How Flawed Is ECE? An Analysis via Logit Smoothing

Muthu Chidambaram* [1] **Holden Lee*** [2] **Colin McSwiggen*** [3] **Semon Rezchikov** [4]

Abstract

Informally, a model is calibrated if its predictions are correct with a probability that matches the confidence of the prediction. By far the most common method in the literature for measuring calibration is the expected calibration error (ECE). Recent work, however, has pointed out drawbacks of ECE, such as the fact that it is discontinuous in the space of predictors. In this work, we ask: how fundamental are these issues, and what are their impacts on existing results? Towards this end, we completely characterize the discontinuities of ECE with respect to general probability measures on Polish spaces. We then use the nature of these discontinuities to motivate a novel *continuous, easily estimated* miscalibration metric, which we term *Logit-Smoothed ECE (LS-ECE)*. By comparing the ECE and LS-ECE of pre-trained image classification models, we show in initial experiments that binned ECE closely tracks LS-ECE, indicating that the theoretical pathologies of ECE may be avoidable in practice.

1. Introduction

The prevalence of machine learning across domains has increased drastically over the past few years, spurred by significant breakthroughs in deep learning for computer vision (Ramesh et al., 2022) and language modeling (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023). Consequently, the underlying deep learning models are increasingly being evaluated for critical use cases such as predicting medical diagnoses (Elmarakeby et al., 2021; Nogales et al., 2021) and self-driving (Hu et al., 2023). In these latter cases, due to the risk associated with incorrect decision-making, it is crucial not only that the models be accurate, but also that they have proper predictive uncertainty.

*Equal contribution [1]Department of Computer Science, Duke University [2]Johns Hopkins University [3]New York University [4]Princeton University. Correspondence to: Muthu Chidambaram <muthu@cs.duke.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

This desideratum is formalized via the notion of *calibration* (Dawid, 1982; DeGroot & Fienberg, 1983), which codifies how well the model-predicted probabilities for events reflect their true frequencies conditional on the predictions. For example, in a medical context, a model that yields the correct diagnosis for a patient 95% of the time when it predicts a probability of ≈ 0.95 for that diagnosis can be considered to be calibrated.

The analysis of whether modern deep learning models are calibrated can be traced back to the influential work of Guo et al. (2017), which showed that recent models exhibit calibration issues not present in earlier models; in particular, they are overconfident when they are incorrect. These findings have been corroborated by a large body of subsequent work in which several training and post-training modifications have been proposed in order to improve calibration (Lakshminarayanan et al., 2017; Kumar et al., 2018; Thulasidasan et al., 2019; Müller et al., 2020; Wang et al., 2021; Wang & Golebiowski, 2023).

However, the validity of these results depends on having an appropriate measure of calibration. The canonical measure of calibration in the machine learning literature has been the *Expected Calibration Error (ECE)* and its binned variants (Naeini et al., 2014; Nixon et al., 2019), and indeed all of the aforementioned works report ECE in some capacity.

Unfortunately, several works have pointed out (seemingly) significant drawbacks of ECE. First, it is discontinuous as a function of the model being considered. In other words, small changes to model predictions can cause large jumps in the ECE (Kakade & Foster, 2008; Foster & Hart, 2018; Błasiok et al., 2023; Błasiok & Nakkiran, 2023). Second, it is not possible to efficiently estimate from samples (Arrieta-Ibarra et al., 2022; Lee et al., 2022), and binned variants can be sensitive to the choice of bin width (Nixon et al., 2019; Kumar et al., 2019; Minderer et al., 2021).

As a result of these drawbacks, a number of authors have recently proposed alternatives to ECE that enjoy better theoretical properties (Arrieta-Ibarra et al., 2022; Lee et al., 2022; Błasiok & Nakkiran, 2023; Błasiok et al., 2023). Despite these proposals, as noted in Błasiok & Nakkiran (2023), ECE continues to be the main metric reported in very recent studies. Błasiok & Nakkiran (2023) hypothesize that the reason for this fact is that ECE can be easily visualized and

interpreted via reliability diagrams.

