y.

For the purposes of illustrating the role of the Jensen gap, we show implementation of the divergences $D_G(p, q)$, which have the arguments in the opposite order to the version implemented by proper losses. Later, we show how to implement the usual ordering of the arguments, $D_G(p, q)$. We begin with the RBB case for clarity of exposition.

D.1. Jensen Gap

Lemma D.1.1. *If $D_G(p, q)$ is a Bregman divergence then it can be RBB-implemented, provided that an unbiased estimator, δ, exists for $G(p)$.*

$$\begin{aligned} L(\hat{p}, q) &= D_G(\hat{p}, q) - [G(\hat{p}) - \delta(\hat{p})] \\ &= G(\hat{p}) - [G(q) + \langle \partial G(q), \hat{p} - q \rangle] - [G(\hat{p}) - \delta(\hat{p})] \\ &= \delta(\hat{p}) - [G(q) + \langle \partial G(q), \hat{p} - q \rangle] \end{aligned}$$

Note that $\mathbb{E}[G(\hat{p}) - \delta(\hat{p})] = \mathbb{E}\, D_G(\hat{p}, p)$, the Jensen gap. This can be interpreted as the expected additional distance the randomness of $\hat{p}$ adds.

Proof. Begin with the law of cosines for Bregman divergences and take the expectation of both sides.

$$\begin{aligned} \mathop{\mathbb{E}}_{H \sim p^n} D_G(\hat{p}, q) &= \mathbb{E}\, D_G(\hat{p}, p) + D_G(p, q) - \mathbb{E}\langle \hat{p} - p, \partial G(q) - \partial G(p) \rangle \\ &= \mathbb{E}\, D_G(\hat{p}, p) + D_G(p, q) - 0 \qquad (1) \\ &= \mathbb{E}\left[G(\hat{p}) - [G(p) + \langle \partial G(p), \hat{p} - p \rangle]\right] + D_G(p, q) \\ &= \mathbb{E}[G(\hat{p})] - [G(p) + \langle \partial G(p), \mathbb{E}\,\hat{p} - p \rangle] + D_G(p, q) \\ &= \mathbb{E}[G(\hat{p})] - G(\mathbb{E}\,\hat{p}) - 0 + D_G(p, q). \qquad (2) \end{aligned}$$

Let us note several things here. First, line (1) formalizes the intuition we have outlined in the lemma. Second we also clearly see that the expected Bregman divergence between $\hat{p}$ and p, $\mathop{\mathbb{E}}_{\hat{p} \sim p^n} D_G(\hat{p}, p)$, is exactly the Jensen gap, $\mathbb{E}[G(\hat{p})] - G(\mathbb{E}\,\hat{p})$, as exhibited by the resulting expression in (2). Hence, rearranging for clarity we see that

$$\begin{aligned} D_G(p, q) &= \mathop{\mathbb{E}}_{H \sim p^n} \left[D_G(\hat{p}, q) - D_G(\hat{p}, p)\right] \\ &= \mathbb{E}\, D_G(\hat{p}, q) - \left[\mathbb{E}\, G(\hat{p}) - G(\mathbb{E}\,\hat{p})\right]. \end{aligned}$$

Which gives us the RBB-implementing loss for D_G if we have an unbiased estimator δ where $\mathbb{E}\,\delta(\hat{p}) = G(\mathbb{E}\,p) = G(p)$.

□

Example D.1.1. As an example, let's look at the Jensen Gap for $G(x) = \|x\|^2$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{H\sim p^n} G(\hat{p}) - G(\mathop{\mathbb{E}}_{H\sim p^n} \hat{p}) &= \mathbb{E}[\|\hat{p}\|^2] - \|\,\mathbb{E}\,\hat{p}\|^2 \\
&= \mathbb{E}[\sum_{\mathcal{X}} \hat{p}_x^2] - \sum_{\mathcal{X}} \mathbb{E}[\hat{p}_x]^2 \\
&= \sum_{\mathcal{X}} \mathbb{E}[\hat{p}_x^2] - \mathbb{E}[\hat{p}_x]^2 \\
&= \sum_{\mathcal{X}} Var(\hat{p}_x).
\end{aligned}$$

Lemma D.1.2. *If $D_G(p,q)$ is a Bregman divergence then an equivalent divergence can be BB-implemented, provided that an unbiased estimators δ, and δ' where $\mathbb{E}\,\delta(\hat{p}) = G(p)$, and $\mathbb{E}\,\delta'(\hat{q}) = \partial G(q)$ exist.*

$$L(\hat{p},\hat{q}) = \delta(\hat{p}) - \delta(\hat{q}) - \langle \hat{p}, \delta'(\hat{q})\rangle.$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\substack{H^p\sim p^n \\ H^q\sim q^m}} L(\hat{p},\hat{q}) &= \mathbb{E}\left[\delta(\hat{p}) - \delta(\hat{q}) - \langle \hat{p}, \delta'(\hat{q})\rangle\right] \\
&= G(p) - G(q) - \langle p, \partial G(q)\rangle \\
&= G(p) - [G(q) + \langle \partial G(q), p-q\rangle] - \langle q, \partial G(q)\rangle \\
&= D_G(p,q) - \langle q, \partial G(q)\rangle.
\end{aligned}$$

□

Note that the divergence implemented by the above loss, $D_G(p,q) - \langle q, \partial G(q)\rangle$, when considered from the candidate distributions point of view, is merely the original divergence, $D_G(p,q)$, minus a constant. Thus to the candidate model, the loss is equivalent up to a constant to the original Bregman divergence. The same idea is used for classical proper losses as exhibited by Theorem D.0.1.

D.2. Implementing $D_G(q,p)$

We will implement the same equivalent divergence as in Theorem D.0.1:

$$D_G(q,p) - G(q) = G(p) - \langle \partial G(p), p\rangle + \langle \partial G(p), q\rangle.$$

Notice that this equivalent divergence can be implemented if each additive term can be unbiasedly estimated on its own. Thus we will need unbiased estimators for q, $G(p)$, $\partial G(p)$, and $\langle p, \partial G(p)\rangle$. In the case of deterministic or Poisson sampling, if $G(p)$ can be unbiasedly estimated then $\partial G(p)$ and $\langle p, \partial G(p)\rangle$ can be unbiasedly estimated. Estimating q is easy.

We will use the characterization of deterministic sampling and Poisson unbiasedly estimable functions. The following is applicable to both Poisson and deterministic sampling.

