Toward a Characterization of Loss Functions for Distribution Learning (Full Version with Appendices)

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work we study loss functions for learning and evaluating probability distributions over large discrete domains. Unlike classification or regression where a wide variety of loss functions are used, in the distribution learning and density estimation literature, very few losses outside the dominant *log loss* are applied. We aim to understand this fact, taking an axiomatic approach to the design of loss functions for learning distributions. We start by proposing a set of desirable criteria that any good loss function should satisfy. Intuitively, these criteria require that the loss function faithfully evaluates a candidate distribution, both in expectation and when estimated on a few samples. Interestingly, we observe that *no loss function* possesses all of these criteria. However, one can circumvent this issue by introducing a natural restriction on the set of candidate distributions. Specifically, we require that candidates are *calibrated* with respect to the target distribution, i.e., they may contain less information than the target but otherwise do not significantly distort the truth. We show that, after restricting to this set of distributions, the log loss, along with a large variety of other losses satisfy the desired criteria. These results pave the way for future investigations of distribution learning that look beyond the log loss, choosing a loss function based on application or domain need.

1 Introduction

Estimating a probability distribution given independent samples from that distribution is a fundamental problem in machine learning and statistics [e.g. 25, 5, 26, 8]. In machine learning applications, the distribution of interest is often over a very large but finite sample space, e.g., the set of all English sentences up to a certain length or images of a fixed size in their RGB format.

A central technique in learning these types of distributions, encompassing, e.g., log likelihood maximization, is evaluation via a *loss function*. Given a distribution $\mathbf{p}$ over a set of outcomes $\mathcal{X}$ and a sample $x \sim \mathbf{p}$, a loss function $\ell(\mathbf{q}, x)$ evaluates the performance of a candidate distribution $\mathbf{q}$ in predicting x. Generally, $\ell(\mathbf{q}, x)$ will be higher if ℓ places smaller probability on x. Thus, in expectation over $x \sim \mathbf{p}$, the loss will be lower for candidate distributions that closely match $\mathbf{p}$.

The dominant loss applied in practice is the log loss ($\ell(\mathbf{q}, x) = \ln(1/q_x)$), which corresponds to log likelihood maximization. Surprisingly, few other losses are ever considered. This is in sharp contrast to other areas of machine learning, including in supervised learning where different applications have necessitated the use of different losses, such as the squared loss, hinge loss, ℓ_1 loss, etc. However, alternative loss functions can be beneficial for density estimation on large domains, as we show with a brief motivating example.

Motivating example. In many learning applications, one seeks to fit a complex distribution with a simple model that cannot fully capture its complexity. This includes e.g., noise tolerant or agnostic

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

Samples from $\mathbf{q}_1$	Samples from $\mathbf{q}_2$
brappost	to
hild	oneems
me	the
on	not
ther	of

$\texttt{log loss}(\mathbf{p}) = 7.45$
$\texttt{log log loss}(\mathbf{p}) = 1.91$
$\texttt{log loss}(\mathbf{q}_1) = 11.25$
$\texttt{log log loss}(\mathbf{q}_1) = 2.22$
$\texttt{log loss}(\mathbf{q}_2) = 12.26$
$\texttt{log log loss}(\mathbf{q}_2) = 2.18$

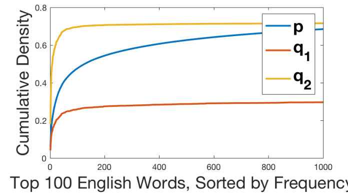

Figure 1: Modeling the distribution of English words, corrupted with 12% French and German words with character trigrams. Distribution $\mathbf{q}_1$ is trained by minimizing log loss. $\mathbf{q}_2$ achieves worse log loss but better *log log loss* and better performance at fitting the 'head' of the the target $\mathbf{p}$, indicating that log log loss may be more appropriate in this application. See Appendix G for more details.

learning. As an example, consider modeling the distribution over English words with a character trigram model. While this model, trained by minimizing log loss, fits the distribution of English words relatively well, its performance significantly degrades if the dataset includes a small fraction of foreign language words. The model is unable to fit the 'tail' of the distribution (corresponding to foreign words), however, in trying to do so it performs significantly worse on the 'head' of the distribution (corresponding to common English words). This is due to the fact that minimizing log loss requires q_x to not be much smaller than p_x for all x. A more robust loss function, such as the *log log loss*, $\ell(\mathbf{q}, x) = \ln(\ln(1/q_x))$, emphasizes the importance of fitting the 'head' and is less sensitive to the introduction of the foreign words. See Figure 1 for an illustration and Appendix G for details.

Loss function properties. In this paper, we start by understanding the desirable properties of log loss and seek to identity other loss functions with such properties that can have applications in various domains. A key characteristic of the log loss is that it is (strictly) *proper*. That is, the true underlying distribution $\mathbf{p}$ (uniquely) minimizes the expected loss on samples drawn from $\mathbf{p}$. Properness is essential for loss functions, as without it minimizing the expected loss leads to choosing an incorrect candidate distribution even when the target distribution is fully known. Log loss is also *local* (sometimes termed *pointwise*). That is, the loss of $\mathbf{q}$ on sample x is a function of the probability q_x and not of $q_{x'}$ for $x' \neq x$. Local losses are preferred in machine learning, where q_x is often implicitly represented as the output of a likelihood function applied to x, but where fully computing $\mathbf{q}$ requires at least linear time in the size of the sample space N and is infeasible for large domains, such as learning the distribution of all English sentences up to a certain length.

It is well-known that *log loss is the unique local and strictly proper loss function* [21, 24, 15]. Thus, requiring strict properness and locality already restricts us to using the log loss. At the same time, these restrictive properties are not sufficient for effective distribution learning, because

- A candidate distribution may be far from the target yet have arbitrarily close to optimal loss. Motivated by this problem, we define *strongly proper* losses that, if given a candidate is far from the target, will give an expected loss significantly worse than optimal.
- A candidate distribution might be far from the target, yet on a small number of samples, it may be likely to have smaller empirical loss than that of the target. This motivates our definition of *sample-proper* losses.
- On a small number of samples, the empirical loss of a distribution may be far from its expected loss, making evaluation impossible. This motivates our definition of *concentrating* losses.

Naively, it seems we cannot satisfy all our desired criteria: our only local strictly proper loss is the log loss, which in fact fails to satisfy the concentration requirement (see Example 4). We propose to overcome this challenge by restricting the set of candidate distributions, specifically to ones that satisfy the reasonable condition of *calibration*. We then consider the properties of loss functions on, not the set of all possible distributions, but the set of calibrated distributions.

Calibration and results. We call a candidate distribution $\mathbf{q}$ calibrated with respect to a target $\mathbf{p}$ if all elements to which $\mathbf{q}$ assigns probability α actually occur on average with probability α in the target distribution.[1] This can also be interpreted as requiring $\mathbf{q}$ to be a coarsening of $\mathbf{p}$, i.e., a calibrated distribution may contain less information than $\mathbf{p}$ but does otherwise not distort information. While for simplicity we focus on exactly calibrated distributions, in Appendix F we extend our results

[1]This definition is an adaptation of the standard calibration criterion applied to sequences of predictions made by a forecaster [11, 13]. See discussion in Appendix H.

to a natural notion of approximate calibration. Our main results show that the calibration constraint overcomes the impossibility of satisfying properness along with the our three desired criteria.

Main results (Informal summary)**.** *Any (local) loss* $\ell(\mathbf{q}, x) := f\left(\frac{1}{q_x}\right)$ *such that* f *is strictly concave and monotonically increasing has the following properties subject to calibration:*

1. ℓ *is* strictly proper*, i.e., the target distribution minimizes expected loss.*
2. *If* f *furthermore satisfies* left-strong-concavity*,* ℓ *is* strongly proper*, i.e., distributions far from the target have significantly worse loss.*
3. *If* f *furthermore grows relatively slowly,* ℓ *is* sample proper *i.e., on few samples, distributions far from the target have higher empirical loss with high probability.*
4. *Under these same conditions,* ℓ concentrates *i.e., on few samples, a distribution's empirical loss is a reliable estimate of its expected loss with high probability.*

The above criteria are formally introduced in Section 3. Each criteria is parameterized and different losses satisfy them with different parameters. We illustrate a few examples in Table 1 below. We emphasize that all losses shown below achieve relatively strong bounds, only depending polylogarithmically on the domain size N. Thus, we view all of these loss functions as viable alternatives to the log loss, which may be useful in different applications.

$\ell(\mathbf{q}, x)$	Strong Properness $\mathbb{E}\,\ell(\mathbf{q}; x) - \mathbb{E}\,\ell(\mathbf{p}; x)$	Concentration sample size $m(\gamma, N)$	Sample Properness sample size $m(\epsilon, N)$
$\ln \frac{1}{q_x}$	$\Omega(\epsilon^2)$	$\tilde{O}\left(\gamma^{-2} \ln\left(\frac{N}{\gamma}\right)^2\right)$	$O\left(\epsilon^{-4} (\ln N)^2\right)$
$\left(\ln \frac{1}{q_x}\right)^p$ for $p \in (0, 1]$	$\Omega\left(\epsilon^2 (\ln N)^{p-1}\right)$	$\tilde{O}\left(\gamma^{-2} \ln\left(\frac{N}{\gamma}\right)^{2p}\right)$	$O\left(\epsilon^{-4} (\ln N)^2\right)$
$\ln \ln \frac{1}{q_x}$	$\Omega\left(\frac{\epsilon^2}{\ln N}\right)$	$\tilde{O}\left(\gamma^{-2} \ln \ln\left(\frac{N}{\gamma}\right)^2\right)$	$O\left(\epsilon^{-4} (\ln \ln N)^2 (\ln N)^2\right)$
$\left(\ln \frac{e^2}{q_x}\right)^2$	$\Omega(\epsilon^2)$	$\tilde{O}\left(\gamma^{-2} \ln\left(\frac{N}{\gamma}\right)^4\right)$	$O\left(\epsilon^{-4} (\ln N)^4\right)$

Table 1: Examples of loss function that demonstrate strong properness, sample properness, and concentration, when restricted to calibrated distributions. In the above, N is the distributions support size, $\epsilon := \|\mathbf{p} - \mathbf{q}\|_1$ is the ℓ_1 distance between $\mathbf{p}$ and $\mathbf{q}$, and γ is an approximation parameter for concentration (see Section 4.2 for details). We assume for simplicity that $\epsilon \geq 1/N$ and hide dependencies on a success probability parameter for sample properness and concentration. $\tilde{O}(\cdot)$ suppresses logarithmic dependence on $1/\epsilon$ and $1/\gamma$.

1.1 Related work

Our work is directly inspired by applications of distribution estimation in very high-dimensional spaces, such as language modeling [20] . However, we do not know of work in this area that takes a systematic approach to designing loss functions.

A conceptually related research problem is that of learning distributions using computationally and statistically efficient algorithms. In addition to loss function minimization, a number of general-purpose methods have been proposed for this problem, including using histograms, nearest neighbor estimators, etc. See [17] for a survey of these methods. Much of the work in this space focuses on learning *structured* or *parametric* distributions [10, 18, 19, 9], e.g., monotone distributions or mixtures of Gaussians. On the other hand, learning an unstructured discrete distribution with support size N within an ℓ_1 distance of ϵ requires $\mathrm{poly}(N, 1/\epsilon)$ samples. Thus, works in this space typically focus on designing computationally efficient algorithms for optimal estimation using large sample sets [26]. In comparison, we focus on unstructured distributions with prohibitively large supports and characterize loss functions that only require $\mathrm{polylog}(N)$ sample complexity to estimate. We do not introduce a general algorithm for distribution learning — as any such algorithm would require $\Omega(N)$ samples. Rather, motivated by tailored algorithms used in complex domains such as natural language processing, our work characterizes loss functions that could be used by a variety of algorithms.

Outside distribution learning, loss functions (termed *scoring rules*) have been studied for decades in the information elicitation literature, which seeks to incentivize experts, such as weather forecasters, to give accurate predictions [e.g. 7, 16, 24, 14, 15]. The notion of loss function properness, for example, comes from this literature. Recent research has made some connections between information elicitation and loss functions in machine learning; however, it has focused mostly on the classification and regression and not distribution learning [4, 14, 22, 23, 12]. Our work can be viewed as a contribution to the literature on evaluating forecasters by showing that, if the forecaster is constrained to be calibrated, then a variety of simple local loss functions become (strongly, sample) proper.

2 Preliminaries

We work with distributions over a finite domain $\mathcal{X}$ with $|\mathcal{X}| = N$. The set of all distributions over $\mathcal{X}$ is denoted by $\Delta_{\mathcal{X}}$. We denote a distribution $\mathbf{p} \in \{0,1\}^N$ over $\mathcal{X}$ by a vector of probabilities, where p_x is the probability $\mathbf{p}$ places on $x \in \mathcal{X}$. For any set $B \subseteq \mathcal{X}$, the total probability $\mathbf{p}$ places on B is denoted by $\mathbf{p}(B) := \sum_{x \in B} p_x$. We use X to denote a random variable on $\mathcal{X}$ whose distribution is specified in context. We also consider point mass distributions $\boldsymbol{\delta}^x \in \Delta_{\mathcal{X}}$ where $\boldsymbol{\delta}^x_{x'} = \mathbf{1}[x = x']$.

Throughout this paper, we typically use $\mathbf{p}$ to denote the true (or target) distribution and $\mathbf{q}$ to denote a candidate or predicted distribution. For any two distributions $\mathbf{p}$ and $\mathbf{q}$, the *total variation distance* between them is defined by $\mathrm{TV}(\mathbf{p},\mathbf{q}) := \sup_{B \subseteq \mathcal{X}} \mathbf{p}(B) - \mathbf{q}(B) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1$, where $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector. Together, ℓ_1 and the total variation distance are two of the most widely used measures of distance between distributions.

To measure the quality of a candidate distribution $\mathbf{q}$ given samples from $\mathbf{p}$, machine learning typically turns to loss functions. A *loss function* is a function $\ell : \Delta_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}$ where $\ell(\mathbf{q}, x)$ is the loss assigned to candidate $\mathbf{q}$ on outcome x. Given a target distribution $\mathbf{p}$, the *expected loss* for candidate $\mathbf{q}$ is defined as $\ell(\mathbf{q}; \mathbf{p}) := \mathbb{E}_{X \sim \mathbf{p}}\left[\ell(\mathbf{q}, X)\right]$. A loss function is called *proper* if $\ell(\mathbf{p}; \mathbf{p}) \le \ell(\mathbf{q}; \mathbf{p})$ for all $\mathbf{p} \neq \mathbf{q}$, and *strictly proper* if the inequality is always strict[2]. Two common examples of proper loss functions are the *log loss* function $\ell(\mathbf{q}, x) = \ln(\frac{1}{q_x})$ (with the logarithm always taken base e in this paper) and the *quadratic loss* $\ell(\mathbf{q}, x) = \frac{1}{2}\|\boldsymbol{\delta}^x - \mathbf{q}\|_2^2$. A loss function ℓ is called *local* if $\ell(\mathbf{q}, x)$ is a function of q_x alone. For example, the log loss is local while the quadratic loss is not.

Our main results will be are characterized by the topology of the loss functions we consider.

For simplicity, we will generally assume functions are differentiable, although our results can be extended.

Definition 1 (Strongly Concave)**.** A function $f : [0, \infty] \to \mathbb{R}$ is β*-strongly concave* if for all z, z' in the domain of f, $f(z) \le f(z') + \nabla f(z') \cdot (z - z') - \frac{\beta}{2}(z - z')^2$.

We also consider a relaxation of strong concavity that helps us in analyzing functions that have a large curvature close to the origin but flatten out as we move farther from it.

Definition 2 (Left-Strongly Concave)**.** A function $f : [0, \infty] \to \mathbb{R}$ is $\beta(z)$*-left-strongly concave* if the function restricted to $[0, z]$ is $\beta(z)$-strongly concave, for all z.

As discussed, a natural assumption on the set of candidate distributions is *calibration*. Formally:

Definition 3 (Calibration)**.** Given a distribution $\mathbf{q} \in \Delta_{\mathcal{X}}$, let $B_t(\mathbf{q}) = \{x : q_x = t\}$. When it is clear from the context, we suppress $\mathbf{q}$ in the definition of B_t. We say that $\mathbf{q}$ is *calibrated with respect to* $\mathbf{p}$, if $\mathbf{q}(B_t(\mathbf{q})) = \mathbf{p}(B_t(\mathbf{q}))$ for all $t \in [0, 1]$. We let $\mathcal{C}(\mathbf{p})$ denote the set of all calibrated distributions with respect to $\mathbf{p}$.