In this work, we propose an alternative explanation that may serve to justify the continued predominance of ECE: besides the fact that ECE is historically established and well supported by standard codebases, the pathologies of ECE are not encountered in practice due to noise inherent to the data and model training process. This paper formalizes one simple variant of such a noise model. In this model, we show that the addition of noise makes ECE continuous and leads to an effective estimation scheme, which moreover does not appreciably differ from direct estimates of ECE performed via binning.

Informed by this perspective, we aim to answer the following questions in this work:

- Can we characterize the points of discontinuity of ECE?
- Can these discontinuities be eliminated by a simple modification of the miscalibration metric?
- Does the discontinuous behavior of ECE actually pose a problem for estimating the calibration of real-world deep learning models?

1.1. Summary of Main Contributions

Our main contributions towards answering the above questions are as follows.

1. In Section 3, we completely characterize the discontinuities of ECE in a very general setting. We illustrate in detail how these considerations apply in the case of discrete data distributions with finite support, and we show that in this case the discontinuities are a measure zero set. The case of continuous distributions is more subtle, and intuitions from the discrete setting do not always carry over; however, we nevertheless provide a necessary and sufficient condition for discontinuity in the case of arbitrary distributions of data taking values in a Polish space, and we show that in this setting the ECE is always a lower semicontinuous functional.
2. Building on the ideas of Section 3, we derive a modified ECE measure in Section 4 which we term *Logit-Smoothed ECE (LS-ECE)*. We show that the LS-ECE is continuous in the space of predictors, for *any* data distribution. Our results rely on establishing strong connections between convergence of the underlying joint probability measures in total variation on one hand, and continuity of the ECE functional on the other, which may be of independent interest.
3. We further propose a consistent estimator of the LS-ECE in Section 5, and show that our estimator can both be efficiently estimated and implemented. As an additional consequence of our estimation result, we show that LS-ECE can be used to produce a consistent estimator of the true ECE when the predictive distribution satisfies mild regularity conditions, which to the best of our knowledge is a stronger result than any pre-existing consistency results for ECE.
4. Lastly, in Section 6, we verify empirically that LS-ECE is continuous even when ECE is not, and also show that for the standard image classification benchmarks of CIFAR-10, CIFAR-100, and ImageNet, both ECE and LS-ECE produce near identical results across various models — indicating that the theoretical pathologies of ECE may not pose an issue in practice.

We note that, in the process of analyzing ECE, we have proposed yet another competing notion in the form of LS-ECE. We wish to stress that we are not trying to claim that LS-ECE is a "better" measure of calibration than recently proposed alternatives, and in fact it shares much in common with the SmoothECE proposal of Błasiok & Nakkiran (2023). Rather, we view LS-ECE as a useful theoretical and empirical tool for sanity-checking ECE in a given setting — our experiments in Section 6.2 use it to suggest that reported ECE results may not be particularly brittle. Furthermore, we hope that the theoretical framework under which we formulate LS-ECE will prove useful in future analyses of calibration.

1.2. Related Work

Estimation of ECE. Perhaps the most common way to estimate ECE in the literature is by binning predictions into uniformly sized bins (Naeini et al., 2014). A similarly popular approach is to bin predictions using equal mass bins, which leads to Adaptive Calibration Error (Nixon et al., 2019). These binning approaches are, however, known not to be consistent (Vaicenavicius et al., 2019), and follow-up works have modified them further via debiasing schemes (Kumar et al., 2019; Roelofs et al., 2022). An alternative to binning is estimating ECE via kernel density/regression estimators (Bröcker, 2008; Zhang et al., 2020; Popordanoska et al., 2022), which trade off bin selection with bandwidth selection.