Suppose $G(p)$ is unbiasedly estimable. Then $G(p) = \sum_{k\in K} a_k \prod_{x\in\mathcal{X}} p_x^{j_{k,x}}$ where $|K|$ is possibly infinite. Then

$$\partial G(p)_y = \sum_{k\in K: j_{k,y}\neq 0} a_k \;\; j_{k,y} p_y^{j_{k,y}-1} \prod_{x\neq y} p_x^{j_{k,x}}$$

is an (infinite) polynomial, thus it is unbiasedly estimable. Note that if $\{k \in K : j_{k,y} \neq 0\} = \emptyset$ then $\partial G(p)_y = 0$, for all p. Furthermore,

$$\langle p, \partial G(p) \rangle = \sum_{\mathcal{X}} p_x \partial G(p)_x$$

is an also an (infinite) polynomial, so unbiasedly estimable.

One can view the following results as corollaries to the characterization of Poisson implementable divergences (Corollary 10).

Lemma D.2.1. *Let r be a Poisson sampling scheme and t be any sampling scheme that can estimate q. If $D_G(q, p)$ is a Bregman divergence then an equivalent divergence can be (r, t)-implemented if and only if G has an equivalent power series expression in the first and second arguments with non-negative integer powers and the power series satisfies 1) every coefficient of the first and second arguments is finite and 2) if the series diverges for any argument, the proper divergence also diverges in the same direction (goes to $+\infty$ or $-\infty$).*

Corollary D.2.1. *Let r be a deterministic size sampling scheme drawing n samples and t be any sampling scheme that can estimate q. If $D_G(q, p)$ is a Bregman divergence then an equivalent divergence can be (n, t)-implemented if and only if G is a polynomial with degree less than or equal to n.*

E. General Sampling Schemes

Definition E.0.1. A *generic-black-box (GBB)* loss is a function $L : \mathbb{N}^{\mathcal{X}} \times \mathbb{N}^{\mathcal{X}} \to \mathbb{R}$ where $L(h^p, h^q)$ is the loss assigned to histogram h^p of samples drawn from the model on histogram h^q of samples drawn from the target distribution.

The difference between a GBB loss and a BB loss (Definition 4.1) is that we allowed BB losses to be a function of histograms of a specific, predetermined size (n and m). In contrast, a GBB loss must be defined for histograms of any size. These functions can also compute N, the sample size.

Definition E.0.2. A *sampling scheme* r is a stopping rule for the process of drawing observations from a black-box generative model. The stopping rule may depend on the history of the seen observations and may also use randomness.

Definition E.0.3. Let r, t be sampling schemes for the report and the target distribution, respectively. A generic-black-box loss L is (r, t)*-black-box proper* if $\bar{L}(p, q) := \mathop{\mathbb{E}}_{r,t} L(H^p, H^q)$ is a proper divergence. If ℓ is some proper divergence and there exists L such that $\bar{L} = \ell$, we will say that L (r, t)*-implements* ℓ and that ℓ is (r, t)*-implementable*.

Given a characterization of the U-estimable functions under certain sampling schemes, we can construct the set of implementable proper divergences. We can also construct the respective implementing losses from these characterizations. We do not investigate the sample complexity of the schemes or define minimally-implementable in the generic setting. While one could consider ordering generic sampling schemes by e.g. expected number of ramples drawn, the most reasonable ordering of sampling schemes is not always clear, and we leave such investigations to future work.

Theorem E.0.1. *Let r, t be sampling schemes. Let $\mathcal{T}$ be the set of all proper distances and let*

$$\mathcal{F}_r = \{\ell(\cdot, \cdot) : \ell \text{ is unbiasedly estimable in the first argument under sampling scheme } r\}$$

$$\mathcal{F}_t = \{\ell(\cdot, \cdot) : \ell \text{ is unbiasedly estimable in the second argument under sampling scheme } t\}.$$

Then the set of all (r, t)-implementable proper divergences is

$$BB_{r,t} = \mathcal{T} \cap \mathcal{F}_r \cap \mathcal{F}_t$$

Proof. If we can characterize $\mathcal{F}_r$ and $\mathcal{F}_t$ then we have a characterization of the unbiasedly estimable functions under sampling schemes r and t, respectively. These characterizations must provide constructions of the unbiased estimators. Thus we can construct an unbiased estimator for each $\ell \in F_r \cap F_t$. Hence $\ell \in BB_{r,t}$ is implementable and a proper divergence, by definition. □

F. Omitted results from section 7

For this section we consider densities on continuous domains. For a density p over $\mathbb{R}$, $F_p(\cdot)$ is the CDF of p.

Definition F.0.1 (Empirical CDF)**.** Given a sample $s = \{X_i\}_{i=1}^n$ the empirical CDF is defined as $F_s(x) := \frac{|i : X_i \leq x|}{n}$.

F.1. Implementation of the Cramér Distance

Corollary F.1.1. *For densities p, q over $[0,1]$, let s and u be samples drawn from p and q, respectively. Then the loss*

$$L(s,u) = \int_{\mathbb{R}} (F_s(x) - F_u(x))^2 - s^2_{|s|}(F_s(x)) - s^2_{|u|}(F_u(x))\, dx$$

$(2,2)$-minimally-implements the Cramér distance, $\ell(p,q) = \int_{\mathbb{R}} (F_p(x) - F_q(x))^2 dx$ (Cramér, 1928).

Proof of Corollary F.1.1. Notice that $F_S(x) = \frac{|i:X_i \le x|}{n}$ is distributed according to $\frac{1}{n}Bin(n, F_p(x))$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,U} L(S,U) &= \mathop{\mathbb{E}}_{S,U} \int_{\mathbb{R}} (F_S(x) - F_U(x))^2 - s^2_{|S|}(F_S(x)) - s^2_{|U|}(F_U(x))\, dx \\
&= \int_{\mathbb{R}} \mathbb{E}[F_S(x)^2] - 2\,\mathbb{E}\,F_U(x)\,\mathbb{E}\,F_S(x) + \mathbb{E}[F_U(x)]^2 - Var(F_S(x)) - Var(F_U(x)) dx \\
&= \int_{\mathbb{R}} \mathbb{E}[F_S(x)]^2 - 2F_q(x)F_p(x) + \mathbb{E}[F_U(x)]^2 dx \\
&= \int_{\mathbb{R}} (F_p(x) - F_q(x))^2 dx.
\end{aligned}$$

Where the second equality is by the independence of S and U and the previously defined variance estimator for the binomial distribution (claim A.2.1). The third equality is by expanding expectation of the squared term and reducing (the second non-centered moment of a binomial). One could also prove this using the technique from the proof of claim 3. □

The energy distance in one dimension is equivalent to twice the Cramér distance. Thus the energy distance also gives a loss that implements the Cramér distance. See appendix G for a discussion of the relationships between different types of losses in the continuous setting

F.2. High Dimensional Extension of the Cramér Distance

Let us now work in a continuous domain where the samples are from $\mathbb{R}^j$. Now instead of distributions $p, q \in \Delta_{\mathcal{X}}$ we will have densities p, q on $\mathbb{R}^j$. The desired score will again be the Harald Cramér distance. However now we will define a CDF with respect to a direction and then integrate over all directions.