In other words, $\mathbf{q}$ is calibrated with respect to $\mathbf{p}$ if points assigned probability $q_x = t$ have average probability t under $\mathbf{p}$. In other words, $\mathbf{p}$ can be "coarsened" to $\mathbf{q}$ by taking subsets of points and replacing their probabilities with the subset average. Note that the uniform distribution $\mathbf{q} = (\frac{1}{N}, \ldots, \frac{1}{N})$ is calibrated with respect to all $\mathbf{p}$, and that $\mathbf{p}$ is calibrated with respect to itself. Also note that there are only finitely many values $t \in [0, 1]$ for which B_t is non-empty. We denote the set of these values by $T(\mathbf{q}) = \{t : B_t \neq \emptyset\}$.

We refer an interested reader to Appendix H for a more detailed discussion of the notion of calibration and its connections to similar notions used in forecasting theory, e.g. [11, 13]. See Appendix F for a discussion of how our results can be extended to a natural notion of approximate calibration.

[2]Our use of "properness" is inspired the literature on *proper scoring rules*. It is not to be confused with "properness" in learning theory where the learned hypothesis must belong to a pre-determined class of hypotheses.

3 Three Desirable Properties of Loss Functions

In this section, we define three criteria and discuss why any desirable loss function should demonstrate them. We use examples of loss functions, such as the log loss $\ell_{log\text{-}loss}(\mathbf{q}, x) = \ln(\frac{1}{q_x})$ and the linear loss $\ell_{lin\text{-}loss}(\mathbf{q}, x) = -q_x$ to help demonstrate the existence or lack of these criteria.

3.1 Strong Properness

Recall that a loss function is strictly proper if all incorrect candidate distributions yield a higher expected loss value than the target distribution. Here, we expand this to *strong* properness where this gap in expected loss grows with distance from the target distribution. We also extend both definitions to hold over a specific domain of candidate distributions, rather than all distributions.

Definition 4 (Calibrated Properness)**.** Let $\mathcal{P} : \Delta_{\mathcal{X}} \to 2^{\Delta_{\mathcal{X}}}$ be a *domain function*, that is, $\mathcal{P}(\mathbf{p}) \subseteq \Delta_{\mathcal{X}}$ is a restricted set of distributions. A loss function ℓ is *proper over* $\mathcal{P}$ if for all $\mathbf{p} \in \Delta_{\mathcal{X}}$, $\mathbf{p} \in \operatorname{argmin}_{\mathbf{q} \in \mathcal{P}(\mathbf{p})} \ell(\mathbf{q}; \mathbf{p})$. A loss function is said to be *strictly proper over* $\mathcal{P}$ if the argmin is always unique. When $\mathcal{P}(\mathbf{p}) = \mathcal{C}(\mathbf{p})$, i.e. is the set of calibrated distributions w.r.t. $\mathbf{p}$, we call such a loss function *(strictly) calibrated proper*.

Example 1. It is well-known that $\ell_{log\text{-}loss}(\mathbf{q}, x) = \ln\left(\frac{1}{q_x}\right)$ is the *unique local* proper loss function (up to scaling) over the unrestricted domain $\mathcal{P}(\mathbf{p}) = \Delta_{\mathcal{X}}$ [6]. Indeed, it is known that the difference in expected log loss of a prediction $\mathbf{q}$ and the target distribution $\mathbf{p}$ is the KL-divergence, i.e.

$$\ell_{log\text{-}loss}(\mathbf{q}; \mathbf{p}) - \ell_{log\text{-}loss}(\mathbf{p}; \mathbf{p}) = \mathrm{KL}(\mathbf{p}, \mathbf{q}) := \sum_x p_x \ln\left(\frac{p_x}{q_x}\right). \tag{1}$$

Furthermore, the KL-divergence is strictly positive for $\mathbf{p} \neq \mathbf{q}$. This proves that the log loss is strictly proper over $\Delta_{\mathcal{X}}$, and as a result, is strictly calibrated proper as well.

On the other hand, $\ell_{lin\text{-}loss}(\mathbf{q}, x) = -q_x$ is not proper over $\Delta_{\mathcal{X}}$. This is due to that fact that the minimizer of this loss is the point mass distribution $\boldsymbol{\delta}^x$ for $x = \operatorname{argmax}_x p_x$. For example, for target distribution $\mathbf{p} = (\frac{1}{3}, \frac{2}{3})$ distribution $\mathbf{q} = (0, 1)$ yields a lower $\ell_{lin\text{-}loss}$ than that of $\mathbf{p}$. Note, however, that such a choice of $\mathbf{q}$ is not calibrated with respect to $\mathbf{p}$. When loss minimization is constrained to the set of calibrated distribution $\mathcal{C}(\mathbf{p}) = \{(\frac{1}{3}, \frac{2}{3}), (\frac{1}{2}, \frac{1}{2})\}$, $\mathbf{p}$ minimizes the expected linear loss. Indeed, in Section 4 we show more generally that the linear loss and in fact many reasonable local loss functions are calibrated proper.

While strict properness is an important baseline guarantee, we would like a "stronger" property: If $\mathbf{q}$ is significantly incorrect in the sense of being far from $\mathbf{p}$, then the expected loss of $\mathbf{q}$ should be significantly worse. This motivates the following definition.

Definition 5 (Strong Calibrated Properness)**.** A loss function ℓ is *β-strongly proper over a domain function* $\mathcal{P}$ if for all $\mathbf{p} \in \Delta_{\mathcal{X}}$, for all $\mathbf{q} \in \mathcal{P}(\mathbf{p})$, $\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) \geq \frac{\beta}{2} \|\mathbf{p} - \mathbf{q}\|_1^2$. When $\mathcal{P}(\mathbf{p}) = \mathcal{C}(\mathbf{p})$, we call such functions *β-strongly calibrated proper* and when $\mathcal{P}(\mathbf{p}) = \Delta_{\mathcal{X}}$, we simply refer to them as *β-strongly proper*.

Example 2. The log loss is 1-strongly proper. This is *equivalent* to Pinsker's inequality, which states that *for all* $\mathbf{p}$ *and* $\mathbf{q}$, $\mathrm{KL}(\mathbf{p}, \mathbf{q}) \geq 2\mathrm{TV}(\mathbf{p}, \mathbf{q})^2$. Together with (1) and the fact that $\mathrm{TV}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1$, this shows that log loss is 1-strongly proper (and thus also 1-strongly calibrated proper.)

As we will see in Section 4, strong calibrated properness relates to the notion of strong concavity (of the inverse loss function) in ℓ_1 norm. We refer the interested reader to Appendix I for a discussion of the use of alternative norms in the definition of strong properness. In Appendix J we extend the study of normed concavity of loss functions to strong properness of a loss function over $\Delta_{\mathcal{X}}$.

3.2 Sample-properness

So far, we have focused on the loss a candidate $\mathbf{q}$ receives in *expectation over* $x \sim \mathbf{p}$. Of course, if one is attempting to learn $\mathbf{p}$, this expectation can generally not be computed. We would like the notion of properness to carry over to the setting when the loss on $\mathbf{q}$ is estimated using a small set of samples from $\mathbf{p}$. We say that a loss function is sample-proper if within a small number, all candidate distributions that are sufficiently far from $\mathbf{p}$ yield a loss that is larger than that of $\mathbf{p}$ on the samples.

In the remainder of this paper, let $\hat{\mathbf{p}}$ denote the empirical distribution corresponding to samples drawn from $\mathbf{p}$. Note that the average loss of any $\mathbf{q}$ on the samples can be written $\ell(\mathbf{q}; \hat{\mathbf{p}})$. Formally:

Definition 6 (Calibrated Sample-Properness)**.** A loss function ℓ is $m(\epsilon, \delta, N)$-*sample proper* over a function domain $\mathcal{P}$ if, for all $\mathbf{p} \in \Delta_{\mathcal{X}}$ and all $\mathbf{q} \in \mathcal{P}(\mathbf{p})$ with $\|\mathbf{p}-\mathbf{q}\|_1 \geq \epsilon$, with probability at least $1-\delta$ over $m(\epsilon, \delta, N)$ i.i.d. samples from $\mathbf{p}$, we have $\ell(\mathbf{p}; \hat{\mathbf{p}}) < \ell(\mathbf{q}; \hat{\mathbf{p}})$. When $\mathcal{P}(\mathbf{p}) = \mathcal{C}(\mathbf{p})$, we call such functions *calibrated* $m(\epsilon, \delta, N)$*-sample proper.*

Example 3. A folklore theorem states that $\ell_{log\text{-}loss}$ is $O\left(\frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right)$-sample proper over $\Delta_{\mathcal{X}}$, and as a result it is calibrated $O\left(\frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right)$-sample proper.

Now consider $\ell_{lin\text{-}loss}(\mathbf{q}, x) = -q_x$. Since it is not a proper loss function over $\Delta_{\mathcal{X}}$, by definition it is not sample proper over $\Delta_{\mathcal{X}}$ for any $m(\epsilon, \delta, N)$. When restricting to calibrated distributions however, as we claimed in Example 1 linear loss is calibrated proper in expectation. It is interesting to note that linear loss is not sample proper for any $m(\epsilon, \delta, N) \in o\left(N^2\right)$. To observe this, consider $\mathbf{p}$ where $p_1 = \frac{1}{4} + \frac{1}{2\sqrt{m}}$, $p_2 = \frac{1}{4} - \frac{1}{2\sqrt{m}}$, and $p_x = \frac{1}{2(N-2)}$ for all $x = 3, \ldots, N$. Consider $\mathbf{q}$ where $q_1 = q_2 = \frac{1}{4}$ and $q_x = \frac{1}{2(N/2-2)}$ for $x = 3, \ldots, N/2$. Let $x_1, \ldots, x_m$ be the samples drawn from $\mathbf{p}$. Then, with a constant probability $m(\frac{1}{4} + \frac{1}{2\sqrt{m}}) \pm \sqrt{m}$ and $m(\frac{1}{4} - \frac{1}{2\sqrt{m}}) \pm \sqrt{m}$ number of these samples are on instances $x = 1$ and $x = 2$, respectively. Therefore, with a constant probability

$$\ell(\mathbf{q}; \hat{\mathbf{p}}) - \ell(\mathbf{p}; \hat{\mathbf{p}}) = \frac{1}{m}\sum_{i=1}^{m}(p_i - q_i) \leq \sqrt{m} \pm 2\sqrt{m} - \Theta\left(\frac{1}{N}\right) < 0,$$

when $m \in o\left(N^2\right)$. Furthermore, note that $\mathbf{q}$ is calibrated w.r.t. $\mathbf{p}$ and $\|\mathbf{p}-\mathbf{q}\|_1 = \Theta(1)$. Thus, for $\ell_{lin\text{-}loss}$ to be calibrated $m(\epsilon, \delta, N)$-sample proper, we must have $m(\Theta(1), \Theta(1), N) \in \Omega\left(N^2\right)$.

3.3 Concentration

Beyond sample properness, when the expected loss $\ell(\mathbf{q}; \mathbf{p})$ is estimated from a small i.i.d. sample from $\mathbf{p}$, we would like the empirical loss to remain faithful to the true value. For example, one might hope that minimizing loss on that sample will result in a distribution that has small loss on $\mathbf{p}$. This will hold as long as the empirical loss well approximates the true expected loss with high probability.

Definition 7 (Calibrated Concentration)**.** A loss function ℓ *concentrates over domain function* $\mathcal{P}$ *with* $m(\gamma, \delta, N)$ *samples* if for all $\mathbf{p} \in \Delta_{\mathcal{X}}$, for all $\mathbf{q} \in \mathcal{P}(\mathbf{p})$, for $m(\gamma, \delta, N)$ i.i.d. samples from $\mathbf{p}$, $\Pr\left[|\ell(\mathbf{q}; \hat{\mathbf{p}}) - \ell(\mathbf{q}; \mathbf{p})| \geq \gamma\right] \leq \delta$. When $\mathcal{P}(\mathbf{p}) = \mathcal{C}(\mathbf{p})$, we say that ℓ *calibrated concentrates* with $m(\gamma, \delta, N)$ samples.[3]

Example 4. We can easily see that log loss does *not* concentrate with $o(N)$ samples over $\Delta_{\mathcal{X}}$. Let $\mathbf{p}$ be the uniform distribution and $\mathbf{q}$ be uniform on $\mathcal{X} \setminus \{x\}$. With high probability, x is not sampled, and $\ell(\mathbf{q}; \hat{\mathbf{p}})$ is finite. Yet $\ell(\mathbf{q}; \mathbf{p}) = \infty$. Note that although this example is extreme, its conclusion is robust: one can make an arbitrarily large finite gap. As we will see, the log loss, along with many other reasonable loss will concentrate with a small number of samples over calibrated distributions.

4 The Main Results

Looking back at the criteria defined in Section 3, we are immediately faced with an impossibility result: no local loss function exists that satisfies properness, $o(N)$-sample properness, and concentration with $o(N)$ samples. This is because log loss is the unique local loss function that satisfies the first property and as shown in Example 4 it does not concentrate. In this section, we show that a broad class of local loss functions with certain niceness properties satisfies the above three criteria over calibrated domains. Specifically, we consider loss functions $\ell(\mathbf{q}, x)$ that are non-increasing in q_x and are inversely concave: $\ell(\mathbf{q}, x) = f(\frac{1}{q_x})$ for some concave function f. Similarly, we say that ℓ is inversely strongly concave if the corresponding f is strongly concave.

4.1 Calibrated and Strong Calibrated Properness

In this section, we show that any (strongly) nice loss function is (strongly) proper over the domain of calibrated distributions. More formally.

Theorem 1 (Strict Properness)**.** *Suppose the local loss function ℓ is such that $\ell(\mathbf{q}, x) = f(\frac{1}{q_x})$ for a concave f function. Then, ℓ is strictly proper over the domain function $\mathcal{C}$.*

[3] We use γ to denote difference in loss to avoid confusion with ϵ, which generally means a distance between distributions.

Theorem 2 (Strong Properness)**.** *Suppose the loss function ℓ is such that $\ell(\mathbf{q}, x) = f(\frac{1}{q_x})$ where f is non-decreasing and is $\frac{C(x)}{x^2}$-left-strongly concave where $C(x)$ is non-increasing and non-negative for $x \geq 1$. Then for all $\mathbf{p} \in \Delta_{\mathcal{X}}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$,*

$$\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) \geq C\left(\frac{4N}{\|\mathbf{p} - \mathbf{q}\|_1}\right) \cdot \frac{\|\mathbf{p} - \mathbf{q}\|_1^2}{128}.$$

We defer the proof of Theorem 2 to Appendix B.1 and only prove Theorem 1 here. To help us with this proof, let us first understand an a key property of calibration in the next lemma, whose proof appears in Appendix A.1. At a high level, this lemma shows that the average value of $1/p_x$ and $1/q_x$ is the same over instances x such that $q_x = t$, which is also equal to $1/t$.