Alternatives to ECE. Recent work has pursued several different directions for developing alternatives to ECE. These include, but are not limited to, proper scoring rules (Gneiting et al., 2007; Gneiting & Raftery, 2007), estimating miscalibration using splines (Gupta et al., 2021), isotonic regression (Dimitriadis et al., 2021), hypothesis tests for miscalibration (Lee et al., 2022), and cumulative plots comparing labels to predicted probabilities (Arrieta-Ibarra et al., 2022). Błasiok et al. (2023) detail several more alternatives to ECE,

along with a theoretical framework based on distance to the nearest calibrated predictor that justifies the use of these alternatives in practice. Very recently, Błasiok & Nakkiran (2023) have proposed a kernel-smoothed ECE that satisfies the constraints of the framework of Błasiok et al. (2023) but still maintains the interpretability of ECE. The approach we take in this paper was developed independently and in parallel, and shares similarities with the approach of Błasiok et al. (2023) that we point out in Section 5.

2. Background

Notation. Given $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, 2, ..., n\}$. For a function $g : \mathbb{R}^n \to \mathbb{R}^m$ we use $g^i(x)$ to denote the i^{th} coordinate function of g. For a probability distribution π we use $\text{supp}(\pi)$ to denote its support. Additionally, if π corresponds to the joint distribution of two random variables X and Y (e.g. data and label), we use π_X and π_Y to denote the respective marginals, and $\pi_{X|Y=y}$ and $\pi_{Y|X=x}$ to denote the conditional distributions. For a random variable X that has a density with respect to the Lebesgue measure, we use $p_X(x)$ to denote its density. For a general random variable X, we use $d\mathbb{P}_X$ to denote its associated probability measure, and use $d\mathbb{P}_X/d\mathbb{P}_Z$ to denote the Radon–Nikodym derivative when $X \ll Z$ (i.e. X is absolutely continuous with respect to Z). We use d_{TV} to denote the total variation distance between probability measures. Lastly, we use $\text{Uni}([a, b])$ to denote the uniform distribution on $[a, b]$ and $\mathcal{N}(\mu, \sigma^2)$ to denote the Gaussian distribution with mean μ and variance σ^2.

We first consider calibration in the context of binary classification and then discuss generalizations to multi-class classifications. For a data distribution π on $\mathbb{R}^d \times \{0, 1\}$, we say that a predictor $g : \mathbb{R}^d \to [0, 1]$ is calibrated if it satisfies the regular conditional probability condition $\mathbb{E}_{(X,Y)\sim\pi}[Y \mid g(X) = p] = p$. This condition corresponds to the idea that for all instances on which our model g predicts probability p, the correct label of those instances is actually 1 with probability p.

The Expected Calibration Error (ECE) with respect to the distribution π is then defined to be the expected absolute deviation from this condition:

$$\text{ECE}_\pi(g) \triangleq \mathbb{E}_{(X,Y)\sim\pi}\left[\left|\mathbb{E}[Y \mid g(X)] - g(X)\right|\right]. \quad (2.1)$$

Although there are several ways to generalize ECE to the multi-class setting, perhaps the most reported generalization in the literature is top-class (or confidence calibration) ECE. Namely, for a k-class classification problem in which we have a predictor $g : \mathbb{R}^d \to \Delta_k$, where Δ_k denotes the simplex of probability measures on a set with k elements, the top-class ECE simply corresponds to computing the ECE with respect to the highest probability prediction:

$$\mathbb{E}\left[\left|\mathbb{E}[Y \in \operatorname*{argmax}_i g^i(X) \middle| \max_i g^i(X)] - \max_i g^i(X)\right|\right]. \quad (2.2)$$

This definition is equivalent to considering the binary ECE with respect to a modified predictor f and a distribution $(X', Y') \sim \pi'$ defined such that $Y' = \mathbb{1}_{Y \in \text{argmax}_i\, g^i(X)}$ and $f(X') = \max_i g^i(X)$. As such, any modifications made to the binary version of ECE can be lifted to the multi-class setting via the top-class formulation of (2.2), so we will henceforth just work with the binary version as defined in (2.1).