Definition F.2.1. (Generalized CDF). Let Y be a random variable taking values in $\mathbb{R}^j$, p be the associated density, and $v \in \mathbb{R}^j$ such that $\|v\| = 1$. Then the direction v CDF of Y is

$$F^v_p(x) = \Pr[\langle v, Y\rangle + x \le 0].$$

Where $x \in \mathbb{R}$.

The distance analogous to the Harald Cramér distance is then

$$\int_{\substack{v \in \mathbb{R}^j \\ ||v||=1}} \int_{\mathbb{R}} (F^v_p(x) - F^v_q(x))^2 dx\, dv.$$

Now to create a sample proper loss we may again introduce a variance correction term as before. However, we also note that if we pick a random direction v, then we would not have to integrate over all v since the expectation of the distance under a random v is the same as the deterministic distance. Below we show the RBB loss, however the result for the BB version is very similar; one can compare between corollary F.1.1 and the following.

Claim F.2.1. *Let p, q be densities over $\mathbb{R}^j$ and s be the sample drawn from p. Then the following loss is RBB proper. First pick a random unit vector $v \in \mathbb{R}^j$ then*

$$L(s, q) = \int_{\mathbb{R}} (F_s^v(x) - F_q^v(x))^2 - s_n^2(F_s^v(x))\, dx.$$

In other words, L implements $\ell(p, q) = \int_{\substack{v \in \mathbb{R}^j \\ ||v||=1}} \int_{\mathbb{R}} (F_p^v(x) - F_q^v(x))^2 dx\, dv$.

Proof. Let $S = (X_1, X_2, \ldots, X_n)$.

$$\begin{aligned}
\mathop{\mathbb{E}}_{v \in \mathcal{S}^{j-1}} \mathop{\mathbb{E}}_{S} L(\hat{p}, q) &= \mathop{\mathbb{E}}_{v \in \mathcal{S}^{j-1}} \mathop{\mathbb{E}}_{S} \int_{\mathbb{R}} (F_S^v(x) - F_q^v(x))^2 - s_n^2(F_p^v(x))\, dx \\
&= \int_{\substack{v \in \mathbb{R}^j \\ ||v||=1}} \mathbb{E} \int_{\mathbb{R}} (F_S^v(x) - F_q^v(x))^2 - s_n^2(F_p^v(x))\, dx\, dv \\
&= \int_{\substack{v \in \mathbb{R}^j \\ ||v||=1}} \int_{\mathbb{R}} \mathbb{E}[(F_S^v(x) - F_q^v(x))^2 - s_n^2(F_p^v(x))]\, dx\, dv \\
&= \int_{\substack{v \in \mathbb{R}^j \\ ||v||=1}} \int_{\mathbb{R}} F_p(x)^2 + \frac{F_p^v(x)(1 - F_p^v(x))}{n} - 2F_q^v(x)F_p^v(x) + F_q^v(x)^2 \\
&\qquad - \frac{F_p^v(x)(1 - F_p^v(x))}{n}\, dx\, dv \\
&= \int_{\substack{v \in \mathbb{R}^j \\ ||v||=1}} \int_{\mathbb{R}} F_p(x)^2 - 2F_q^v(x)F_p^v(x) + F_q^v(x)^2\, dx\, dv \\
&= \int_{\substack{v \in \mathbb{R}^j \\ ||v||=1}} \int_{\mathbb{R}} (F_p^v(x) - F_q^v(x))^2 dx\, dv.
\end{aligned}$$

Let $C \sim Bin(n, F_p^v(x))$. Once again note that $F_S^v(x) = \frac{1}{n}|\{X_i \in S : \langle v, X_i \rangle + x \leq 0\}| = \frac{1}{n}C$. Hence we expand the expectation with the first and second moment as in Claim F.1.1. □

G. Discussion of other continuous losses

We discuss our results in the previous section in relation to two other methods of generative model evaluation in the continuous setting. Our results rely on computing losses based on the empirical CDF whether in one or many dimensions.

First, unless estimation/smoothing is done on the empirical density, it is not possible to work with losses that integrate over a function of the two densities at every point in the outcome space. There is a large body of work on density estimation for evaluating generative models. However, losses based on kernel density estimation are beyond the scope of this work.

Second, one can trivially construct proper losses based on functions of the random variables associated with densities p and q. For example the energy distance is

$$D^2(F, G) = 2\,\mathbb{E}\,\|X - Y\| - \mathbb{E}\,\|X - X'\| - \mathbb{E}\,\|Y - Y'\|.$$

Where X, X' and Y, Y' are independent copies of the random variable associated with density p and q, respectively.

The number of independent copies of a random variable in the expression is exactly the number of independent samples from that random variable required to unbiasedly estimate the loss. For the energy distance, we need 2 independent samples from p and q each. In one dimension these can also be written as functions of the empirical CDF.

G.1. Connection to energy distance in one dimension

In the one dimensional continuous setting, we have densities p, q on $\mathbb{R}$. We repeat the proof that the energy distance is equal to twice the Cramér distance. We can see from our approach or the form of the energy distance that this loss is $(2, 2)$-minimally implementable.