Lemma 1. *For any distribution $\mathbf{p} \in \Delta_{\mathcal{X}}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$, and for any $t \in [0, 1]$, we have $\mathbb{E}_{X \sim \mathbf{p}}\left[\frac{1}{p_X} \;\middle|\; X \in B_t\right] = \frac{1}{t}$, where $B_t = \{x : q_x = t\}$.*

Proof of Theorem 1. Suppose $\ell(\mathbf{q}, x) = f(\frac{1}{q_x})$ for a strictly concave f. Consider any $\mathbf{q}$ that is calibrated with respect to $\mathbf{p}$. Recall that $B_t = \{x : q_x = t\}$ and $T(\mathbf{q}) = \{t : |B_t| \neq \emptyset\}$ is a finite set.

$$\ell(\mathbf{p}; \mathbf{p}) = \sum_{t \in T(\mathbf{q})} \mathbf{p}(B_t)\, \mathbb{E}\left[f\left(\frac{1}{p_X}\right) \;\middle|\; X \in B_t\right] \leq \sum_{t \in T(\mathbf{q})} \mathbf{p}(B_t) f\left(\mathbb{E}\left[\frac{1}{p_X} \;\middle|\; X \in B_t\right]\right)$$
$$= \sum_{t \in T(\mathbf{q})} \mathbf{p}(B_t) f\left(\frac{1}{t}\right) = \sum_{t \in T(\mathbf{q})} \sum_{x \in B_t} p_x f\left(\frac{1}{q_x}\right) = \ell(\mathbf{q}; \mathbf{p}),$$

where the second transition is by Jensen's inequality and the third transition is by Lemma 1. If f is strictly concave and there exists a B_t where $\mathbf{q}$ and $\mathbf{p}$ disagree, then the inequality is strict. □

4.2 Concentration

The (strong) properness of a loss function, as discussed in Section 4.1, is only concerned with loss functions in expectation. In this section, we consider finite sample guarantees. Recall that ℓ concentrates over $\mathcal{P}(\mathbf{p})$ (Definition 7) if, with $m(\gamma, \delta, N)$ samples, the empirical loss $\ell(\mathbf{q}; \hat{\mathbf{p}})$ of a distribution $\mathbf{q} \in \mathcal{P}(\mathbf{p})$ is γ-close to its true loss $\ell(\mathbf{q}; \mathbf{p})$ with probability $1 - \delta$. Concentration can be difficult to achieve: by Example 4, even the log loss does not concentrate for any sample size $o(N)$ for general $\mathbf{q} \in \Delta_{\mathcal{X}}$. However, as we show below, when $\mathbf{q}$ is *calibrated*, many natural loss functions, including log loss, indeed concentrate. All that is needed is that the loss function is inverse concave, increasing, and does not grow too quickly as $q_x \to 0$.

Theorem 3 (Concentration)**.** *Suppose ℓ is a local loss function with $\ell(\mathbf{q}, x) = f\left(\frac{1}{q_x}\right)$ for nonnegative, increasing, concave $f(z)$. Suppose further that $f(z) \leq c\sqrt{z}$ for all $z \geq 1$ and some constant c. Then ℓ concentrates over the domain function $\mathcal{C}$ for any $m(\gamma, \delta, N) \leq N$, such that*

$$m(\gamma, \delta, N) \geq \frac{c_1 \cdot f\,(\beta)^2 \ln \frac{1}{\delta}}{\gamma^2},$$

where c_1 is a fixed constant and $\beta := \frac{16N^8}{\delta \cdot \min(1, \gamma^2/c^2)}$. That is, for any $\mathbf{p} \in \Delta_{\mathcal{X}}, \mathbf{q} \in \mathcal{C}(\mathbf{p})$, drawing at least $m(\gamma, \delta, N)$ samples guarantees $|\ell(\mathbf{q}; \hat{\mathbf{p}}) - \ell(\mathbf{q}; \mathbf{p})| \leq \gamma$ with probability $\geq 1 - \delta$.

Note that γ bounds the absolute difference between $\ell(\mathbf{q}; \hat{\mathbf{p}})$ and $\ell(\mathbf{q}; \mathbf{p})$. The desired difference may depend on the relative scale of the loss function. If e.g., we take $\ell(\mathbf{q}, x)$ and scale to obtain $\ell'(\mathbf{q}, x) = \alpha \cdot \ell(\mathbf{q}, x)$ for some α, the desired error γ scales by α, $f(\beta)$ and c both scale by α, and thus we can see that the sample complexity remains fixed.

We defer the proof of Theorem 3 to Appendix C. At a high level, calibration helps us avoid worst-case instances (as in Example 4) using a very simple fact: when $\mathbf{q}$ is calibrated, we have $\frac{q_x}{p_x} \geq \frac{1}{N}$ for all x (see Lemma 3). This rules out very low probability events that contribute significantly to $\ell(\mathbf{q}; \mathbf{p})$ but require many samples to identify. The main idea in proving Theorem 3 is to partition $\mathcal{X}$ into Ω containing elements of very small probability, and $\mathcal{X} \setminus \Omega$. With high probability, no element of Ω is ever sampled from $\mathbf{p}$. Conditioned on this, the loss is bounded (and its expectation does not change much), so a concentration result can be applied.

4.3 Sample Properness

Lastly, we turn our attention to calibrated sample properness. Recall that a loss function is sample proper if all candidate distributions that are sufficiently far from $\mathbf{p}$ have a loss that is larger $\mathbf{p}$ on the empirical distribution $\hat{\mathbf{p}}$ corresponding to a small number of samples from $\mathbf{p}$. It is not hard to see that sample properness of a loss function is a direct consequence of its concentration and strong properness. For any candidate distribution $\mathbf{q}$ for which $\|\mathbf{q}-\mathbf{p}\|_1$ is large, strong properness (Theorem 2) implies that $\ell(\mathbf{q};\mathbf{p})$ is significantly larger than $\ell(\mathbf{p};\mathbf{p})$. Furthermore, concentration (Theorem 3) implies that with high probability $\ell(\mathbf{q};\mathbf{p}) \approx \ell(\mathbf{q};\hat{\mathbf{p}})$ and $\ell(\mathbf{p};\mathbf{p}) \approx \ell(\mathbf{p};\hat{\mathbf{p}})$. Therefore, with high probability, $\ell(\mathbf{q};\hat{\mathbf{p}}) > \ell(\mathbf{p};\hat{\mathbf{p}})$. Formally in Appendix D we prove:

Theorem 4 (Sample properness)**.** *Suppose ℓ is a local loss function with $\ell(\mathbf{q},x) = f(\frac{1}{q_x})$ for nonnegative, increasing, concave $f(z)$. Suppose further that $f(z) \le c\sqrt{z}$ for all $z \ge 1$ and some constant c and that f is $\frac{C(x)}{x^2}$-left-strongly concave for where $C(x)$ is nonincreasing and nonnegative for $x \ge 1$. Then for all $\mathbf{p} \in \Delta_{\mathcal{X}}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$, if $\hat{\mathbf{p}}$ is the empirical distribution constructed from m independent samples of $\mathbf{p}$ with $m \le N$ and*

$$m \ge \frac{c_1 \cdot f(\beta)^2 \ln \frac{1}{\delta}}{\left(C\left(\frac{4N}{\|\mathbf{p}-\mathbf{q}\|_1}\right)\|\mathbf{p}-\mathbf{q}\|^2\right)^2},$$

where c_1 is constant and $\beta := \frac{288N^8}{\delta \cdot \min\left(1, \left[C\left(\frac{4N}{\|\mathbf{p}-\mathbf{q}\|_1}\right)\frac{\|\mathbf{p}-\mathbf{q}\|_1^2}{128c}\right]^2\right)}$, then $\ell(\mathbf{q};\hat{\mathbf{p}}) > \ell(\mathbf{p};\hat{\mathbf{p}})$ with prob. $\ge 1-\delta$.

4.4 Application of the Main Results to Loss Functions

We now instantiate Theorems 2, 3, and 4 for one example of a natural loss function $\ell(\mathbf{q},x) = \ln\ln(\frac{1}{q_x})$. Refer to Table 1 for other loss functions and see Appendix E for details on its derivation.

First, note that $\ln\ln(z)$ is $C(z)/z^2$-left-strongly concave for $C(z) = \frac{(1+\ln(z))}{\ln(z)^2}$.[4] Moreover, $C(z)$ is non-increasing and non-negative for $z \ge 1$ and $\ln\ln(z) \le \sqrt{z}$. Using these, for any $\mathbf{p}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$ such that $\|\mathbf{p}-\mathbf{q}\|_1 \ge \epsilon$ we have

- By Theorem 2, $\ell(\mathbf{q};\mathbf{p}) - \ell(\mathbf{p};\mathbf{p}) \ge \Omega(\frac{\epsilon^2}{\ln(N/\epsilon)})$.
- By Theorem 3, an empirical distribution $\hat{\mathbf{p}}$ of $\tilde{O}\left(\gamma^{-2}\ln\ln(N)^2\ln(1/\delta)\right)$ i.i.d samples from $\mathbf{p}$ is sufficient such that $|\ell(\mathbf{q};\hat{\mathbf{p}}) - \ell(\mathbf{q};\mathbf{p})| \le \gamma$ with probability $1-\delta$.
- By Theorem 4, an empirical distribution $\hat{\mathbf{p}}$ of $\tilde{O}\left(\epsilon^{-4}\ln\ln(N\ln(N))^2\ln(1/\delta)\ln(N)\right)$ i.i.d samples from $\mathbf{p}$ is sufficient such that $\ell(\mathbf{q};\hat{\mathbf{p}}) > \ell(\mathbf{p};\hat{\mathbf{p}})$ with probability $1-\delta$.

5 Discussion

In this work, we characterized loss functions that meet three desirable properties: properness in expectation, concentration, and sample properness. We demonstrated that no local loss function meets all of these properties over the domain of all candidate distributions. But, if one enforces the criterion of *calibration* (or approximate calibration as discussed in Appendix F), then many simple loss functions have good properties for evaluating learned distributions over large discrete domains. We hope that our work provides a starting point for several future research directions.

One natural question is to understand how to select a loss function based on the application domain. Our example for language modeling, from the introduction, motivates the idea that log loss is not the best choice always. Understanding this more formally, for example in the framework of robust distribution learning, could provide a systematic approach for selecting loss functions based on the needs of the domain. Our work also leaves open the question of designing compuationally and statistically efficient learning algorithms for different loss functions under the constraint that the candidate $\mathbf{q}$ is (approximately) calibrated. One challenge in designing computationally efficient algorithms is that the space of calibrated distributions is not convex. We present some advances towards dealing with this challenge in Appendix F by providing an efficient procedure for ‘projecting’ a non-calibrated distribution on the space of approximately calibrated distribution. It remains to be seen if iteratively applying this procedure could be useful in designing an efficient algorithm for minimizing the loss on calibrated distributions.

[4]In Appendix E, we show that function f is $b(z)$-left-strongly concave if for all z, $f''(z) \le -b(z)$.

References

[1] Open Subtitles french frequent words lists. Obtained at https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#French. Original project url: www.opensubtitles.org.

[2] Open Subtitles german frequent words lists. Obtained at https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#German. Original project url: www.opensubtitles.org.

[3] Project Gutenberg frequent words lists. Obtained at https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#Top_English_words_lists. Original project url: https://www.gutenberg.org.

[4] Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of the 28th Conference on Computational Learning Theory (COLT)*, pages 4–22, 2015.

[5] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, , and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, 2000.

[6] José M Bernardo. Expected information as expected utility. *Annals of Statistics*, pages 686–690, 1979.

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

[8] Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22, pages 1–1, 2015.

[9] Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Learning mixtures of structured distributions over discrete domains. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1380–1394, 2013.

[10] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of gaussians. In *Proceedings of the 27th Conference on Computational Learning Theory (COLT)*, pages 1183–1213, 2014.

[11] A. Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

[12] Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):505–562, 2016.

[13] Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

[14] Rafael Frongillo and Ian Kash. Vector valued property elicitation. In *Proceedings of the 28th Algorithmic Learning Theory (ALT)*, pages 710–727, 2015.

[15] Tilman Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[16] Irving J. Good. Rational decisions. *Journal of the Royal Statistical Society*, 14(1):107–114, 1952.

[17] Alan Julian Izenman. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.

[18] Adam Kalai, Ankur Moitra, and Gregory Valiant. Disentangling Gaussians. *Communications of the ACM*, 55(2):113–120, February 2012.

[19] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the 26th Annual ACM Symposium on Theory of Computing (STOC)*, pages 273–282, 1994.

[20] Christopher D Manning, Christopher D Manning, and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

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

[22] Harikrishna Narasimhan, Harish G. Ramaswamy, Aadirupa Saha, and Shivani Agarwal. Consistent multiclass algorithms for complex performance measures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 2398–2407, 2015.

[23] Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17:1–45, 2016.

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

[25] Bernard W Silverman. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, 1986.

[26] Gregory Valiant and Paul Valiant. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pages 142–155, 2016.

A Additional Proofs for Strict Proper Losses

A.1 Proof of Lemma 1

Lemma 1 (restated). *For any distribution $\mathbf{p} \in \Delta_{\mathcal{X}}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$, and for any $t \in [0,1]$, we have $\mathbb{E}_{X\sim\mathbf{p}}\left[\frac{1}{p_X} \,\middle|\, X \in B_t\right] = \frac{1}{t}$, where $B_t = \{x : q_x = t\}$.*

Proof. We have

$$\mathbb{E}\left[\frac{1}{p_X} \,\middle|\, X \in B_t\right] = \sum_{x\in B_t} \frac{p_x}{\mathbf{p}(B)} \frac{1}{p_x} = \frac{|B_t|}{\mathbf{p}(B_t)} = \frac{1}{t}.$$

□

A.2 Proof of Theorem 1

Suppose $\ell(\mathbf{q}, x) = f(\frac{1}{q_x})$ for a strictly concave f. Consider any $\mathbf{q}$ that is calibrated with respect to $\mathbf{p}$. Recall that $B_t = \{x : q_x = t\}$ and $T(\mathbf{q}) = \{t : |B_t| \neq \emptyset\}$ is a finite set.

$$\begin{aligned}
\ell(\mathbf{p};\mathbf{p}) &= \sum_{t\in T(\mathbf{q})} \mathbf{p}(B_t)\,\mathbb{E}\left[\ell(\mathbf{p}, X) \,\middle|\, X \in B_t\right] \\
&= \sum_{t\in T(\mathbf{q})} \mathbf{p}(B_t)\,\mathbb{E}\left[f\left(\frac{1}{p_X}\right) \,\middle|\, X \in B_t\right] \\
&\leq \sum_{t\in T(\mathbf{q})} \mathbf{p}(B_t) f\left(\mathbb{E}\left[\frac{1}{p_X} \,\middle|\, X \in B_t\right]\right) \quad \text{(By Jensen's inequality)} \\
&= \sum_{t\in T(\mathbf{q})} \mathbf{p}(B_t) f\left(\frac{1}{t}\right) \quad \text{(By Lemma 1)} \\
&= \sum_{t\in T(\mathbf{q})} \sum_{x\in B_t} p_x f\left(\frac{1}{t}\right) \\
&= \sum_{t\in T(\mathbf{q})} \sum_{x\in B_t} p_x f\left(\frac{1}{q_x}\right) \\
&= \ell(\mathbf{q};\mathbf{p}).
\end{aligned}$$

If f is strictly concave and there exists a B_t where $\mathbf{q}$ and $\mathbf{p}$ disagree, then the inequality is strict.

B Additional Proofs for Strongly Proper Losses

B.1 Proof of Theorem 2

Let us start by an analogous result to Lemma 1. We defer the proof of this lemma to Appendix B.2.

Lemma 2. *Suppose $f(z)$ is $b(z)$-left-strongly concave. Let $B \subseteq \mathcal{X}$ be any set and let $t(B) := \frac{\mathbf{p}(B)}{|B|}$,*[5] *and suppose $\sum_{x\in B} |p_x - t(B)| \geq \epsilon$. Let $\mu = \frac{1}{t(B)}$. Then*

$$\mathop{\mathbb{E}}_{X\sim\mathbf{p}}\left[f\left(\frac{1}{p_X}\right) \,\middle|\, X \in B\right] \leq f(\mu) + \frac{b(\mu)}{32} \frac{\epsilon^2}{\mathbf{p}(B)^2 t(B)^2}.$$

Proof of Theorem 2. Note that a calibrated distribution $\mathbf{q}$ can be thought of as a piecewise uniform distribution with pieces $\{B_t\}_{t\in T(\mathbf{q})}$ and $\mathbf{q}(B_t) = \mathbf{p}(B_t)$. Let $\epsilon_t = \sum_{x\in B_t} |p_x - q_x|$, with

[5] When $B = B_t(\mathbf{q})$ for some $t \in [0,1]$, $t(B) = t$.