In practice, $\text{ECE}_\pi(g)$ is estimated via binning. Given a set of data points $\{(x_1, y_1), ..., (x_n, y_n)\}$, one specifies a partition $B_1, B_2, ..., B_m$ of $[0, 1]$ and then computes

$$\text{ECE}_{\text{BIN},\pi}(g) \triangleq \sum_{j=1}^{m} \frac{|B_j|}{n} |\bar{y}(B_j) - \bar{g}(B_j)|, \quad (2.3)$$

where $\bar{g}(B_j)$ corresponds to the average of all $g(x_i)$ such that $g(x_i) \in B_j$, and $\bar{y}(B_j)$ denotes the average over corresponding labels. In the multi-class case, one simply replaces $\bar{y}(B_j)$ with average accuracy and $\bar{g}(B_j)$ with average top-class probability.

3. Continuity Properties of ECE

Having provided the necessary background regarding ECE, we now analyze its continuity properties. We begin first with the case where $|\text{supp}(\pi_X)| = n < \infty$, i.e. discrete data distributions with finite support. In this case, we can provide a necessary and sufficient condition for ECE_π to be discontinuous at g, which implies that g can only be a point of discontinuity if it predicts the same probability at multiple points that each have positive measure under π_X. We subsequently show, however, that this intuition does not extend to the case where π_X is supported on a more general (infinite) set. Nevertheless, with careful analysis, we can still extend the necessary and sufficient condition from the discrete case to a much more general setting.

3.1. Discrete Distributions

To get a sense of the continuity issues that arise with ECE_π, we introduce an example adapted from (Błasiok et al., 2023) that we will refer to multiple times throughout the rest of the paper.

Definition 3.1. [2-Point Distribution] Let π be the distribution on $\{-1/2, 1/2\} \times \{0, 1\}$ such that $\pi_Y(0) = \pi_Y(1) = 1/2$ and $\pi_{X|Y=y}(x) = \mathbb{1}_{x=y-1/2}$ (i.e. $X \mid Y = y$ is a point mass on $y - 1/2$).

It is straightforward to see that the predictor $g(x) = 1/2$ satisfies $\mathrm{ECE}_\pi(g) = 0$ for π as in Definition 3.1. However, perturbing g such that $g(-1/2) = 1/2 - \varepsilon$ and $g(1/2) = 1/2 + \varepsilon$ yields $\mathrm{ECE}_\pi(g) = 1/2 - \varepsilon$ (for $\varepsilon \in (0, 1/2)$). The important idea here is that we split the level sets of g by making an arbitrarily small perturbation, which causes $\mathbb{E}[Y \mid g(X)]$ to jump from $1/2$ to 1.[1]

In fact, we can show that *all* discontinuities of ECE_π for finitely supported π occur at predictors that have non-singleton level sets with positive measure under π_X. This is part of the following full characterization of the discontinuities of ECE_π for discrete π. (When we refer to discrete data distributions below, we always mean distributions with finite support.)

Theorem 3.2. *[Discontinuities for Discrete ECE] Let π be any distribution such that $\mathrm{supp}(\pi_X) = [n]$ for an arbitrary positive integer n, and let $g^*(x) = P(Y = 1 \mid X = x)$ denote the ground truth conditional distribution. Then the set of discontinuities of ECE_π (in the space of predictors $g : [n] \to [0, 1]$ endowed with the ℓ^∞ norm) is exactly the set of g such that there exists $m \in [n]$ with $P(X = m) \neq 0$ and*

$$\big|g^*(m) - g(m)\big| \neq \big|\mathbb{E}[Y \mid g(X) = g(m)] - g(m)\big|. \quad (3.1)$$

Note that the choice of norm on the space of predictors makes no difference in this case, since when $\mathrm{supp}(\pi_X)$ is finite g is just an n-dimensional vector, and all norms on $\mathbb{R}^n$ are equivalent. As promised, the proof of Theorem 3.2 relies on the following lemma, which shows that a discontinuity can only occur if g predicts identical probabilities for at least two distinct points in $\mathrm{supp}(\pi_X)$.