Lemma G.1.1. *(Székely & Rizzo, 2005) Let X, X' be i.i.d. with CDF $F(x)$ and Y, Y' be i.i.d. with CDF $G(y)$. Then the energy distance in one dimension is equal to twice the Cramér distance.*

$$2\,\mathbb{E}\,|X - Y| - \mathbb{E}\,|X - X'| - \mathbb{E}\,|Y - Y'| = 2\int_{\mathbb{R}} (F(x) - G(x))^2 dx.$$

Proof. We will convert the energy distance into the Cramer distance. First we use the identity

$$|X - Y| = \int_{\mathbb{R}} \mathbb{1}(X \le u < Y) + \mathbb{1}(Y \le u < X) du$$

Now let $A = \mathbb{E}\,|X - Y|, B = \mathbb{E}\,|X - X'|, C = |Y - Y'|$. We then use Fubini's theorem.

$$\begin{aligned}
A &= \mathbb{E}\,|X - Y| \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}(X \le u < Y) + \mathbb{1}(Y \le u < X) du dx dy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}(X \le u < Y) + \mathbb{1}(Y \le u < X) dx dy du \\
&= \int_{\mathbb{R}} \Pr[X \le u]\Pr[Y > u] + \Pr[X > u]\Pr[Y \le u] du \\
&= \int_{\mathbb{R}} F(u)(1 - G(u)) + (1 - F(u))G(u) du \\
&= \int_{\mathbb{R}} F(u) - 2F(u)G(u) + G(u) du
\end{aligned}$$

Hence by similar derivation, $B = \int_{\mathbb{R}} 2F(u) - 2F(u)^2 du$ and $C = \int_{\mathbb{R}} 2G(u) - 2G(u)^2 du$. The lemma follows by simple algebra.

□

G.2. Connection to the CRPS

We derived the Cramér distance via extending the Continuously Ranked Probability Score from the proper scoring rules literature. Intuitively, one can think CRPS as evaluating a distribution against an empirical distribution consisting of a single sample (Gneiting & Raftery, 2007). Let F_r be the CDF of a density r and again p be the reported distribution and q the true distribution. Then the CRPS (in terms of a loss to be minimized) for a outcome particular y drawn from the density q is

$$\int_{-\infty}^{\infty} (F_p(x) - \mathbb{1}\{x \ge y\})^2 dx = \int_{-\infty}^{\infty} (F_p(x) - F_{\hat{q}}(x))^2 dx. \tag{3}$$

Where $\hat{q}(x) = H^q$ is the empirical distribution of the data consisting of the single sample y. Note that $F_{\hat{q}}(x)$ is 0 below y and 1 when greater than or equal to y. Hence $F_{\hat{q}}(x) = \mathbb{1}\{x \ge y\}$. It is easy to see from the form of the CRPS that CRPS is also $(2, 1)$-minimally-implementable since the LHS of (3) contains a polynomial of degree 2 in F_p and it requires only

1 sample from q. There also exists a form of the CRPS derived from the form of the equivalent energy distance that also shows $(2, 1)$-minimal-implementability (Gneiting & Raftery, 2007).

To extend CRPS to our setting, in which we have an empirical densities for $\hat{p}$ and $\hat{q}$, we derived

$$\ell(p, q) = \int_0^1 (F_p(x) - F_q(x))^2 dx.$$

Which is the Harald Cramér distance (Cramér, 1928). The CRPS is a special case when we draw only 1 sample from q. We give a BB loss that implements this distance in claim F.1.1.

H. Omitted Experiments

H.1. Distribution Definitions

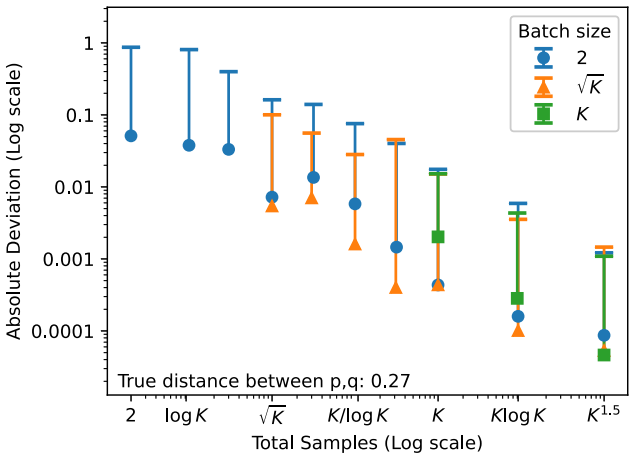
Definition H.1.1 (Spiked Uniform)**.**

$$p_x = \begin{cases} 0.1 & \text{if } x \in \{1, 2, 3, 4, 5\} \\ \frac{1-0.5}{K-5} & \text{otherwise} \end{cases}$$

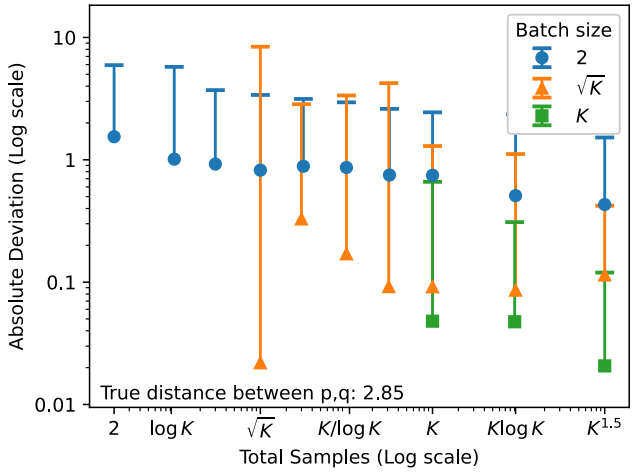
Definition H.1.2 (Spiked Zipfian(r))**.** Let z_x be the probability mass of x in a Zipfian(r) distribution over an outcome space $\{1, \ldots, K\}$

$$p_x = \begin{cases} \frac{0.05}{1.15} & \text{if } x \in \{5, 10, 20\} \\ \frac{z_x}{1.15} & \text{otherwise} \end{cases}$$