$\sum_{t\in T(\mathbf{q})} \epsilon_t = \epsilon = \|\mathbf{p}-\mathbf{q}\|_1$. Let $\alpha = \frac{\|\mathbf{p}-\mathbf{q}\|_1}{4}$ and let $H = \{t \in T(\mathbf{q}) : t \geq \frac{\alpha}{N}\}$ refer to indices of pieces in which the two distributions place reasonably high probability. We have:

$$\ell(\mathbf{q};\mathbf{p}) - \ell(\mathbf{p};\mathbf{p}) = \sum_x p_x \left[f\left(\frac{1}{q_x}\right) - f\left(\frac{1}{p_x}\right)\right] = \sum_{t\in T(\mathbf{q})} \mathbf{p}(B_t) \left[f\left(\frac{1}{t}\right) - \mathop{\mathbb{E}}_{X|B_t}\left[f\left(\frac{1}{p_X}\right)\right]\right]$$

where $\mathbb{E}_{X|B_t}[\cdot]$ refers to the expectation over $X \sim \mathbf{p}$ conditioned on $X \in B_t$. Now consider any fixed component B_t. The difference inside the brackets is $f\left(\frac{1}{t}\right) - \mathbb{E}_{X|B_t}\left[f\left(\frac{1}{p_X}\right)\right]$. Intuitively, strong concavity implies there should be a significant "Jensen gap". This is formalized in Lemma 2 of Appendix B that shows that if $\sum_{x\in B_j} |p_x - q_x| = \epsilon_j$, then

$$f\left(\frac{1}{t}\right) - \mathop{\mathbb{E}}_{X|B_t}\left[f\left(\frac{1}{p_X}\right)\right] \geq \frac{b\left(\frac{1}{t}\right)}{32} \cdot \frac{\epsilon_t^2}{t^2 \mathbf{p}(B_t)^2}. \tag{2}$$

Summing over all $t \in T(\mathbf{q})$ and Applying the assumption that $b(x) \geq \frac{C(x)}{x^2}$ where $C(x)$ is nonincreasing along with the fact that $t \geq \frac{\alpha}{N}$ for $t \in H$ gives

$$\ell(\mathbf{q};\mathbf{p}) - \ell(\mathbf{p};\mathbf{p}) \geq \sum_{t\in T(\mathbf{q})} \mathbf{p}(B_t) \frac{b(\frac{1}{t})}{32} \frac{\epsilon_t^2}{t^2\mathbf{p}(B_t)^2} \geq \sum_{t\in H} \mathbf{p}(B_t) \frac{b(\frac{1}{t})}{32} \frac{\epsilon_t^2}{t^2\mathbf{p}(B_t)^2} \geq \frac{C\left(\frac{N}{\alpha}\right)}{32} \sum_{t\in H} \frac{\epsilon_t^2}{\mathbf{p}(B_t)}. \tag{3}$$

For $t \notin H$, since $\mathbf{q}(B_t) = \mathbf{p}(B_t) \leq \frac{\alpha|B_t|}{N}$ we have $\epsilon_t \leq \frac{2\alpha|B_t|}{N}$. Thus we have $\sum_{t\notin H} \epsilon_t \leq \frac{2\alpha}{N}|T(\mathbf{q}) \setminus H| \leq 2\alpha$, and so correspondingly, $\sum_{t\in H} \epsilon_t \geq \epsilon - 2\alpha$. Since the bound of (3) is increasing in each ϵ_t and decreasing in each $\mathbf{p}(B_t)$ we can obtain a lower bound by considering its minimum when $\sum_{t\in H} \epsilon_t = \epsilon - 2\alpha$ and $\sum_{t\in H} \mathbf{p}(B_t) = 1$. By the convexity of $(\cdot)^2$ this minimum is obtained at $\epsilon_t = \mathbf{p}(B_t) \cdot (\epsilon - 2\alpha)$.

This gives an overall bound of $\ell(\mathbf{q};\mathbf{p}) - \ell(\mathbf{p};\mathbf{p}) \geq \frac{C\left(\frac{N}{\alpha}\right)}{32} \cdot (\epsilon - 2\alpha)^2$. Replacing $\alpha = \frac{\|\mathbf{p}-\mathbf{q}\|_1}{4}$ in this bound completes theorem. □

B.2 Proof of Lemma 2

Proof. We draw $X \sim \mathbf{p}$ conditioned on $X \in B$. Let $S = \{x \in B : p_x > t(B)\}$. We upper-bound $f(\frac{1}{p_X})$ for each realization of X. If $p_X \leq t(B)$, then we simply use concavity. Otherwise, if $X \in S$, we use $b(z)$-left-strong-concavity. Furthermore, note that by Lemma 1, $\mathbb{E}_{X\sim\mathbf{p}|B}\left[\frac{1}{p_X}\right] = \mu$. We have:

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim\mathbf{p}|B}\left[f\left(\frac{1}{p_X}\right)\right] &\leq \mathbb{E}\left[f(\mu) + df(\mu)\cdot\left(\frac{1}{p_X} - \mu\right) - \mathbf{1}[X \in S]\frac{b(\mu)}{2}\left(\frac{1}{p_X} - \mu\right)^2\right] \\ &= f(\mu) - \frac{b(\mu)}{2}\frac{1}{\mathbf{p}(B)}\sum_{x\in S} p_x \left(\frac{1}{p_x} - \mu\right)^2,\end{aligned}$$

Note the $\frac{1}{\mathbf{p}(B)}$ term arises from conditioning on $X \in B$. We now lower-bound the sum, using the constraint that $\sum_{x\in B} |p_x - t(B)| \geq \epsilon$, which implies that $\sum_{x\in S} p_x - t(B) \geq \frac{\epsilon}{2}$.

$$\sum_{x\in S} p_x \left(\frac{1}{t(B)} - \frac{1}{p_x}\right)^2 = \frac{\mathbf{p}(S)}{t(B)^2} - \frac{2|S|}{t(B)} + \sum_{x\in S}\frac{1}{p_x}.$$

Fixing $\mathbf{p}(S)$ and $|S|$, we get by convexity that this is minimized by p_x constant on S, therefore equal to $t(B) + \frac{\epsilon}{2|S|}$. So we have

$$\begin{aligned}|S|\left(t(B) + \frac{\epsilon}{2|S|}\right)\left(\frac{1}{t(B)} - \frac{1}{t(B) + \frac{\epsilon}{2|S|}}\right)^2 &= \left(|S|t(B) + \frac{\epsilon}{2}\right)\left(\frac{\epsilon}{2|S|\left(t(B)^2 + \frac{\epsilon t(B)}{2|S|}\right)}\right)^2 \\ &= \frac{|S|t(B)\epsilon^2 + \frac{\epsilon^3}{2}}{4|S|^2 t(B)^2\left(t(B) + \frac{\epsilon}{2|S|}\right)^2}.\end{aligned}$$

We consider the two cases for the larger term in the denominator. In the case $\frac{\epsilon}{2|S|} > t(B)$, we get

$$\begin{aligned}
&\geq \frac{|S|t(B)\epsilon^2 + \frac{\epsilon^3}{2}}{4|S|^2 t(B)^2 \left(\frac{\epsilon}{|S|}\right)^2} \\
&\geq \frac{|S|t(B) + \frac{\epsilon}{2}}{4t(B)^2} \\
&\geq \frac{\epsilon}{4t(B)^2} \\
&\geq \frac{\epsilon^2}{4\mathbf{p}(B)t(B)^2}
\end{aligned}$$

where the last line follows because we must have $\epsilon \leq \mathbf{p}(B)$ from the definition of ϵ. In the remaining case, we get

$$\begin{aligned}
&\geq \frac{|S|t(B)\epsilon^2 + \frac{\epsilon^3}{2}}{4|S|^2 t(B)^2 \left(2t(B)\right)^2} \\
&\geq \frac{\epsilon^2}{16|S|t(B)^3} \\
&\geq \frac{\epsilon^2}{16|B|t(B)^3} \\
&= \frac{\epsilon^2}{16\mathbf{p}(B)t(B)^2}.
\end{aligned}$$

□

C Additional Proofs for Concentration of Losses

We first give a simple lemma that will be used to prove our main calibrated distribution concentration result, Theorem 3.

Lemma 3 (Calibrated Distribution Probability Lower Bound)**.** *For any $\mathbf{p} \in \Delta_{\mathcal{X}}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$, for any $x \in \mathcal{X}$,*

$$q_x \geq \frac{p_x}{N}. \tag{4}$$

Proof. Let $B = \{x' : q_{x'} = q_x\}$. Then by calibration we have:

$$q_x = \frac{\mathbf{q}(B)}{|B|} \geq \frac{\mathbf{q}(B)}{N} = \frac{\mathbf{p}(B)}{N} \geq \frac{p_x}{N}.$$

□

This bound is achieved when $\mathbf{q}$ is the uniform distribution and $\mathbf{p}$ is a point distribution.

C.1 Proof of Theorem 3

We prove the following stronger result, Proposition 1, that only uses the lower-bound property $q_x \geq \Omega(\frac{p_x}{N})$ and does not require a distribution to be calibrated. Combining this proposition with Lemma 3 immediately proves Theorem 3.

Proposition 1. *Suppose ℓ is a local loss function with $\ell(\mathbf{q}, x) = f\left(\frac{1}{q_x}\right)$ for nonnegative, increasing, concave $f(z)$. Suppose further that $f(z) \leq cz^r$ for all $z \geq 1$, some constant $c > 0$, and some constant $r < 1$. Given $\mathbf{p}$, suppose $\mathbf{q}$ is **any** distribution such that $q_x \geq \frac{c_2 p_x}{N}$ for all x and some constant $c_2 \in (0, 1]$. Then, drawing at least $m(\gamma, \delta, N)$ samples guarantees that $\left|\ell(\mathbf{q}; \hat{p}_) - \ell(\mathbf{q}; \mathbf{p})\right| \leq \gamma$ with probability $\geq 1 - \delta$ if*

$$m(\gamma, \delta, N) \geq \frac{c_1 \cdot f(\beta)^2 \ln \frac{1}{\delta}}{\gamma^2},$$

where c_1 is a fixed constant and $\beta := \frac{2^{2/(1-r)} N^{3/(1-r)+2}}{c_2^{r/(1-r)} \delta \cdot \min(1, [\gamma/c]^{1/(1-r)})}$.

Proof. Fix a sample size $m \leq N$. Let $\Omega \subseteq \mathcal{X}$ be the set of x's that occur with non-negligible probability:

$$\Omega = \left\{ x : p_x \geq \frac{c_2^{r/(1-r)} \cdot \delta \cdot \min(1, [\gamma/c]^{1/(1-r)})}{2^{2/(1-r)} N^{3/(1-r)+1}} \right\}.$$

we have $\mathbf{p}(\mathcal{X} \setminus \Omega) \leq N \cdot \frac{c_2^{r/(1-r)} \delta}{4N^4} \leq \frac{\delta}{4N}$ and thus for $x_1, \ldots x_m$ drawn i.i.d. from $\mathbf{p}$. By a union bound, letting $\mathcal{E}$ be the event that $x_1, \ldots, x_m \in \Omega$ and using that $m \leq N$:

$$\Pr\left[\mathcal{E}\right] \geq 1 - \frac{\delta}{4}. \tag{5}$$

We will condition on $\mathcal{E}$ going forward. First note that for $x \in \Omega$, we can bound $\ell(\mathbf{q}, x)$ using Lemma 3. Specifically, since $q_x \geq \frac{c_2 p_x}{N}$ and f is nondecreasing, we have:

$$\ell(\mathbf{q}, x) = f\left(\frac{1}{q_x}\right) \leq f\left(\frac{2^{2/(1-r)} N^{3/(1-r)+2}}{c_2^{r/(1-r)} \cdot \delta \cdot \min(1, [\gamma/c]^{1/(1-r)})}\right).$$

Denote

$$\beta := \frac{2^{2/(1-r)} N^{3/(1-r)+2}}{c_2^{r/(1-r)} \cdot \delta \cdot \min(1, [\gamma/c]^{1/(1-r)})}.$$

Letting z_i be the random variable

$$z_i = \frac{1}{m}\left(\ell(\mathbf{q}, x_i) - \mathop{\mathbb{E}}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x) | x \in \Omega]\right),$$

we have for $x_i \in \Omega$, $|z_i| \leq \frac{f(\beta)}{m}$ (where we use that $\ell(\mathbf{q}, x)$ is nonnegative by assumption.) So $\mathbb{E}[z_i^2 \mid x_i \in \Omega] \leq f(\beta)^2/m^2$. Then by a standard Bernstein inequality:

$$\Pr\left[\left|\frac{1}{m}\sum_{j=1}^{m} \ell(\mathbf{q}, x_j) - \mathop{\mathbb{E}}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x) | x \in \Omega]\right| \geq \frac{\gamma}{2} \mid \mathcal{E}\right] \leq \exp\left(-\frac{\gamma^2/8}{f(\beta)^2/m + f(\beta)/m \cdot \gamma/3}\right) \leq \frac{\delta}{2} \tag{6}$$

where the second inequality follows if we have $m \geq \frac{c_1 f(\beta)^2 \log(1/\delta)}{\gamma^2}$ for sufficiently large c_1. By a union bound, from (5) and (6) we have:

$$\Pr\left[\left|\frac{1}{m}\sum_{j=1}^{m} \ell(\mathbf{q}, x_j) - \mathop{\mathbb{E}}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x) | x \in \Omega]\right| \geq \frac{\gamma}{2}\right] \leq \delta.$$

It remains to show that the conditional expectation $\mathbb{E}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x) | x \in \Omega]$ is very close to $\ell(\mathbf{q}; \mathbf{p}) = \mathbb{E}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x)]$, which will give us the lemma. Intuitively, by conditioning on $x \in \Omega$ we are only removing very low probability events, which do not have a big effect on the loss. Specifically, we need to show that:

$$\left|\mathop{\mathbb{E}}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x) | x \in \Omega] - \ell(\mathbf{q}; \mathbf{p})\right| \leq \frac{\gamma}{2} \tag{7}$$

Since $\mathbf{p}(\mathcal{X} \setminus \Omega) \leq N \cdot \frac{c_2^{r/(1-r)}\delta \cdot \min(1,\gamma/c)}{4N^4} \leq \frac{c_2^{r/(1-r)}\gamma}{4N^3} \leq \frac{c_2^r \gamma}{4N^3}$, using that f is nondecreasing, $f(z) \leq cz^r$ for some c and $r < 1$, and $q_x \geq \frac{c_2 p_x}{N}$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x) \mid x \in \Omega] &= \sum_{x \in \Omega} \frac{p_x}{\mathbf{p}(\Omega)} \cdot \ell(\mathbf{q}, x) \\
&\leq \frac{1}{1 - \frac{c_2^r \min(1,\gamma/c)}{4N^3}} \cdot \sum_{x \in \Omega} p_x \cdot \ell(\mathbf{q}, x) \\
&\leq \left(1 + \frac{c_2^r \min(1,\gamma/c)}{2N^3}\right) \cdot \sum_{x \in \mathcal{X}} p_x \cdot \ell(\mathbf{q}, x) \\
&\leq \ell(\mathbf{q}; \mathbf{p}) + \frac{c_2^r \min(1,\gamma/c)}{2N^3} \cdot \sum_{x \in \mathcal{X}} p_x \cdot f\left(\frac{N}{c_2 p_x}\right) \\
&\leq \ell(\mathbf{q}; \mathbf{p}) + \frac{c_2^r \min(1,\gamma/c)}{2N^3} \cdot c \cdot \frac{N^r}{c_2^r} \sum_{x \in \mathcal{X}} p_x^{1-r} \\
&= \ell(\mathbf{q}; \mathbf{p}) + \frac{\min(1,\gamma/c)}{2N^3} \cdot c \cdot N^{2r} \\
&\leq \ell(\mathbf{q}; \mathbf{p}) + \frac{\gamma}{2}.
\end{aligned}$$

This gives us one side of (7). On the other side we have:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x \sim \mathbf{p}}[\ell(\mathbf{q}, x) \mid x \in \Omega] &= \sum_{x \in \Omega} \frac{p_x}{\mathbf{p}(\Omega)} \cdot \ell(\mathbf{q}, x) \\
&\geq \sum_{x \in \Omega} p_x \cdot \ell(\mathbf{q}, x) \\
&= \ell(\mathbf{q}; \mathbf{p}) - \sum_{x \notin \Omega} p_x \cdot \ell(\mathbf{q}, x).
\end{aligned} \tag{8}$$

Again using that $f(z) \leq cz^r$ for $r < 1$, that $q_x \geq \frac{c_2 p_x}{N}$, and that for $x \notin \Omega$ we have $p_x \leq \frac{c_2^{r/(1-r)}\delta \cdot \min(1,[\gamma/c]^{1/(1-r)})}{2^{2/(1-r)} N^{3/(1-r)+1}}$:

$$\sum_{x \notin \Omega} p_x \cdot \ell(\mathbf{q}, x) = \sum_{x \notin \Omega} p_x \cdot f\left(\frac{1}{q_x}\right) \leq c \sum_{x \notin \Omega} p_x^{1-r} \cdot \frac{N^r}{c_2^r} \leq c \cdot N^{r+1} \cdot \frac{\gamma/c}{4N^3} \leq \frac{\gamma}{4N}.$$