Lemma 3.3. *Let $S(g, p) = \{j \in [n] : g(j) = p \text{ and } P(X = j) \neq 0\}$. If $P(X = m) \neq 0$ and (3.1) holds, then $|S(g, g(m))| > 1$.*

Proof. Clearly $|S(g, g(m))| \geq 1$ since $m \in S(g, g(m))$, so it suffices to show that (3.1) fails if we assume $|S(g, g(m))| = 1$. Under this assumption, we have

$$\begin{aligned}\mathbb{E}[Y \mid g(X) = g(m)] &= \frac{\sum_{j \in S(g,g(m))} \pi(j) g^*(j)}{\sum_{j \in S(g,g(m))} \pi(j)} \\ &= \frac{\pi(m) g^*(m)}{\pi(m)} = g^*(m), \quad (3.2)\end{aligned}$$

so (3.1) indeed fails. □

A consequence of Lemma 3.3 is the following corollary, which shows that the set of discontinuities for ECE_π in the discrete case is negligible.

[1] How the level sets of a predictor impact different loss functions applied to that predictor has also been studied more generally in the literature on scoring rules, in particular in the work of Kull & Flach (2015) which discusses the notion of a grouping loss.

Corollary 3.4. *If $|\mathrm{supp}(\pi)| = n$, then the set of predictors g (identified with vectors in $\mathbb{R}^n$) at which ECE_π is discontinuous has measure zero with respect to the Lebesgue measure on $\mathbb{R}^n$.*

Proof. By Lemma 3.3, the set of discontinuities is a subset of the union of all sets $S_{i,j} = \{g : g(i) = g(j)\}$, where $i, j \in [n]$ and $i \neq j$. Since each $S_{i,j}$ has measure zero and we are considering a finite union of such sets, the set of discontinuities has measure zero. □

3.2. General Probability Measures

Unfortunately, a key piece of intuition from the discrete setting fails to extend to the continuous case, as the following proposition shows.

Proposition 3.5. *Take π with $\mathrm{supp}(\pi_X) = [0, 1]^2$ such that $\pi_{X_1} = \mathrm{Uni}([0, 1])$,*

$$\pi_{X_2 | X_1 = x_1} = x_1 \mathrm{Uni}([0.5, 1]) + (1 - x_1)\mathrm{Uni}([0, 0.5)),$$

and $\mathbb{P}(Y = 1 \mid X_2 = x_2) = \mathbb{1}_{x_2 \geq 0.5}$. Then ECE_π is discontinuous at the predictor $g(x) = x_1$, despite the fact that g has no level sets of positive measure.

The proof of Proposition 3.5, as well as the tricky aspect of the continuous case, rests on the fact that we can make a small perturbation to g at every x that greatly changes the behavior of g despite each point x being measure zero. In order to tackle the additional subtleties that arise when dealing with continuously distributed data, we take a completely general perspective. For the remainder of this section, we allow the data variable X to take values in an arbitrary Polish space Ω_X, endowed with its Borel σ-algebra $\mathcal{B}(\Omega_X)$. Recall that a Polish space is, by definition, a separable completely metrizable topological space, so that this setting subsumes the case of finite discrete distributions treated above and also includes the case of continuously distributed vector-valued data (corresponding to $\Omega_X = \mathbb{R}^d$). In this general setting, we show that ECE_π is always lower semicontinuous on $L^p(\Omega_X; \pi_X)$ for $1 \leq p \leq \infty$, and we give a precise characterization of its points of discontinuity.

We start with two lemmas. The first is a straightforward and well-known result. The second lemma is, to our knowledge, new. The proof relies on a construction due to Kudô (1974).