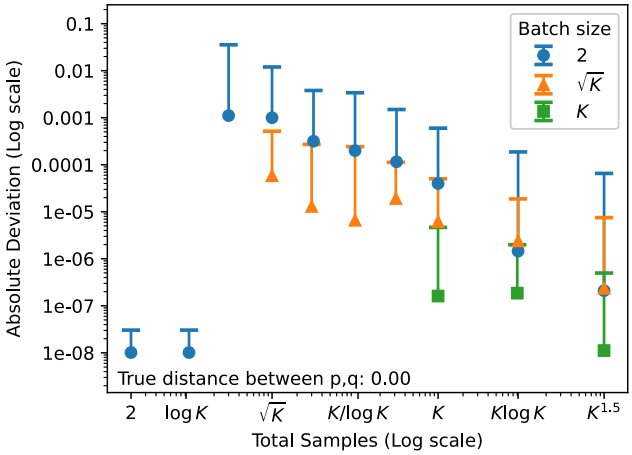
H.2. Absolute Deviation



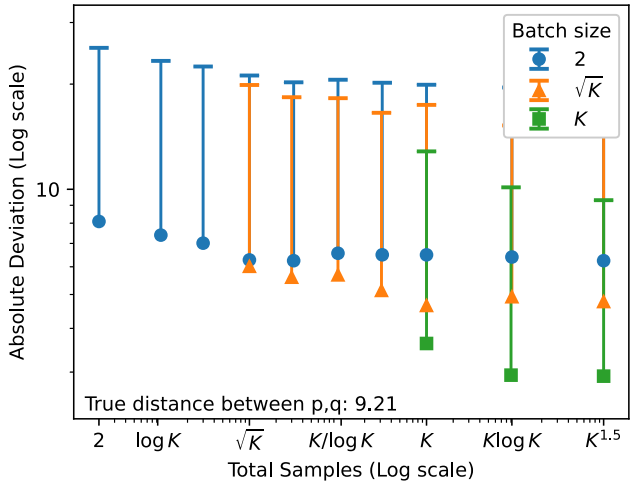
(a) Squared Distance, p: Zipfian(1), q: Zipfian(2)



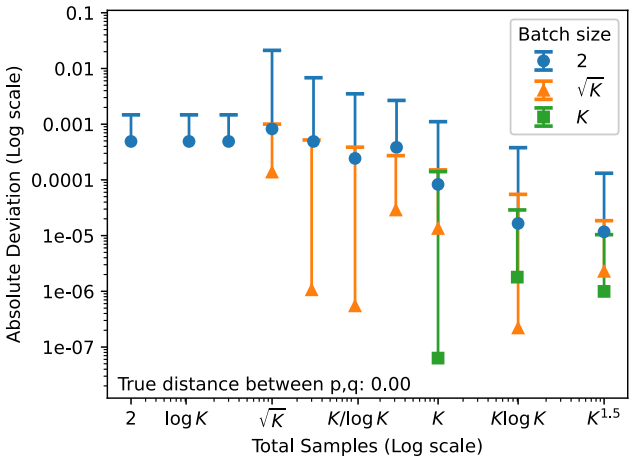
(b) Cross Entropy, p: Zipfian(1), q: Zipfian(2)



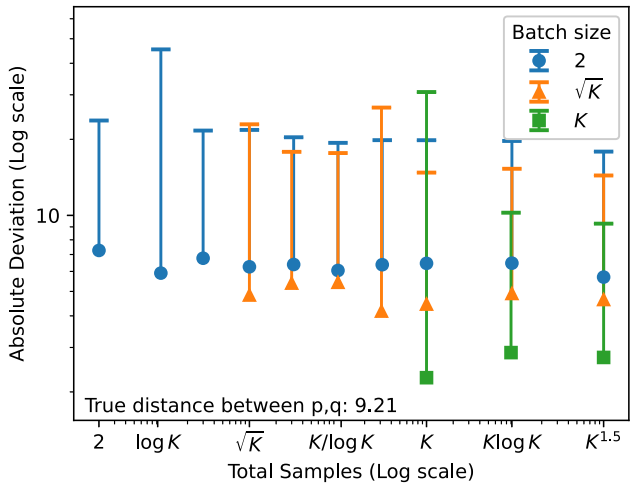
(c) Squared Distance, p: Uniform, q: Zipfian(0.01)



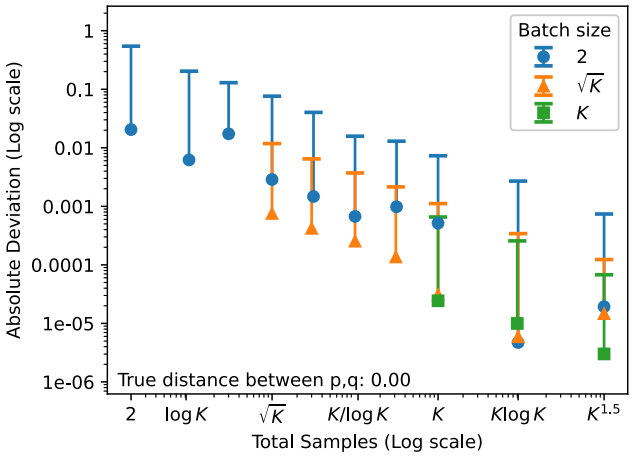
(d) Cross Entropy, p: Uniform, q: Zipfian(0.01)



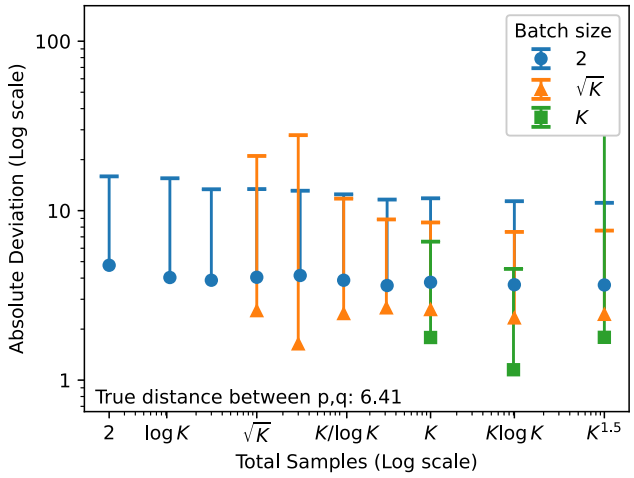
(e) Squared Distance, p: Uniform, q: Spiked Uniform (definition H.1.1)



(f) Cross Entropy, p: Uniform, q: Spiked Uniform (definition H.1.1)



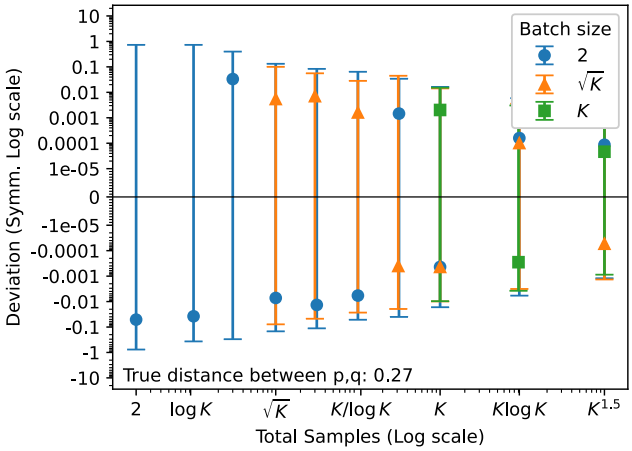
(g) Squared Distance, p: Zipfian(1), q: Spiked Zipfian(1) (definition H.1.2)