Combined with (8) this yields the other side of (7), completing the bound and the proof. □

Proof of Theorem 3. Using Lemma 3, we have that for a calibrated distribution $\mathbf{q}$, $q_x \in \Omega(p_x/N)$ for all $x \in \mathcal{X}$. Together with the assumption that $f(z) \leq c\sqrt{z}$, we can directly apply Proposition 1 to prove the claim. □

D Additional Proofs for Sample Proper Losses

D.1 Proof of Theorem 4

In the statement of Theorem 4 we require that $\ell(\mathbf{q}, x) = f\left(\frac{1}{q_x}\right)$ for f that is nonnegative, increasing, and $\frac{C(x)}{x^2}$-left-strongly concave. Further we require that $C(x)$ is non-decreasing and non-negative for $x \geq 1$. Directly applying Theorem 2 we thus have:

$$\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) \geq C\left(\frac{4N}{\|\mathbf{p} - \mathbf{q}\|_1}\right) \cdot \frac{\|\mathbf{p} - \mathbf{q}\|_1^2}{128}. \tag{9}$$

Let $\gamma := C\left(\frac{4N}{\|\mathbf{p}-\mathbf{q}\|_1}\right) \cdot \frac{\|\mathbf{p}-\mathbf{q}\|_1^2}{128}$. Additionally, since $f(x) \leq c\sqrt{z}$ for $z \geq 1$ and since $\mathbf{q}, \mathbf{p} \in \mathcal{C}(\mathbf{p})$, applying Theorem 3 with error parameter $\gamma/3$ and failure parameter $\delta/2$, we have for $\beta := \frac{288 N^8}{\delta \cdot \min(1, \gamma^2/c^2)}$, if $m \geq \frac{c_1 f(\beta)^2 \lg \frac{2}{\delta}}{(\gamma/3)^2}$ for large enough constant c_1 then the following hold, each with probability $\geq 1-\delta/2$:

$$|\ell(\mathbf{q};\hat{\mathbf{p}}) - \ell(\mathbf{q};\mathbf{p})| \leq \frac{\gamma}{3} \text{ and } |\ell(\mathbf{p};\hat{\mathbf{p}}) - \ell(;\mathbf{p})| \leq \frac{\gamma}{3}.$$

By a union bound, with probability $\geq 1-\delta$ both bounds hold simultaneously and by (9) we have:

$$\ell(\mathbf{q};\hat{\mathbf{p}}) - \ell(\mathbf{p};\hat{\mathbf{p}}) \geq \ell(\mathbf{q};\mathbf{p}) - \ell(\mathbf{p};\mathbf{p}) - \frac{2\gamma}{3} \geq \gamma - \frac{2\gamma}{3} > 0,$$

which completes the theorem. Plugging the value of γ in we see that the bound holds for

$$m \geq \frac{c_1 f(\beta)^2 \ln \frac{1}{\delta}}{\left(C\left(\frac{4N}{\|\mathbf{p}-\mathbf{q}\|_1}\right) \cdot \frac{\|\mathbf{p}-\mathbf{q}\|_1^2}{128}/3\right)^2} = \frac{c_1' f(\beta)^2 \ln \frac{1}{\delta}}{\left(C\left(\frac{4N}{\|\mathbf{p}-\mathbf{q}\|_1}\right) \cdot \|\mathbf{p}-\mathbf{q}\|_1^2\right)^2}$$

for large enough constant c_1'. Additionally, we see that:

$$\beta = \frac{288N^8}{\delta \cdot \min\left(1, \left[C\left(\frac{4N}{\|\mathbf{p}-\mathbf{q}\|_1}\right) \cdot \frac{\|\mathbf{p}-\mathbf{q}\|_1^2}{128c}\right]^2\right)}.$$

E Instantiation of Theorems 2, 3, and 4

Let us start with two observations regarding loss functions, characterizing inverse concave loss functions and inverse left-concave functions.

Observation 1. *Let $\ell(\mathbf{q}, x) = f\left(\frac{1}{q_x}\right)$ be such that ℓ is nonnegative, twice differentiable, decreasing, and convex. Then, $f(x)$ is concave.*

Proof. For ease of exposition, let $\ell(z) = f(\frac{1}{z})$.

$$\begin{aligned}
\frac{df}{dy} &= \frac{d\ell(\frac{1}{y})}{dz}\left(\frac{-1}{y^2}\right) \\
\frac{d^2 f}{dz^2} &= \frac{d^2\ell(\frac{1}{y})}{dz^2}\left(\frac{-1}{y^2}\right) + \frac{d\ell(\frac{1}{y})}{dz}\left(\frac{2}{y^3}\right).
\end{aligned}$$

Decreasing and convex gives a negative derivative and positive second derivative. Given that $y > 0$, we obtain a negative second derivative, hence concavity. $\square$

Observation 2. *Consider a nonincreasing function $b(z)$. A function f is $b(z)$-left-strongly concave if for all z, $f''(z) \leq -b(z)$.*

Proof. We need to show that f restricted to $[0, z]$ is $b(z)$-strongly concave. Consider $z_1 \geq z_2$. Since $b(z)$ is non-increasing we have for $t \in [z_2, z_1]$:

$$f'(t) = f'(z_2) + \int_{z_2}^{t} f''(s)ds \leq f'(z_2) - b(z) \cdot (t - z_2).$$

We thus have:

$$\begin{aligned}
f(z_1) - f(z_2) = \int_{z_2}^{z_1} f'(t)dt &\leq \int_{z_2}^{z_1} [f'(z_2) - b(z)(t-z_2)]dt \\
&\leq f'(z_2)\cdot(z_1 - z_2) - b(z)\cdot\frac{(z_1-z_2)^2}{2}.
\end{aligned}$$

Rearranging gives:

$$D_{-f}(z_1, z_2) := f(z_2) + f'(z_2)\cdot(z_1 - z_2) - f(z_1) \geq \frac{b(z)}{2}\cdot(z_1 - z_2)^2.$$

For $z_1 \le z_2$, analogously for $t \in [z_1, z_2]$ we have:

$$f'(t) = f'(z_2) - \int_t^{z_2} f''(s)ds \ge f'(z_2) - b(z) \cdot (t - z_2)$$

and so

$$f(z_1) - f(z_2) = -\int_{z_2}^{z_1} f'(t)dt \le \int_{z_2}^{z_1} [f'(z_2) - b(z)(t - z_2)]dt$$
$$\le f'(z_2) \cdot (z_1 - z_2) - b(z) \cdot \frac{(z_1 - z_2)^2}{2}.$$

Rearranging gives again gives:

$$f(z_2) + f'(z_2) \cdot (z_1 - z_2) - f(z_1) \ge \frac{b(z)}{2} \cdot (z_1 - z_2)^2,$$

completing the lemma. □

E.1 Deriving Table 1

For $\ell(\mathbf{q}, x) = (\ln(1/q_x))^p$ for a constant $p \in (0, 1]$. By Observation 2, we have that $(\ln(z))^p$ is $C(z)/z^2$-left-strongly concave for

$$C(z) = p \ln(z)^{p-1} + p(1-p) \ln(z)^{p-2} \in \Theta\left(\ln(z)^{p-1}\right).$$

Moreover, $C(z)$ is non-increasing and non-negative for $z \ge 1$ and $\ln(z)^{p-1} \le \sqrt{z}$. Using these, for any $\mathbf{p}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$ such that $\|\mathbf{p} - \mathbf{q}\|_1 \ge \epsilon$ we have

- By Theorem 2, $\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) = \Omega\left(\epsilon^2 \ln(N/\epsilon)^{p-1}\right)$.
- By Theorem 3, an empirical distribution $\hat{\mathbf{p}}$ of $O\left(\gamma^{-2} \ln(1/\delta) \ln(N/\delta\gamma)^{2p}\right)$ i.i.d samples from $\mathbf{p}$ is sufficient such that $|\ell(\mathbf{q}; \hat{\mathbf{p}}) - \ell(\mathbf{q}; \mathbf{p})| \le \gamma$ with probability $1 - \delta$.
- By Theorem 4, an empirical distribution $\hat{\mathbf{p}}$ of
$$O\left(\frac{1}{\epsilon^4} \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{N}{\delta\epsilon^2 \ln(N/\epsilon)^p}\right)^{2p} \ln(N/\epsilon)^{-2p+2}\right) \in O\left(\frac{1}{\epsilon^4} \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{N}{\delta\epsilon}\right)^2\right)$$
i.i.d samples from $\mathbf{p}$ is sufficient such that $\ell(\mathbf{q}; \hat{\mathbf{p}}) > \ell(\mathbf{p}; \hat{\mathbf{p}})$ with probability $1 - \delta$.

For $\ell(\mathbf{q}, x) = \ln(e^2/q_x)^2$. By Observation 2, we have that $\ln(e^2 \cdot z)^2$ is $\frac{2+2\ln(z)}{z^2}$-left-strongly concave. Since Theorem 2 requires that $C(z)$ is nonincreasing we cannot set $C(z) = 2 + 2\ln(z)$ as might be expected. Instead we set $C(z) = 2$. Additionally, using that $\ln(e^2 \cdot z)^2 \le \sqrt{z}$, for any $\mathbf{p}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$ such that $\|\mathbf{p} - \mathbf{q}\|_1 \ge \epsilon$ we have

- By Theorem 2, $\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) = \Omega\left(\epsilon^2\right)$.
- By Theorem 3, an empirical distribution $\hat{\mathbf{p}}$ of $O\left(\gamma^{-2} \ln(1/\delta) \ln(N/\delta\gamma)^4\right)$ i.i.d samples from $\mathbf{p}$ is sufficient such that $|\ell(\mathbf{q}; \hat{\mathbf{p}}) - \ell(\mathbf{q}; \mathbf{p})| \le \gamma$ with probability $1 - \delta$.
- By Theorem 4, an empirical distribution $\hat{\mathbf{p}}$ of
$$O\left(\frac{1}{\epsilon^4} \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{N}{\delta\epsilon^2 \ln(N/\epsilon)}\right)^4\right) \in O\left(\frac{1}{\epsilon^4} \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{N}{\delta\epsilon}\right)^4\right)$$
i.i.d samples from $\mathbf{p}$ is sufficient such that $\ell(\mathbf{q}; \hat{\mathbf{p}}) > \ell(\mathbf{p}; \hat{\mathbf{p}})$ with probability $1 - \delta$.

E.2 Other Loss Functions

We also instantiate Theorem 2 for a few natural loss functions that do not obtain strong finite sample bounds (Theorems 3, and 4).

For the linear loss $\ell_{lin\text{-}loss}(\mathbf{q}, x) = -q_x$, we have by Observation 2 that $-\frac{1}{z}$ is $\frac{2}{z^3}$-left-strongly-concave. Thus setting $C(z) = 1/z$, by Theorem 2 for any $\mathbf{p}$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p})$ with $\|\mathbf{p} - \mathbf{q}\|_1 \ge \epsilon$:

$$\ell_{lin\text{-}loss}(\mathbf{q}; \mathbf{p}) - \ell_{lin\text{-}loss}(\mathbf{p}; \mathbf{p}) = \Omega\left(\frac{\epsilon}{N} \cdot \epsilon^2\right) = \Omega\left(\frac{\epsilon^3}{N}\right).$$

We can improve the dependence on N and ϵ by considering e.g., $\ell(\mathbf{q}, x) = -\sqrt{q_x}$. In this case we have that $-1/\sqrt{z}$ is $\frac{3}{4z^{5/2}}$-left-strongly-concave. Thus setting $C(z) = \frac{3}{4\sqrt{z}}$, by Theorem 2 we have:

$$\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) = \Omega\left(\sqrt{\frac{\epsilon}{N}} \cdot \epsilon^2\right) = \Omega\left(\frac{\epsilon^{2.5}}{\sqrt{N}}\right).$$

F Approximate Calibration

In this section we show that our results are robust to a notion of approximate calibration and that we can construct distributions that satisfy approximate calibration using a small number of samples.

Definition 8 (Approximate Calibration)**.** For $\mathbf{q} \in \Delta_{\mathcal{X}}$, for any $t \in [0,1]$, let $B_t = \{x : q_t = t\}$. $\mathbf{q}$ is (α_1, α_2)-*approximately calibrated with respect to* $\mathbf{p}$ if there is some subset $T \subseteq [0,1]$ such that $\mathbf{q}(B_t) \in (1 \pm \alpha_1)\mathbf{p}(B_t)$ for all $t \notin T$, $\mathbf{q}(B_t) \geq (1-\alpha_1)\mathbf{p}(B_t)$ for all $t \in T$, and $\mathbf{q}(\cup_{t \in T} B_t) \leq \alpha_2$. Let $\mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$ denote the set of all (α_1, α_2)-approximately calibrated distributions w.r.t. $\mathbf{p}$.

Intuitively, $\mathbf{q} \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$ is calibrated up to $(1 \pm \alpha_1)$ multiplicative error on any bucket B_t where $\mathbf{q}$ and hence $\mathbf{p}$ place reasonably large mass. There is some set of buckets (corresponding to $t \in T$) where $\mathbf{q}$ may significantly overestimate the probability assigned by $\mathbf{p}$, however, the total mass placed on these buckets will still be small – at most α_2.

F.1 Efficiently Constructing Approximately Calibrated Distributions

We now demonstrate that, given a candidate distribution $\mathbf{q}$ and sample access to $\mathbf{p}$, it is possible to efficiently construct $\mathbf{q}' \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$. Further, if $\mathbf{q} \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$ we will have $\|\mathbf{q} - \mathbf{q}'\|_1 \leq O(\alpha_1 + \alpha_2)$. In this way, if $\mathbf{q}$ is approximately calibrated, we can certify at least that it is close to another approximately calibrated distribution. Of $\mathbf{q}$ is not approximately calibrated, we return a distribution that is approximately calibrated, which of course, may be far from $\mathbf{q}$.

Theorem 5. *Given any $\mathbf{q} \in \Delta_{\mathcal{X}}$, sample access to $\mathbf{p} \in \Delta_{\mathcal{X}}$, and parameters $\alpha_1, \alpha_2, \delta \in (0,1]$ there is an algorithm that takes $O\left(\frac{\log\left(\frac{N}{\alpha_1}\right)^2 \cdot \log\left(\frac{\log N}{\delta \alpha_1}\right)}{\alpha_1^4 \cdot \alpha_2^2}\right)$ samples from $\mathbf{p}$ and returns, with probability $\geq 1-\delta$, $\mathbf{q}' \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$. Further, if $\mathbf{q} \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$ then $\|\mathbf{q} - \mathbf{q}'\|_1 \leq O(\alpha_1 + \alpha_2)$.*

The main idea of the algorithm achieving Theorem 5 is to round $\mathbf{q}$'s probabilities into buckets of multiplicative width $(1 \pm \alpha_1)$. We can then efficiently approximate the total probability mass in each bucket, excluding those that may have very small mass. On these buckets, we may over approximate the true mass, and thus they are included in the set T in Definition 8.

We start with a simple lemma that shows, using a standard concentration bound, how well we can approximate the probability of any event under any distribution.

Lemma 4. *For any $\mathbf{p} \in \Delta_{\mathcal{X}}$ and $B \subseteq \mathcal{X}$, given m independent samples $x_1, \ldots x_m \sim \mathbf{p}$, there is some fixed constant c such that, for any $\epsilon, \delta \in (0,1]$, if $m \geq \frac{3\ln(2/\delta)}{\epsilon^2}$, then with probability $\geq 1-\delta$:*

$$\left| \mathbf{p}(B) - \frac{|\{x_i : x_i \in B\}|}{m} \right| \leq \epsilon.$$

Proof. $\mathbb{E}\,|\{x_i : x_i \in B\}| = m \cdot \mathbf{p}(B)$. By a standard Chernoff bound:

$$\begin{aligned}
\Pr\left[\left||\{x_i : x_i \in B\}| - m \cdot \mathbf{p}(B)\right| \geq m \cdot \epsilon\right] &\leq e^{-\frac{\left(\frac{\epsilon}{\mathbf{p}(B)}\right)^2}{2+\frac{\epsilon}{\mathbf{p}(B)}} m\mathbf{p}(B)} + e^{-\frac{\left(\frac{\epsilon}{\mathbf{p}(B)}\right)^2}{2} m\mathbf{p}(B_i)} \\
&\leq e^{-\frac{\epsilon^2 m}{2\mathbf{p}(B)+\epsilon}} + e^{-\frac{\epsilon^2 m}{2}} \\
&\leq 2e^{-\frac{\epsilon^2 m}{3}},
\end{aligned}$$

which is $\leq \delta$ as long as $m \geq \frac{3\ln(2/\delta)}{\epsilon^2}$. □

With Lemma 4 in hand, we proceed to the proof of Theorem 5.