Lemma 3.6. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, and let $\mathcal{F} \subseteq \mathcal{G} \subseteq \Sigma$ be sub-σ-algebras. Then for any $f \in L^1(\Omega; \mathbb{P})$, $\|\mathbb{E}[f|\mathcal{F}]\|_1 \leq \|\mathbb{E}[f|\mathcal{G}]\|_1$.*

Lemma 3.7. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, let $f \in L^1(\Omega; \mathbb{P})$, and let $g, g_1, g_2, \ldots$ be real-valued random variables such that $g_n \to g$ in probability. Then $\liminf_{n \to \infty} ||\mathbb{E}[f|g_n]||_1 \geq ||\mathbb{E}[f|g]||_1$.*

With these two lemmas we can now can prove the first main result of this section.

Theorem 3.8. *The functional* ECE_π *is lower semicontinuous on* $L^p(\Omega_X; \pi_X)$ *for* $1 \le p \le \infty$.

Proof. Let $g, g_1, g_2, \ldots \in L^p(\Omega_X; \pi_X)$ and suppose $g_n \to g$. We will show that $\liminf_{n\to\infty} \mathrm{ECE}_\pi(g_n) \ge \mathrm{ECE}_\pi(g)$.

Since $g_n \to g$ in $L^p(\Omega_X; \pi_X)$, we have the convergence $g_n(X) \to g(X)$ of L^p random variables on the background probability space, and in particular, $g_n(X) \to g(X)$ in probability. Then, by Lemma 3.7 and L^p-continuity of conditional expectation, we have

$$\begin{aligned}
\liminf_{n\to\infty} \mathrm{ECE}_\pi(g_n) &= \liminf_{n\to\infty} \big\|\mathbb{E}[Y - g_n(X)|g_n(X)]\big\|_1 \\
&= \liminf_{n\to\infty} \big\|\mathbb{E}[Y - g(X)|g_n(X)] \\
&\quad + \mathbb{E}[g(X) - g_n(X)|g_n(X)]\big\|_1 \\
&= \liminf_{n\to\infty} \big\|\mathbb{E}[Y - g(X)|g_n(X)]\big\|_1 \\
&\ge \big\|\mathbb{E}[Y - g(X)|g(X)]\big\|_1 = \mathrm{ECE}_\pi(g),
\end{aligned}$$

which is the desired result. □

Below we use Theorem 3.8 to prove the necessity direction of our general condition for ECE_π to be continuous at a point in L^p. To prove that this same condition is also sufficient, we will need the following lemma.

Recall that a standard Borel space is a Polish space equipped with its Borel σ-algebra. A measurable bijection between standard Borel spaces is always an isomorphism (i.e., its inverse is also measurable). The Kuratowski isomorphism theorem states that two standard Borel spaces are isomorphic if and only if they have the same cardinality; in particular, every standard Borel space is isomorphic to one of $\mathbb{R}$, $\mathbb{Z}$, or a finite discrete space.

Lemma 3.9. *Let* $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ *be a probability space, where* Ω *is a Polish space and* $\mathcal{B}(\Omega)$ *is its Borel* σ*-algebra. For* $1 \le p \le \infty$ *define*

$$L^p_{\mathrm{inj}}(\Omega; \mathbb{P}) = \big\{\, f \in L^p(\Omega; \mathbb{P}) \;\big|\; f \text{ is almost surely equal to a bijection onto a standard Borel space} \,\big\}.$$

Then $L^p_{\mathrm{inj}}(\Omega; \mathbb{P})$ *is dense in* $L^p(\Omega; \mathbb{P})$.

Putting everything together, we prove the second main result of this section, which gives a full characterization of the points of continuity of ECE_π.