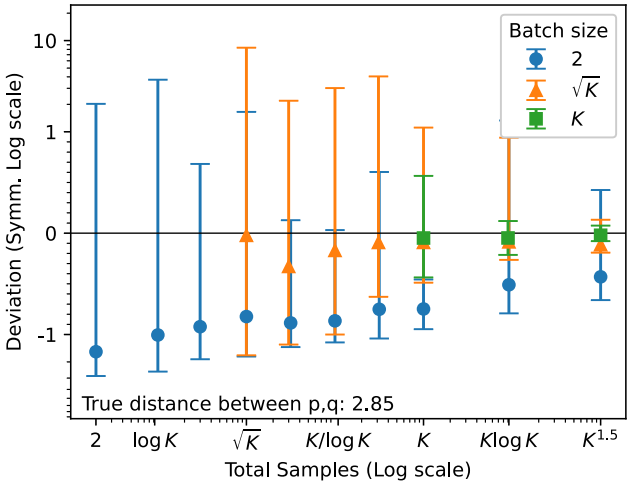
(h) Cross Entropy, p: Zipfian(1), q: Spiked Zipfian(1) (definition H.1.2)

Figure 2. The y-axis is the absolute error deviation the black-box loss value and the true distances (Squared distance and Cross Entropy). $K = 10,000$ for all trials. 30 trials for each parameter setting was recorded (including batch size). The horizontal bars represent the maximum absolute deviation of any of the 30 trials. The solid markers represent the average of the trials. Squared distance was estimated using the loss from claim 3. Cross entropy was estimated with a poisson sampling loss. Batch sizes were the same between draws from p and q.

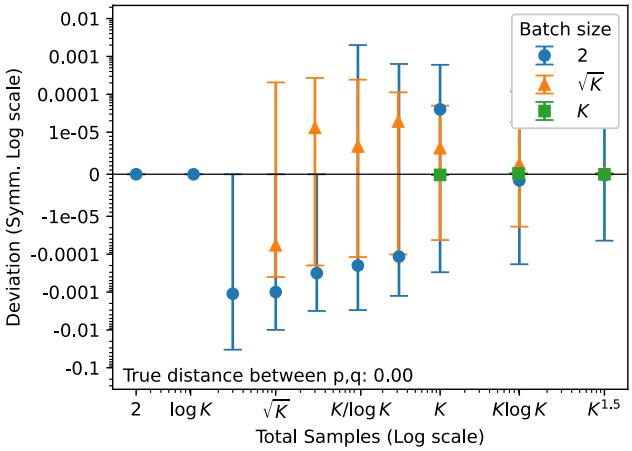
H.3. Two-sided deviation



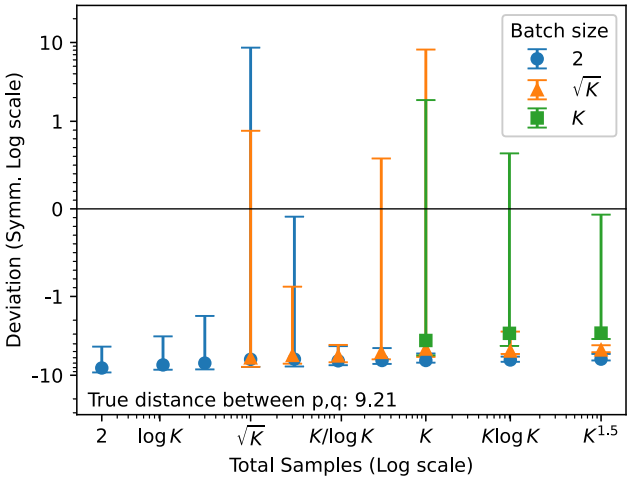
(a) Squared Distance, p: Zipfian(1), q: Zipfian(2)



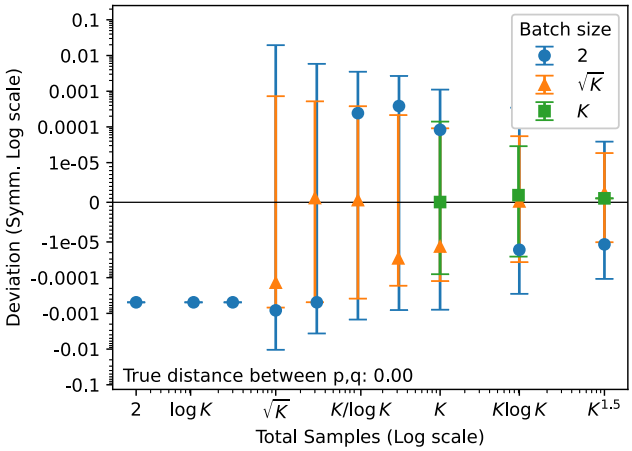
(b) Cross Entropy, p: Zipfian(1), q: Zipfian(2)



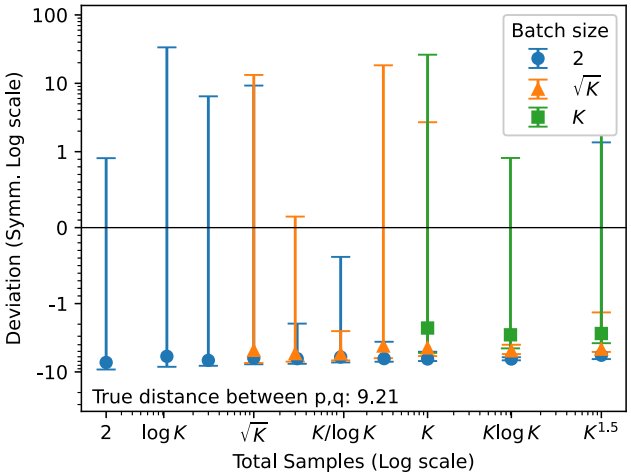
(c) Squared Distance, p: Uniform, q: Zipfian(0.01)