Proof of Theorem 5. For convenience, define $\gamma_1 = \frac{\alpha_1}{3}$, and $b = \lceil \log_{1-\frac{\gamma_1}{8}} \frac{\gamma_1}{8N} \rceil$. Note that $b = O\left(\frac{\log \frac{N}{\alpha_1}}{\alpha_1}\right)$. For $i \in \{1, \ldots, b\}$, define:

$$\bar{B}_i = \left\{ x : q_x \in \left(\left(1 - \frac{\gamma_1}{8}\right)^i, \left(1 - \frac{\gamma_1}{8}\right)^{i-1} \right] \right\}.$$

Let $\bar{B}_{b+1} = \left\{x : q_x \leq \left(1 - \frac{\gamma_1}{8}\right)^b\right\}$.[6] Note that $\bar{B}_1 \cup \ldots \cup \bar{B}_b \cup \bar{B}_{b+1} = \mathcal{X}$. Now, via Lemma 4, with $O\left(\frac{b^2 \cdot \log b/\delta}{\alpha_2^2 \cdot \alpha_1^2}\right)$ samples from $\mathbf{p}$ it is possible to compute $\tilde{\mathbf{p}}(\bar{B}_1), \ldots, \tilde{\mathbf{p}}(\bar{B}_{b+1})$ such that, with probability $\geq 1 - \delta$,

$$|\mathbf{p}(\bar{B}_i) - \tilde{\mathbf{p}}(\bar{B}_i)| \leq \frac{\gamma_1 \cdot \alpha_2}{8(b+1)}$$

for all i simultaneously. Let $\mathcal{E}$ be the event that these approximations hold, and assume that $\mathcal{E}$ occurs. Then for any i with $\tilde{\mathbf{p}}(\bar{B}_i) \leq \frac{\alpha_2}{4(b+1)}$, it must be that

$$\mathbf{p}(\bar{B}_i) \leq \frac{\alpha_2}{4(b+1)} + \frac{\gamma_1 \cdot \alpha_2}{8(b+1)} \leq \frac{\alpha_2}{2(b+1)}. \tag{10}$$

Let $L \subseteq \{1, \ldots, b+1\}$ be the set of all such i. Similarly, for i with $\tilde{\mathbf{p}}(\bar{B}_i) > \frac{\alpha_2}{4(b+1)}$, it must be that:

$$\mathbf{p}(B_i) > \frac{\alpha_2}{4(b+1)} - \frac{\gamma_1 \cdot \alpha_2}{8(b+1)} > \frac{\alpha_2}{8(b+1)}. \tag{11}$$

Let $H = \{1, \ldots, b+1\} \setminus L$ be the set of all such i.

Define $\mathbf{w}$ as follows: for $x \in \cup_{i \in L} \bar{B}_i$ set $w_x = \frac{\alpha_2}{2\left|\cup_{i \in L} \bar{B}_i\right|}$. For $i \in H$, for $x \in \bar{B}_i$ let $w_x = \frac{\tilde{\mathbf{p}}(\bar{B}_i)}{|\bar{B}_i|}$. We have the following facts about $\mathbf{w}$:

1. For $i \in H$, $\mathbf{w}(\bar{B}_i) = \tilde{\mathbf{p}}(\bar{B}_i) \in \mathbf{p}(\bar{B}_i) \pm \frac{\gamma_1 \cdot \alpha_2}{8(b+1)}$, which by the fact that $\mathbf{p}(\bar{B}_i) \geq \frac{\alpha_2}{8(b+1)}$ (equation (11)) gives for all $i \in H$:
$$\mathbf{w}(\bar{B}_i) \in (1 \pm \gamma_1)\, \mathbf{p}(\bar{B}_i). \tag{12}$$
2. $\mathbf{w}(\cup_{i \in L} B_i) = \frac{\alpha_2}{2}$ and by (10), $\mathbf{p}(\cup_{i \in L} \bar{B}_i) = \sum_{i \in L} \mathbf{p}(\bar{B}_i) \leq (b+1) \cdot \frac{\alpha_2}{2(b+1)} = \frac{\alpha_2}{2}$.

In combination, the above facts give that $\|\mathbf{w}\|_1 \in (1 \pm \gamma_1)$. Thus, letting $\mathbf{q}' = \frac{1}{\|\mathbf{w}\|_1} \cdot \mathbf{w}$, we have:

1. Applying (12), for all $i \in H$, $\left(\frac{1-\gamma_1}{1+\gamma_1}\right) \mathbf{p}(\bar{B}_i) \leq \mathbf{q}'(\bar{B}_i) \leq \left(\frac{1+\gamma_1}{1-\gamma_1}\right) \mathbf{p}(\bar{B}_i)$. Since $\gamma_1 = \frac{\alpha_1}{3}$ we have $\frac{1-\gamma_1}{1+\gamma_1} \geq 1 - \alpha_1$ and $\frac{1+\gamma_1}{1-\gamma_1} \leq 1 + \alpha_1$, which gives for all $i \in H$:
$$\mathbf{q}'(\bar{B}_i) \in (1 \pm \alpha_1)\, \mathbf{p}(\bar{B}_i). \tag{13}$$
2. $\mathbf{q}'(\cup_{i \in L} \bar{B}_i) \geq \frac{1}{1+\gamma_1} \cdot \frac{\alpha_2}{2} \geq (1 - \alpha_1) \cdot \frac{\alpha_2}{2} \geq (1 - \alpha_1) \cdot \mathbf{p}(\cup_{i \in L} \bar{B}_i)$. Additionally, $\mathbf{q}'(\cup_{i \in L} \bar{B}_i) \leq \frac{1}{1-\gamma_1} \cdot \frac{\alpha_2}{2} \leq \alpha_2$.
3. $\|\mathbf{q}' - \mathbf{w}\|_1 \leq \gamma_1$.

Properties (1) and (2) together give that $\mathbf{q}' \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$ where we define the set T to be $\{\bar{q}_x\}$ for $x \in \cup_{i \in L} \bar{B}_i$.

Recalling that $b = O\left(\frac{\log N/\alpha_1}{\alpha_1}\right)$, the overall sample complexity used to construct $\mathbf{q}'$ is:

$$O\left(\frac{b^2 \cdot \log b/\delta}{\alpha_2^2 \cdot \alpha_1^2}\right) = O\left(\frac{\log(N/\alpha_1)^2 \log(b/\delta)}{\alpha_1^4 \cdot \alpha_2^2}\right) = O\left(\frac{\log(N/\alpha_1)^2 \cdot \log\left(\frac{\log N}{\delta \alpha_1}\right)}{\alpha_1^4 \cdot \alpha_2^2}\right).$$

Finally, it remains to show that if $\mathbf{q} \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$, then $\|\mathbf{q} - \mathbf{q}'\|_1 \leq O(\alpha_1 + \alpha_2)$.

For every $j \leq b$, since $\mathbf{q}$ places all probabilities within $(1 \pm \frac{\gamma_1}{8}) = (1 \pm \frac{\alpha_1}{24})$ of each other on this bucket, for every $x \in \bar{B}_j$, $q_x \in (1 \pm \frac{\alpha_1}{24}) \cdot \frac{\mathbf{q}(\bar{B}_j)}{|\bar{B}_j|}$. We thus have:

$$\sum_{x \in \bar{B}_j} |q_x - q'_x| \leq |\mathbf{q}(\bar{B}_j) - \mathbf{q}'(\bar{B}_j)| + O(\alpha_1) \cdot \mathbf{q}(\bar{B}_j).$$

[6] Note that this this is different that the usual definition of $B_t = \{x : q_x = t\}$, but it is still within the same spirit of bucketing the elements based on their q_x values.

For $\bar{B}_{b+1}$ since $\mathbf{q}(\bar{B}_{b+1}) \leq \frac{\alpha}{24}$, we simply have $\sum_{x \in \bar{B}_b+1} |q_x - q'_x| \leq |\mathbf{q}(\bar{B}_j) - \mathbf{q}'(\bar{B}_j)| + O(\alpha_1)$. Thus overall:

$$\|\mathbf{q} - \mathbf{q}'\|_1 = \sum_{j=1}^{b+1} \sum_{x \in \bar{B}_j} |q_x - q'_x| \leq \sum_{j=1}^{b+1} |\mathbf{q}(\bar{B}_j) - \mathbf{q}'(\bar{B}_j)| + O(\alpha_1).$$

We now bound the above sum using that both $\mathbf{q}$ and $\mathbf{q}'$ are in $\mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$. Let T be the set of probabilities for which $\mathbf{q}$ may significantly overestimate $\mathbf{p}$ but places mass $\leq \alpha_2$. Let T' be analogous set for $\mathbf{q}'$ (see Definition 8). Let $\bar{\mathbf{q}}$ be vector obtained by setting $q_x = p_x$ for $\{x : q_x \in T\}$. Let $\bar{\mathbf{q}}'$ be defined analogously for $\mathbf{q}'$. We have:

$$\|\mathbf{q} - \mathbf{q}'\|_1 \leq \sum_{j=1}^{b+1} |\mathbf{q}(\bar{B}_j) - \mathbf{q}'(\bar{B}_j)| + O(\alpha_1) \leq \sum_{j=1}^{b+1} |\bar{\mathbf{q}}(\bar{B}_j) - \bar{\mathbf{q}}'(\bar{B}_j)| + O(\alpha_1 + \alpha_2).$$

Additionally, we can see that both $\bar{\mathbf{q}}$ and $\bar{\mathbf{q}}'$ are calibrated up to error $(1 \pm \alpha_1)$ on all $\bar{B}_j$ ($\bar{\mathbf{q}}$ is calibrated up to this error on all its level sets, which form a refinement of $\{\bar{B}_j\}$.) Thus we have:

$$\|\mathbf{q} - \mathbf{q}'\|_1 \leq \sum_{j=1}^{b+1} O(\alpha_1) \cdot \mathbf{p}(\bar{B}_j) + O(\alpha_1 + \alpha_2) = O(\alpha_1 + \alpha_2).$$

which completes the claim. □

F.2 Strong Properness Under Approximate Calibration

We now show that Theorem 2 is robust to approximation calibration, using a similar proof strategy. See Table 2 for a sampling of results that this implies, which essentially match those given by Table1 in the case of exact calibration.

Theorem 6. *Suppose $\ell(\mathbf{q}, x) = f(\frac{1}{q_x})$ where f is non-decreasing, and for $z \geq \frac{1}{\max_x q_x}$ is non-negative and satisfies $f'(z) \leq \frac{D(z)}{z}$ for some non-decreasing function D. Also suppose that f is $\frac{C(z)}{z^2}$-left-strongly concave for C that is non-increasing and non-negative for $z \geq 1$. Then for all $p \in \Delta_{\mathcal{X}}$, $\alpha_1 \leq 1/2$ and $\mathbf{q} \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$:*

$$\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) \geq \frac{C\left(\frac{N}{2\alpha_2}\right)}{32} \cdot \left(\|\mathbf{p} - \mathbf{q}\|_1 - \alpha_1 - 5\alpha_2\right)^2 - 2\alpha_1 \cdot D\left(\frac{N}{2\alpha_2}\right) - 3\alpha_2 \cdot f\left(\frac{N}{3\alpha_2}\right).$$

Proof. Let $\mathbf{q} \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$ be piecewise uniform with pieces $\{B_t\}_{t \in T(\mathbf{q})}$. Let $L_1 = \left\{t : \frac{\mathbf{p}(B_t)}{|B_t|} \leq \frac{\alpha_2}{N}\right\}$. Let $H \subseteq T(\mathbf{q}) \setminus L_1$ contain all remaining t for which $\mathbf{q}(B_t) \in (1 \pm \alpha_1)\mathbf{p}(B_t)$. Finally, let $L_2 = T(\mathbf{q}) \setminus (H \cup L_1)$ contain all remaining $t \in T(\mathbf{q})$. Let $\epsilon_t = \sum_{x \in B_t} |p_x - q_x|$, with $\sum_{t \in T(\mathbf{q})} \epsilon_t = \epsilon = \|\mathbf{p} - \mathbf{q}\|_1$. Finally, consider $\mathbf{q}' \in \mathcal{C}(\mathbf{p})$ that is exactly calibrated and piecewise uniform on $B_t(\mathbf{q})$, that is, $q'_x = \mathbf{p}(B_t(\mathbf{q}))/|B_t|$ for all $x \in B_t(\mathbf{q})$ and $t \in T(\mathbf{q})$.

By definition of L_1 we have $\mathbf{p}(\cup_{t \in L_1} B_t) = \mathbf{q}'(\cup_{t \in L_1} B_t) \leq \alpha_2$. Additionally, by our definition of approximate calibration, for any $t \in L_1$, either $\mathbf{q}(B_t) \in (1 \pm \alpha_1)\mathbf{p}(B_t)$ or else $t \in T$ is in the set of buckets for which the total mass $\mathbf{q}(\cup_{t \in T} B_t) \leq \alpha_2$. We have

$$\mathbf{q}(\cup_{t \in L_1} B_t) \leq (1 + \alpha_1)\alpha_2 + \alpha_2 \leq 3\alpha_2.$$

Similarly, using the definition of approximate calibration we have:

$$\mathbf{q}(\cup_{t \in L_2} B_t) \leq \alpha_2 \text{ and } \mathbf{p}(\cup_{t \in L_2} B_t) = \mathbf{q}'(\cup_{t \in L_2} B_t) \leq \frac{\alpha_2}{1 - \alpha_1} \leq 2\alpha_2.$$

This gives us that the truly calibrated $\mathbf{q}'$ is close to the approximately calibrated $\mathbf{q}$:

$$\|\mathbf{q} - \mathbf{q}'\|_1 \leq \sum_{t \in H} \alpha_1 \cdot \mathbf{p}(B_t) + \sum_{t \in L_1 \cup L_2} (\mathbf{q}(B_t) + \mathbf{q}'(B_t))) \leq \alpha_1 + 5\alpha_2.$$

Thus, by triangle inequality we have

$$\|\mathbf{p}-\mathbf{q}'\|_1 \geq \|\mathbf{p}-\mathbf{q}\|_1 - \alpha_1 - 5\alpha_2. \tag{14}$$

We can thus bound $\ell(\mathbf{q}';\mathbf{p}) - \ell(\mathbf{p};\mathbf{p})$ following the proof of Theorem 2. Let $\epsilon' = \|\mathbf{p}-\mathbf{q}'\|_1$ and $\epsilon'_t = \sum_{x\in B_t} |p_x - q'_x|$. Let $\ell_H(\mathbf{q};\mathbf{p}) = \sum_{j\in H}\sum_{x\in B_t} p_x f\left(\frac{1}{q_x}\right)$ be the loss restricted to the buckets in H. By (2) we can bound:

$$\ell(\mathbf{q}';\mathbf{p}) - \ell(\mathbf{p};\mathbf{p}) \geq \ell_H(\mathbf{q}';\mathbf{p}) - \ell_H(\mathbf{p};\mathbf{p}) \geq \sum_{t\in H} \mathbf{p}(B_t) \frac{b(\frac{|B_t|}{\mathbf{q}'(B_t)})}{32} \frac{(\epsilon'_t)^2}{\left(\frac{\mathbf{q}'(B_t)}{|B_t|}\right)^2 \mathbf{p}(B_t)^2}.$$

Since H excludes call elements in L_1, for all $t\in H$, $\frac{\mathbf{q}'(B_t)}{|B_t|} \geq \frac{\alpha_2}{N}$. Thus by our assumption on $b(\cdot)$:

$$\ell_H(\mathbf{q}';\mathbf{p}) - \ell_H(\mathbf{p};\mathbf{p}) \geq \sum_{t\in H} \frac{C\left(\frac{N}{\alpha_2}\right)}{32} \frac{(\epsilon'_t)^2}{\mathbf{p}(B_t)}.$$

and applying the same argument as in Theorem 2 can lower bound this quantity using (14) by:

$$\ell_H(\mathbf{q}';\mathbf{p}) - \ell_H(\mathbf{p};\mathbf{p}) \geq \frac{C\left(\frac{N}{\alpha_2}\right)}{32} \cdot \left(\|\mathbf{p}-\mathbf{q}\|_1 - \alpha_1 - 5\alpha_2\right)^2. \tag{15}$$

We next show that $\ell_H(\mathbf{q}';\mathbf{p}) - \ell_H(\mathbf{q};\mathbf{p})$ is not too large. Since $\mathbf{q}$ and $\mathbf{q}'$ are both piecewise uniform on $\{B_t\}_{t\in T(\mathbf{q})}$ and since $\mathbf{q}'$ is calibrated (i.e, $\mathbf{q}'(B_t) = \mathbf{p}(B_t)$ for all t),

$$\ell_H(\mathbf{q}';\mathbf{p}) - \ell_H(\mathbf{q};\mathbf{p}) = \ell_H(\mathbf{q}';\mathbf{q}') - \ell_H(\mathbf{q};\mathbf{q}').$$

We have using that f is nondecreasing:

$$\ell_H(\mathbf{q}';\mathbf{q}') = \sum_{t\in H}\sum_{x\in B_t} \mathbf{q}'(B_t)\cdot f\left(\frac{|B_t|}{\mathbf{q}'(B_t)}\right) \leq \sum_{t\in H}\sum_{x\in B_t} \mathbf{q}'(B_t)\cdot f\left(\frac{1}{(1-\alpha_1)\cdot q_x}\right) \tag{16}$$

Using the concavity of f along with the assumption that $f'(z) \leq \frac{D(z)}{z}$, we have:

$$\begin{aligned} f\left(\frac{1}{(1-\alpha_1)\cdot q_x}\right) &\leq f\left(\frac{1}{q_x}\right) + f'\left(\frac{1}{q_x}\right)\cdot\left(\frac{1}{(1-\alpha_1)q_x} - \frac{1}{q_x}\right) \\ &\leq f\left(\frac{1}{q_x}\right) + D\left(\frac{1}{q_x}\right)\cdot q_x \cdot \frac{\alpha_1}{(1-\alpha_1)q_x} \\ &\leq f\left(\frac{1}{q_x}\right) + D\left(\frac{1}{q_x}\right)\cdot 2\alpha_1. \end{aligned}$$

Plugging back into (16), using that $q_x \geq (1-\alpha)q'_x \geq \frac{\alpha_2(1-\alpha_1)}{N} \geq \frac{\alpha_2}{2N}$ for all $x \in \cup_{t\in H} B_t$ we have:

$$\begin{aligned} \ell_H(\mathbf{q}';\mathbf{q}') &\leq \sum_{t\in H}\sum_{x\in B_t} \mathbf{q}'(B_t)\left[f\left(\frac{1}{q_x}\right) + D\left(\frac{N}{2\alpha_2}\right)\cdot 2\alpha_1\right] \\ &\leq \ell_H(\mathbf{q};\mathbf{q}') + D\left(\frac{N}{2\alpha_2}\right)\cdot 2\alpha_1. \end{aligned}$$

Combined with (15) this gives:

$$\ell_H(\mathbf{q};\mathbf{p}) - \ell_H(\mathbf{p};\mathbf{p}) \geq \frac{C\left(\frac{N}{2\alpha_2}\right)}{32}\cdot\left(\|\mathbf{p}-\mathbf{q}\|_1 - \alpha_1 - 5\alpha_2\right)^2 - 2\alpha_1\cdot D\left(\frac{N}{2\alpha_2}\right). \tag{17}$$

Finally, let $\ell_L(\mathbf{q};\mathbf{p})$ be the loss restricted to buckets in $L_1\cup L_2$. As shown, $\sum_{t\in L_1\cup L_2}\sum_{x\in B_t} p_x \leq 3\alpha_2$. By the concavity of $f(z)$ we thus have:

$$\ell_L(\mathbf{p};\mathbf{p}) = \sum_{t\in L_1\cup L_2}\sum_{x\in B_t} p_x\cdot f\left(\frac{1}{p_x}\right) \leq 3\alpha_2\cdot f\left(\frac{N}{3\alpha_2}\right).$$

Combined with (17) this finally gives:

$$\ell(\mathbf{q};\mathbf{p}) - \ell(\mathbf{p};\mathbf{p}) \geq \ell_H(\mathbf{q};\mathbf{p}) - \ell_H(\mathbf{p};\mathbf{p}) - \ell_L(\mathbf{p};\mathbf{p})$$
$$\geq \frac{C\left(\frac{N}{2\alpha_2}\right)}{32} \cdot \left(\|\mathbf{p}-\mathbf{q}\|_1 - \alpha_1 - 5\alpha_2\right)^2 - 2\alpha_1 \cdot D\left(\frac{N}{2\alpha_2}\right) - 3\alpha_2 \cdot f\left(\frac{N}{3\alpha_2}\right),$$

which completes the theorem. ☐

$\ell(\mathbf{q},x)$	$f(z)$	$D(z)$	$C(z)$	α_1	α_2	$\frac{\ell(\mathbf{q};\mathbf{p})-\ell(\mathbf{p};\mathbf{p})}{\epsilon^2}$
$\ln\frac{1}{q_x}$	$\ln(z)$	1	1	$\Theta\left(\epsilon^2\right)$	$\Theta\left(\frac{\epsilon^2}{\ln N}\right)$	$\Omega(1)$
$\ln\frac{1}{q_x}^p, p \in (0,1]$	$(\ln(z))^p$	1	$\ln(z)^{p-1}$	$\Theta\left(\epsilon^2\right)$	$\Theta\left(\frac{\epsilon^2}{(\ln N)^p}\right)$	$\Omega\left((\ln N)^{p-1}\right)$
$\ln\left(\ln\frac{1}{q_x}\right)$	$\ln(\ln(z))$	1	$1/\ln(z)$	$\Theta\left(\epsilon^2\right)$	$\Theta\left(\frac{\epsilon^2}{\ln(\ln N)}\right)$	$\Omega\left(\frac{1}{\ln N}\right)$
$\frac{1}{\sqrt{q_x}}$	$\sqrt{z}$	$2\sqrt{z}$	$\frac{1}{4\sqrt{z}}$	$\Theta\left(\frac{\epsilon^4}{N}\right)$	$\Theta\left(\frac{\epsilon^4}{N}\right)$	$\Omega\left(\frac{\epsilon^2}{N}\right)$
$\left(\ln\frac{e^2}{q_x}\right)^2$	$\ln(e^2 z)^2$	$2\ln(z)+2$	2	$\Theta\left(\frac{\epsilon^2}{\ln N}\right)$	$\Theta\left(\frac{\epsilon^2}{(\ln N)^2}\right)$	$\Omega\left(1\right)$

Table 2: Examples of loss functions that are strongly proper over $\mathcal{C}(\mathbf{p},\alpha_1,\alpha_2)$. We let $\epsilon := \|\mathbf{p}-\mathbf{q}\|_1$ and assume for simplicity that $\epsilon \geq 1/N$. We fix values of α_1 and α_2 that yield a strong properness bound nearly matching that of Theorem 3 for truly calibrated distributions. Note that in the theorem $D(z)$ is required to be nondecreasing and thus we set it to 1 for all loss functions considered that grow slower than the log loss.

F.3 Concentration Under Approximate Calibration

It is also easy to show that our main concentration result, Theorem 3, is robust to approximate calibration, since this result just uses that calibration ensures $\frac{q_x}{p_x}$ is not too small for any x (Lemma 3). In particular, using an identical argument to what is used in Lemma 3 we can see from Definition 8 that for $\mathbf{q} \in \mathcal{C}(\mathbf{p},\alpha_1,\alpha_2)$, for all x, $q_x \geq \frac{(1-\alpha_1)p_x}{N} \geq \frac{p_x}{2N}$ for $\alpha_1 \leq 1/2$. Following the proof of Theorem 3 using this bound in place of Lemma 3 gives:

Theorem 7. *Suppose ℓ is a local loss function with $\ell(\mathbf{q},x) = f\left(\frac{1}{q_x}\right)$ for non-negative, non-decreasing, concave $f(z)$. Suppose further that $f(z) \leq c\sqrt{z}$ for all $z \geq 1$ and some constant c. Then ℓ concentrates over $\mathcal{C}(\mathbf{p},\alpha_1,\alpha_2)$ for any $\alpha_1 \leq 1/2$ and $m(\gamma,\delta,N) \leq N$ satisfying*

$$m(\gamma,\delta,N) \geq \frac{c_1 \cdot f\left(\beta\right)^2 \ln\frac{1}{\delta}}{\gamma^2},$$

where c_1 is a fixed constant and $\beta := \frac{32N^8}{\delta \cdot \min(1,\gamma^2/c^2)}$.

That is, for any $\mathbf{p} \in \Delta_{\mathcal{X}}, \mathbf{q} \in \mathcal{C}(\mathbf{p},\alpha_1,\alpha_2)$, drawing at least $m(\gamma,\delta,N)$ samples guarantees $|\ell(\mathbf{q};\hat{\mathbf{p}}) - \ell(\mathbf{q};\mathbf{p})| \leq \gamma$ with probability $\geq 1-\delta$.

First, the analogue of Lemma 3.

Lemma 5. *For all $\mathbf{p}$ and all $\mathbf{q} \in \mathcal{C}(\mathbf{p},\alpha_1,\alpha_2)$ with $\alpha_1 \leq 1/2$, for all x, we have $q_x \geq \frac{p_x}{N(1-\alpha_1)} \geq \frac{p_x}{2N}$.*

Proof. Given x, let $B = \{x' : q_{x'} = q_x\}$. By calibration,

$$q_x = \frac{\mathbf{q}(B)}{|B|} \geq \frac{\mathbf{q}(B)}{N} \geq \frac{(1-\alpha_1)\mathbf{p}(B)}{N} \geq \frac{(1-\alpha_1)p_x}{N}.$$

If $\alpha_1 \leq 1/2$, we get $q_x \geq \frac{p_x}{2N}$. ☐

Proof of Theorem 7. By Lemma 5, we have $q_x \geq \frac{c_2 p_x}{N}$ for all x with $c_2 = 0.5$. We apply Proposition 1, with all parameters exactly as in Theorem 3 except with $c_2 = 0.5$ rather than 1. ☐

Note that Theorem 7 is essentially identical to Theorem 3, up to a constant factor in β. Thus, all of our concentration results hold, up to constant factors, when $\mathbf{q} \in \mathcal{C}(\mathbf{p}, \alpha_1, \alpha_2)$ for $\alpha_1 \leq 1/2$ and *any* α_2. Also note that Theorem 7 gives a high probability bound for any $\mathbf{q} \in C(\mathbf{p})$. If for example, we wish to minimize $\ell(\mathbf{q}; \mathbf{p})$ over some set of candidate calibrated distributions, we could form an ϵ-net over these distributions and apply the theorem to all elements of this net, union bounding to obtain a bound on the probability that the empirical loss is close to the true loss on all elements. Optimizing would then yield a distribution with loss within γ of the minimal.

G Details on Motivating Example

We now give details on the motivating example for considering alternatives to the log loss in the introduction (see Figure 1.)

Dataset: Our primary data set is a list of 36663 of the most frequent English words, along with their frequencies in a count of all books on Project Gutenberg [3]. We then obtained a list of the 10000 most frequent French [1] and German [2] words. All capitals were converted to lower case, all accents removed, and all duplicates from the French and German lists removed. After preprocessing, the data consisted of the original 36663 English words along with 16409 French/German words. We gave the French and German words uniform frequency values, with the total frequency of these words comprising 12.23% of the probability mass of the word distribution.

Our tests are relatively insensitive to the exact frequency chosen for the French/German words within the reasonable range of 5-30%. Low frequency ($< 5\%$ of the total probability mass) is not sufficient noise to make the log loss minimizing distribution to perform poorly. On the other hand, high frequency ($> 30\%$ of the total probability mass) is too large and forces even our loglog loss minimizing distribution to perform poorly –due to its poor performance on the French and German words.

Learning $\mathbf{q}_1$ and $\mathbf{q}_2$: We trained the candidate distribution $\mathbf{q}_1$ by minimizing log loss for a basic character trigram model. Minimizing log loss here simply corresponds to setting the trigram probabilities to their relative frequencies in the dataset. These frequencies were computed via a scan over all words in the dataset, taking into account the word frequencies. Note that we have full access to the target $\mathbf{p}$ and thus $\mathbf{q}_1$ exactly minimizes $\ell(\mathbf{q}; \mathbf{p}) = \mathbb{E}_{x \sim \mathbf{p}} \left[\ln \frac{1}{q_x} \right]$ over all trigram models.

We trained $\mathbf{q}_2$ by distorting the optimization to place higher weight on the head of the distribution. In particular, we let $\bar{\mathbf{p}}$ be the distribution with $\bar{p}_x \propto p_x^\alpha$ for $\alpha = 1.4$. and minimized log loss over $\bar{\mathbf{p}}$. We saw similar performance for $\alpha \in [1.3, 2]$. Below this range, there was not significant difference between $\mathbf{q}_1$ and $\mathbf{q}_2$. Above this range, $\mathbf{q}_2$ placed very large mass on the head of the distribution, e.g., outputting the most common word `the` with probability $\geq .40$.

Results: Our results are summarized in Figure 1. We can see that $\mathbf{q}_2$ seems to give more natural word samples and, while it achieves worse log loss than $\mathbf{q}_1$ (it must since $\mathbf{q}_1$ minimizes this loss over all trigram models), it achieves better log log loss. This indicates that in this setting, the log log loss may be a more appropriate measure to optimize. Our approach to training $\mathbf{q}_2$ via a reweighting of $\mathbf{p}$ can be viewed a heuristic for minimizing log log loss. Developing better algorithms for doing this, especially under the constraint that $\mathbf{q}_2$ is (approximately) calibrated is an interesting direction.

One way to see the improved performance of $\mathbf{q}_2$ is that its cumulative distribution more closely matches that of $\mathbf{p}$. See plot in Figure 1. Overall $\mathbf{p}$ places 87.77% of its mass on the English words in the input distribution. $\mathbf{q}_1$ places 45.56% of its mass on these words and $\mathbf{q}_2$ places 83.40% of its mass on them. Note that the cumulative distribution plot and these statistics are *deterministic*, since $\mathbf{q}_1$ and $\mathbf{q}_2$ are trained by exactly minimizing log loss over the distributions $\mathbf{p}$ and $\bar{\mathbf{p}}$ without sampling. Thus no error bars are shown.

Below we show an extended sampling of words from $\mathbf{q}_1$, $\mathbf{q}_2$ and $\mathbf{p}$, evidencing $\mathbf{q}_2's$ superior performance on the task of generating natural English words. In this single run, e.g., $\mathbf{q}_1$ generates 6 distinct commonly used English words $\{$`and`, `the`, `why`, `soon`, `caps`, `of`$\}$. $\mathbf{q}_2$ generates 10: $\{$`all`, `the`, `which`, `on`, `take`, `and`, `be`, `in`, `of`, `he`$\}$. $\mathbf{p}$ generates 19, all with the except of the German word `verweigert`. More quantitatively, in a run of 10000 random samples, $\mathbf{p}$ generates 2497 distinct English words (the word distribution is very skewed so many duplicates of common words are generated). In comparison, $\mathbf{q}_1$ generates 815 distinct words and $\mathbf{q}_2$ generates 957.

Of course there are many methods of evaluating the performance of $\mathbf{q}_1$ and $\mathbf{q}_2$, which generally will be application specific. Our experiments are designed to give just a simple example, motivating the idea that minimizing log loss may not always be the optimal choice, and, like in classification and regression, there is room for alternative loss functions to be considered.

Samples from $\mathbf{q}_1$	**Samples from $\mathbf{q}_2$**	**Samples from $\mathbf{p}$**
and	all	old
tiest	the	verweigert
rike	which	five
agal	nesell	common
the	on	ny
itunge	whostionespirs	significance
cand	the	friend
ho	take	i
aren	the	with
why	and	museum
soon	be	the
ca	frould	without
caps	in	in
der	the	ethan
connestand	the	pointed
of	goich	def
per	of	down
shicy	ithe	the
theared	he	sky
introt	ong	the

H Calibration Definition

In this section we give further discussion on our definition of calibration. Most typically in forecasting, calibration is a property of a *sequence* of forecasts $\mathbf{q}^{(1)}, \ldots$, evaluated against a *sequence* of samples $x^{(1)}, \ldots$. So our definition may require some background. First, we give a justification based on $\mathbf{q}$ as a coarsening of $\mathbf{p}$. Then, we show how formalizations of calibration for sequences of forecasts can be related to our definition.