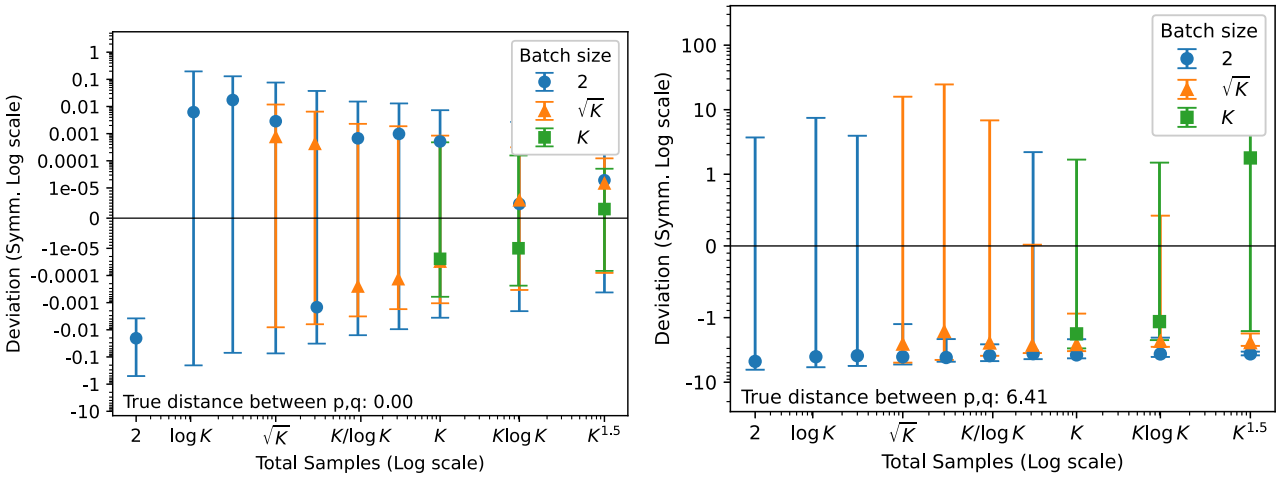
(d) Cross Entropy, p: Uniform, q: Zipfian(0.01)



(e) Squared Distance, p: Uniform,
q: Spiked Uniform (definition H.1.1)



(f) Cross Entropy, p: Uniform,
q: Spiked Uniform (definition H.1.1)

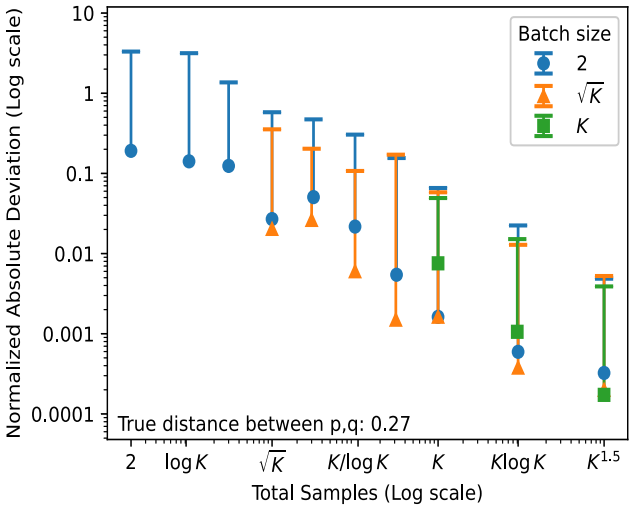


(g) Squared Distance, p: Zipfian(1), q: Spiked Zipfian(1) (definition H.1.2)

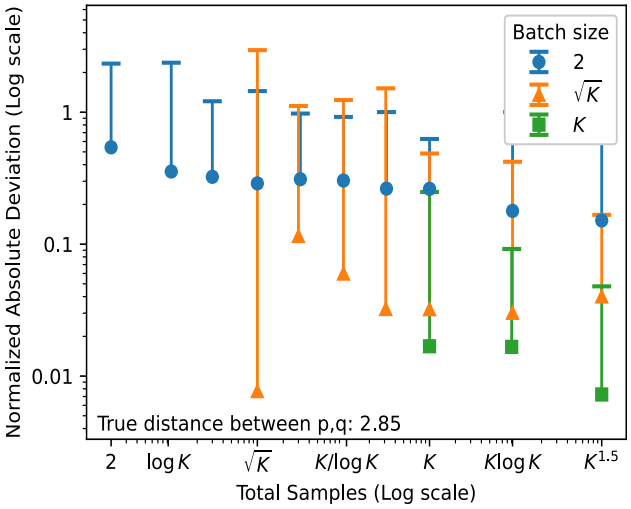
(h) Cross Entropy, p: Zipfian(1), q: Spiked Zipfian(1) (definition H.1.2)

Figure 3. The y-axis is the deviation between the black-box loss value and the true distances (Squared distance and Cross Entropy). $K = 10,000$ for all trials. 30 trials for each parameter setting was recorded (including batch size). The horizontal bars represent the maximum deviation on either side of any of the 30 trials. The solid markers represent the average of the trials. Squared distance was estimated using the loss from claim 3. Cross entropy was estimated with a poisson sampling loss. Batch sizes were the same between draws from p and q.

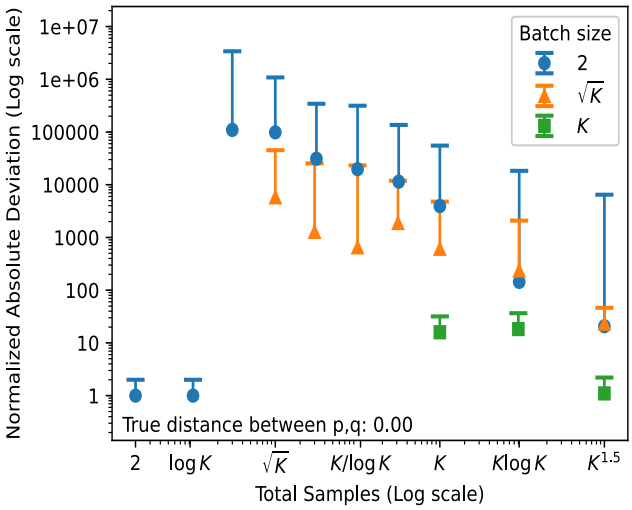
H.4. Normalized Absolute Deviation



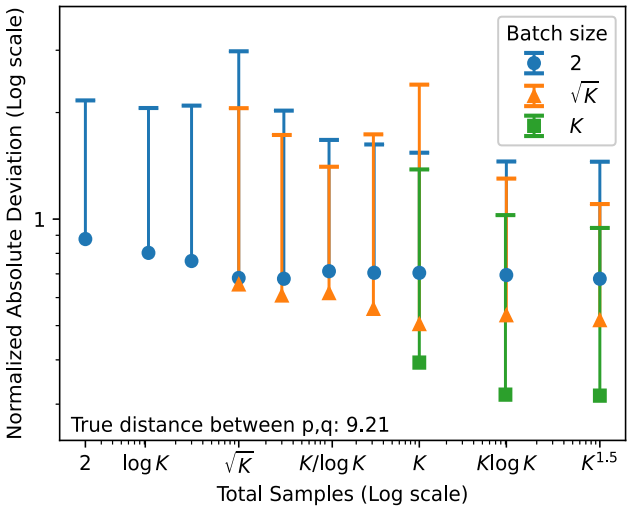
(a) Squared Distance, p: Zipfian(1), q: Zipfian(2)