As a coarsening. One way to view the forecast $\mathbf{q}$ is as a coarsening of $\mathbf{p}$ in the sense of assigning probabilities to certain events $B_\alpha \subseteq \mathcal{X}$, but remaining agnostic as to the relative probabilities of various elements of B_α, assigning all of them equal weight α. By dividing $\mathcal{X}$ into maximal pieces B_α on which $\mathbf{q}$ is piecewise uniform, in this way one obtains that $\mathbf{q}$ is literally a coarsening of $\mathbf{p}$ if $\mathbf{p}(B_\alpha) = \mathbf{q}(B_\alpha)$ for each piece (as the pieces partition $\mathcal{X}$). This is our definition of calibration.

This directly captures the typical informal definition of calibration as "events that are assigned probability β occur a β-fraction of the time", where the pieces B_α are the events and $\beta = \mathbf{q}(B_\alpha) = \mathbf{p}(B_\alpha)$ are the probabilities assigned to them.

It is also consistent with standard formalizations of calibration for sequences (see below), as if $x^{(s)} \sim \mathbf{p}$ i.i.d. each round and $\mathbf{q}^{(s)} = \mathbf{q}$ each round, one has that in the limit, each piece B_α will be represented as often as $\mathbf{q}$ predicts.

Sequences of forecasts. Calibration of sequences can be formalized, for example, as follows. If each $x^{(t)} \in \mathcal{X} = \{0, 1\}$, then we can let R_t be the set of rounds $s \leq t$ where $x^{(s)} = 1$ and $S_t(\mathbf{q})$ be the set of rounds $s \leq t$ where $\mathbf{q}^{(s)} = \mathbf{q}$. In this case, the sequence is termed *calibrated* if, on rounds where $\mathbf{q}$ was predicted, the fraction of times that $x^{(s)} = 1$ converges to q_1:

$$\forall \mathbf{q} : \qquad \lim_{t \to \infty} \frac{|S_t(\mathbf{q}) \cap R_t|}{|S_t(\mathbf{q})|} = q_1.$$

One way to obtain our definition is by "flattening" this one: let there be a finite number of rounds and suppose $\mathbf{p}, \mathbf{q}$ are probability distributions over rounds (so $\mathbf{p}$ will pick exactly one round to occur, and $\mathbf{q}$ assigns a binary prediction to each round). In this case we can let $S(\alpha) = \{t : q_t = \alpha\}$ be the set of rounds assigned a probability α by the forecast, then naturally the round $t \sim \mathbf{p}$ lies in this set with probability $\mathbf{p}(S(\alpha))$. So the flattened definition of calibration requires that for each α, $\mathbf{p}(S(\alpha)) = \mathbf{q}(S(\alpha))$, which is exactly our definition.

Our definition can also be obtained as described above by letting $\mathcal{X}$ be general, letting $\mathbf{q}$ be forecast on each round while $x^{(s)} \sim \mathbf{p}$ i.i.d. each round. If one interprets $\mathbf{q}$ as a distribution over events B_α that partition $\mathcal{X}$, one obtains the requirement that in the limit $\mathbf{p}(B_\alpha) = \mathbf{q}(B_\alpha)$ for each α.

I Strong Properness in ℓ_2 Norm

Our criteria can be extended to utilize different distance measures than our choice of ℓ_1 or total variation distance. However, justifying and investigating other measures requires further work. In particular, this section shows why a choice of ℓ_2 distance can be problematic.

Following our main definitions, one can define a loss to be strongly proper in ℓ_2 if, for all $\mathbf{p}, \mathbf{q}$,

$$\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) \geq \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_2^2.$$

In particular, consider the quadratic loss $\ell(\mathbf{q}, x) = \frac{1}{2} \|\boldsymbol{\delta}^x - \mathbf{q}\|_2^2$, which can be shown to be 1-strongly-proper in ℓ_2 (Corollary 3). However, the usefulness of this guarantee can be limited, as the following example shows.

Proposition 2. *Given a* 1*-strongly proper loss in* ℓ_2 *norm,* $\mathbf{q}$ *can assign probability zero to the entire support of* $\mathbf{p}$*, yet have expected loss within* $\frac{2}{N}$ *of optimal.*

Proof. Let $\mathcal{X} = \{1, \dots, N\}$ for even N. Let $\mathbf{p}$ be uniform on $\{1, \dots, \frac{N}{2}\}$ and let $\mathbf{q}$ be uniform on $\{\frac{N}{2} + 1, \dots, N\}$.

The point is that for any such "thin" distributions (small maximum probability), their ℓ_2 norms $\|\mathbf{p}\|, \|\mathbf{q}\|$ are vanishing and by the triangle inequality so is the distance $\|\mathbf{p} - \mathbf{q}\|$ between them.

In this example, $\|\mathbf{p} - \mathbf{q}\|_2^2 = N \left(\frac{2}{N}\right)^2 = \frac{4}{N}$. So strong properness only guarantees that the difference in loss is $\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) \geq \frac{2}{N}$. In fact, this is exactly matched by the quadratic loss, where the difference in expected score (the Bregman divergence of the two-norm) is exactly $\frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_2^2 = \frac{2}{N}$. □

Thus, strongly proper losses in ℓ_2 can converge to optimal expected loss at the rapid rate of $O(\frac{1}{N})$ even when making completely incorrect predictions.

J Strongly Proper Losses and Scoring Rules on the Full Domain

In this section, for completeness, we investigate the strongly proper criterion in the traditional setting of proper losses (equivalently, scoring rules). The main result is that, just as (strictly) proper losses are Bregman divergences of (strictly) convex functions, so are *strongly* proper losses Bregman divergences of *strongly* convex functions. We derive some non-local strongly proper losses. These results may be of independent interest.

Terminology. Given a function $f : \mathbb{R}^d \to \mathbb{R}$, the vector $v \in \mathbb{R}^d$ is a *supergradient* of f at z if for all z', we have $f(z') \leq f(z) + v \cdot (z' - z)$. (In other words, there is a tangent hyperplane lying above f at z with slope v.) A function is *concave* if it has at least one supergradient at every point. (If exactly one, it is differentiable.) In this case, use $df(z)$ to denote a choice of a supergradient of f at z.

Given a concave f, the *divergence function of* f is

$$D_{-f}(z, z') := [f(z') + df(z') \cdot (z - z')] - f(z),$$

the gap between $f(z)$ and the linear approximation of f at z' evaluated at z. The reason for this notation is that D_{-f} is the Bregman divergence of the convex function $-f$.

Definition 9 (Strongly Concave)**.** A function $f : \mathbb{R}^d \to \mathbb{R}$ is β*-strongly concave* with respect to a norm $\|\cdot\|$ if for all z, z',

$$D_{-f}(z, z') \geq \frac{\beta}{2} \|z - z'\|^2 .$$

J.1 Background: proper loss characterization

We first recall some background from theory of proper scoring rules, phrased in the loss setting. Given a loss $\ell(\mathbf{q}, x)$, the expected loss function is $H_\ell(\mathbf{p}) = \ell(\mathbf{p}; \mathbf{p})$. The following classic characterization says that (strict) properness of ℓ is equivalent to (strict) concavity of H_ℓ.

Theorem 8 ([21, 24, 15])**.** *ℓ is a (strictly) proper loss if and only if H_ℓ is (strictly) concave. If so, we must have*

$$\ell(\mathbf{q}, x) = H_\ell(\mathbf{q}) + dH_\ell(\mathbf{q}) \cdot (\boldsymbol{\delta}^x - \mathbf{q})$$

where $dH_\ell(\mathbf{q})$ is any supergradient of H_ℓ at $\mathbf{q}$ and $\boldsymbol{\delta}^x$ is the point mass distribution on x.

Corollary 1. *The expected loss of $\mathbf{q}$ under true distribution $\mathbf{p}$ is the linear approximation of H_ℓ at $\mathbf{q}$, evaluated at $\mathbf{p}$:*

$$\ell(\mathbf{q}; \mathbf{p}) = H_\ell(\mathbf{q}) + dH_\ell(\mathbf{q}) \cdot (\mathbf{p} - \mathbf{q}).$$

Corollary 2. *When the true distribution is p, the improvement in expected loss for reporting $\mathbf{p}$ instead of $\mathbf{q}$ is the divergence function of H_ℓ (the Bregman divergence of $-H_\ell$), i.e.*

$$\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) = D_{-H_\ell}(\mathbf{p}, \mathbf{q}).$$

Example 5. Recall from Example 1 the log loss $\ell(\mathbf{q}, x) = \ln \frac{1}{\mathbf{q}}$ has expected loss equal to Shannon entropy. The associated Bregman divergence is the KL-divergence, so the difference in expected log loss between $\mathbf{p}$ and $\mathbf{q}$ under true distribution $\mathbf{p}$ is $KL(\mathbf{p}, \mathbf{q}) := \sum_x p_x \ln \frac{p_x}{q_x}$. The quadratic loss has expected loss $H_{\text{quad}}(\mathbf{p}) = \frac{1}{2} - \frac{1}{2}\|\mathbf{p}\|_2^2$, so the associated Bregman divergence is $D_{-H_{\text{quad}}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_2^2$.

The above are all well-known, although in the literature on proper scoring rules everything is negated (a score is used equal to negative loss, the expected score is convex, etc.).

J.2 Strongly concave functions and strong properness

Given the above characterization and our (carefully chosen) definition of *strongly proper*, the classic characterization of proper losses extends easily:

Theorem 9. *A proper loss function ℓ is β-strongly proper (with respect to a norm) if and only if H_ℓ is β-strongly concave (with respect to that norm).*

Proof. We have $\ell(\mathbf{q}; \mathbf{p}) - \ell(\mathbf{p}; \mathbf{p}) = D_{-H_\ell}(\mathbf{p}, \mathbf{q})$ by Corollary 2. H_ℓ is β-strongly concave if and only if $D_{-H_\ell}(\mathbf{p}, \mathbf{q}) \geq \frac{\beta}{2}\|\mathbf{p} - \mathbf{q}\|$ for all $\mathbf{p}, \mathbf{q}$, which is the condition that ℓ is β-strongly proper. ☐

Though the proof is trivial once the definitions are set up and followed through, the statement is powerful. It completely characterizes the proper loss functions satisfying that, if $\mathbf{q}$ is significantly wrong (far from $\mathbf{p}$), then its expected loss is significantly worse. It also gives an immediate recipe for constructing such losses: Start with any concave function $H(\mathbf{q})$ that is strongly concave in your norm of choice, and set $\ell(\mathbf{q}, x) = H(\mathbf{q}) + dH(\mathbf{q}) \cdot (\boldsymbol{\delta}^x - \mathbf{q})$. All strongly proper losses satisfy this construction for some such H.

J.3 Known examples

Recall that the log scoring rule's expected loss function is Shannon entropy. Hence, the fact that log loss is 1-strongly-proper (Example 2) turns out to be equivalent to the statement that Shannon entropy is 1-strongly convex in ℓ_1 norm. As described in Section 3, this fact (perhaps surprisingly) is equivalent to Pinsker's inequality.

However, ℓ_1-strong properness seems difficult to satisfy over the simplex. In particular,

Proposition 3. *The quadratic scoring rule is not strongly proper in ℓ_1 norm.*

Proof. Consider $\mathbf{q}$ as the uniform distribution and let $p_x \in \frac{1 \pm \epsilon}{N}$, such that $\|\mathbf{p} - \mathbf{q}\|_1 = \epsilon$. Then $\ell(q; p) - \ell(p; p) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_2^2 = \frac{1}{2}(N)\left(\frac{\epsilon}{N}\right)^2 = \frac{\epsilon^2}{2N}$. As $N \to \infty$, this difference in loss goes to zero while $\|\mathbf{p} - \mathbf{q}\|_1 = \epsilon$, so there is no fixed β such that the loss is β-strongly proper. ☐

We can show that it is strongly proper in ℓ_2 norm. However, the usefulness of ℓ_2 strong properness is less clear, as is demonstrated in Appendix I.

Lemma 6. *The function $-\frac{1}{2}\|\mathbf{p}\|_2^2$ is 1-strongly concave with respect to the ℓ_2 norm.*

Proof. The associated Bregman divergence is $\frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_2^2$, which is equal to $\frac{1}{2}\|\mathbf{p}-\mathbf{q}\|_2^2$, so it is 1-strongly convex in ℓ_2 norm. □

Corollary 3. *The quadratic loss is 1-strongly proper with respect to the ℓ_2 norm.*

J.4 New proper losses

Because the ℓ_1 norm is especially preferred when measuring distances between probability distributions, we seek losses that are 1-strongly proper with respect to the L_1 norm. By the characterization of Theorem 9, this is equivalent to seeking ℓ_1 β-strongly-convex functions of probability distributions.

Lemma 7. *Let $M \in \mathbb{R}^{N\times N}$ be the* negative *of the Hessian of a function $H : \Delta_{\mathcal{X}} \to \mathbb{R}$. Then H is β-strongly concave in ℓ_1 norm if and only if*

$$\min_{w:\|w\|_1=1} w^\intercal M w \geq \beta$$

Proof. M is the Hessian of the (presumably convex) function $-H$. □

We focus on *separable, symmetric* concave functions: $H(\mathbf{q}) = \sum_x h(q_x)$ for some concave function h. In this case the Hessian of H is a diagonal matrix with (x,x) entry $\frac{d^2h(z)}{dz^2}$. Call its negative M as in Lemma 7 and for convenience later, let us define $f(z)$ as

$$\frac{1}{f(z)} := \frac{-d^2h(z)}{dz^2}.$$

Then by Lemma 7, $H(\mathbf{q})$ is β-strongly concave if

$$\begin{aligned}\beta &\leq \min_{w:\|w\|_1=1} w^\intercal M w \\ &= \min_{w:\|w\|_1=1} \sum_x \frac{w_x^2}{f(q_x)}.\end{aligned}$$

This is solved by setting $w_x \propto f(q_x)$, where the normalizing constant is $C := \sum_x f(q_x)$. So we have

$$\begin{aligned}\beta &\leq \sum_x \left(\frac{f(q_x)}{C}\right)^2 \frac{1}{f(q_x)} \\ &= \frac{1}{C^2}\sum_x f(q_x) \\ &= \frac{1}{C} \\ &= \frac{1}{\sum_x f(q_x)}.\end{aligned}$$

So for 1-strong concavity, we require $\sum_x f(q_x) \leq 1$ for all $\mathbf{q}$. Now choose $f(q_x) = q_x^{1+\alpha}$.

- If $\alpha < 0$, then $\sum_x f(q_x)$ can be arbitrarily large and the resulting function is not strongly concave in ℓ_1 norm.
- If $\alpha = 0$, then we have $\frac{d^2h(z)}{dz^2} = \frac{-1}{z}$ and we recover $h(z) = z\ln(\frac{1}{z})$, which gives H as Shannon entropy; the log scoring rule.
- If $\alpha \geq 1$, we get $h(z)$ is unbounded on $[0,1]$, so we obtain an expected loss function that is unbounded on the simplex.
- For $0 < \alpha < 1$, we get a class of apparently-new proper loss functions that are 1-strongly proper. Here $\frac{d^2h(z)}{dz^2} = \frac{-1}{z^{1+\alpha}}$, so $h(z) = z^{1-\alpha}$ and $H(z) = \sum_x q_x^{1-\alpha}$.

In particular, for the last class, we identify the appealing case $\alpha = 0.5$. It gives the following "inverse root" loss function:

- $H(\mathbf{q}) = 2\sum_x \sqrt{q_x}$.
- $\ell(\mathbf{q}, x) = \frac{1}{\sqrt{q_x}} + \sum_{x'} \sqrt{q_{x'}}$.
- $\ell(\mathbf{q}; \mathbf{p}) = \sum_x \frac{1}{\sqrt{q_x}} (p_x + q_x)$.
- $D_{-H}(\mathbf{p}, \mathbf{q}) = \sum_x \frac{1}{\sqrt{q_x}} \left(\sqrt{p_x} - \sqrt{q_x}\right)^2$.

We are not aware of this loss having been used before, but it seems to have nice properties. There is an apparent similarity to the squared Hellinger distance $H(\mathbf{p}, \mathbf{q})^2 := \frac{1}{2} \sum_x \left(\sqrt{p_x} - \sqrt{q_x}\right)^2$, we are not aware of a closer formal connection. For example, Hellinger distance is symmeteric.