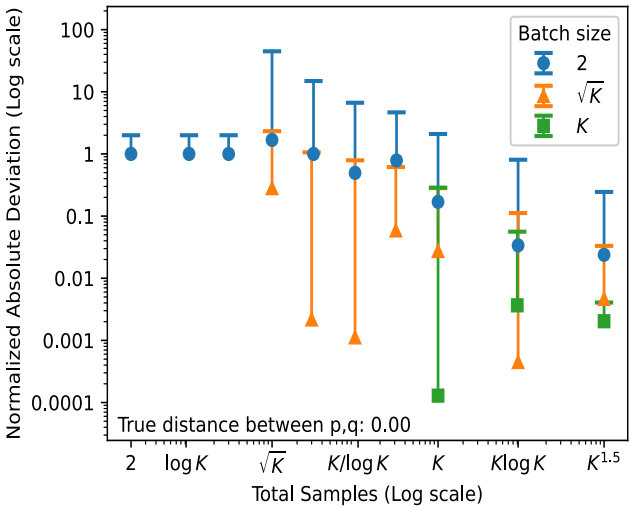
(b) Cross Entropy, p: Zipfian(1), q: Zipfian(2)



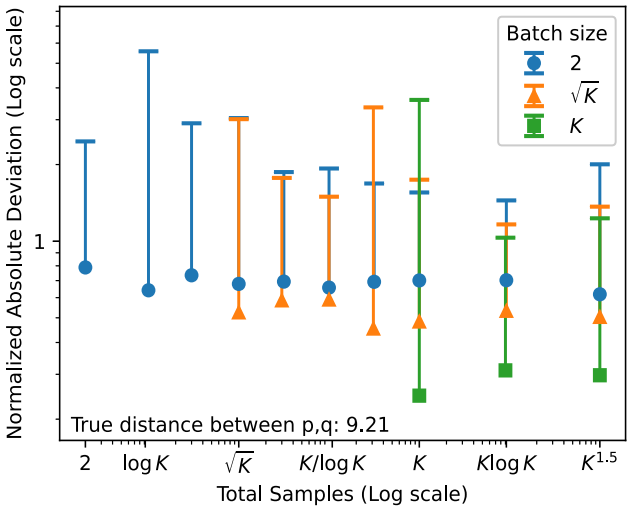
(c) Squared Distance, p: Uniform, q: Zipfian(0.01)



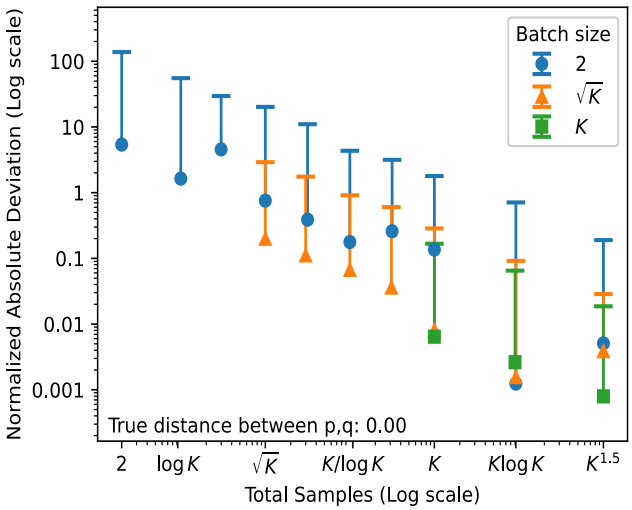
(d) Cross Entropy, p: Uniform, q: Zipfian(0.01)



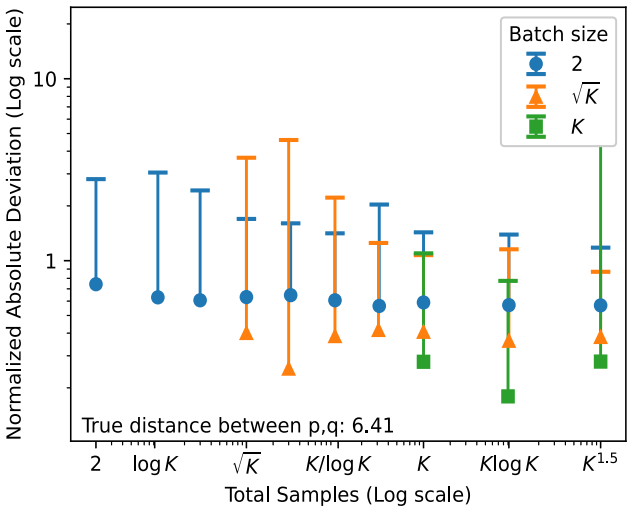
(e) Squared Distance, p: Uniform, q: Spiked Uniform (definition H.1.1)



(f) Cross Entropy, p: Uniform, q: Spiked Uniform (definition H.1.1)



(g) Squared Distance, p: Zipfian(1), q: Spiked Zipfian(1) (definition H.1.2)



(h) Cross Entropy, p: Zipfian(1), q: Spiked Zipfian(1) (definition H.1.2)

Figure 4. The y-axis is the normalized absolute deviation between the black-box loss value and the true distances (Squared distance and Cross Entropy). $K = 10,000$ for all trials. The deviations are normalized by the true distance between p and q. 30 trials for each parameter setting was recorded (including batch size). The horizontal bars represent the maximum normalized absolute deviations of any of the 30 trials. The solid markers represent the average of the trials. Note that when the squared distance is close to 0, the normalized error becomes very difficult to keep low. The un-normalized error in the previous section is more appropriate in this case. Squared distance was estimated using the loss from claim 3. Cross entropy was estimated with a poisson sampling loss. Batch sizes were the same between draws from p and q.