Theorem 3.10. *Let* $1 \le p \le \infty$, *let* $g \in L^p(\Omega_X; \pi_X)$, *and set* $g^*(x) = \mathbb{E}[Y \mid X = x]$ *for* $x \in \Omega_X$. *Then* ECE_π *is continuous at* g *in the topology of* L^p *if and only if*

$$\mathrm{ECE}_\pi(g) = \|g^* - g\|_{L^1(\Omega_X; \pi_X)}. \tag{3.3}$$

Proof. We first show that if (3.3) holds, then ECE_π is continuous at g. Suppose that $\mathrm{ECE}_\pi(g) = \|g^* - g\|_{L^1(\Omega_X;\pi_X)}$, and let $g_1, g_2, \ldots \in L^p(\Omega_X; \pi_X)$ be any sequence converging to g. By Theorem 3.8, $\liminf_{n\to\infty} \mathrm{ECE}_\pi(g_n) \ge \mathrm{ECE}_\pi(g)$. Since $\sigma(g_n(X)) \subset \sigma(X)$, by Lemma 3.6 we have

$$\begin{aligned}
\limsup_{n\to\infty} \mathrm{ECE}_\pi(g_n) &= \limsup_{n\to\infty} \big\|\mathbb{E}[Y - g_n|g_n(X)]\big\|_1 \\
&\le \limsup_{n\to\infty} \big\|\mathbb{E}[Y - g_n|X]\big\|_1 \\
&= \limsup_{n\to\infty} \big\|g^* - g_n\big\|_{L^1(\Omega_X;\pi_X)} \\
&= \big\|g^* - g\big\|_{L^1(\Omega_X;\pi_X)} = \mathrm{ECE}_\pi(g).
\end{aligned}$$

Therefore $\mathrm{ECE}_\pi(g_n) \to \mathrm{ECE}_\pi(g)$, which shows that ECE_π is continuous at g.

For the opposite direction of implication, we show that if (3.3) does not hold, then ECE_π is discontinuous at g. Suppose that $\mathrm{ECE}_\pi(g) \ne \|g^* - g\|_{L^1(\Omega_X;\pi_X)}$. By Lemma 3.9, we can choose a sequence $g_n \to g$ in $L^p(\Omega_X; \pi_X)$ such that each g_n is a bijection onto a standard Borel space. Since a measurable bijection between standard Borel spaces has a measurable inverse, this implies that g_n^{-1} is a measurable bijection from the image of g_n onto Ω_X, which in turn implies that $\sigma(g_n(X)) = \sigma(X)$. Thus

$$\mathbb{E}[Y \mid g_n(X) = g_n(x)] = \mathbb{E}[Y \mid X = x] = g^*(x),$$

so that

$$\mathrm{ECE}_\pi(g_n) = \|g^* - g_n\|_{L^1(\Omega_X;\pi_X)}.$$

The right-hand side above is a continuous functional of g_n in $L^p(\Omega_X; \pi_X)$, so that

$$\lim_{n\to\infty} \mathrm{ECE}_\pi(g_n) = \|g^* - g\|_{L^1(\Omega_X;\pi_X)} \ne \mathrm{ECE}_\pi(g),$$

implying that ECE_π is discontinuous at g. □

We leave it as an exercise to verify that if $\Omega_X = [n]$ with the discrete topology, then the condition in Theorem 3.10 is equivalent to the condition in Theorem 3.2.

4. Logit Smoothed Calibration

We proved in Corollary 3.4 above that for finitely supported distributions, the discontinuities of ECE have measure zero. Although this statement only holds as written for discrete data, it nonetheless provides helpful intuition that we can use to mitigate the discontinuities of ECE in a more general setting: namely, we expect that predictors at which ECE is discontinuous should be, in some sense, rare. Therefore, if we add some independent continuously distributed random noise ξ to $g(X)$ before taking the conditional expectation

in (2.1), we can hope that the resulting functional of g will be continuous. We show below that indeed this is the case.

However, in order to preserve the interpretation of ECE_π, we need to ensure that $g(X) + \xi \in [0, 1]$. To do so, we assume that g can be decomposed as $\rho \circ h$ where $h : \mathbb{R}^d \to \mathbb{R}$ and $\rho : \mathbb{R} \to [0, 1]$ is a strictly increasing function (e.g., ρ can be the sigmoid function), observing that this is virtually always the case in practice.[2] We can then add the noise ξ to $h(X)$ rather than $g(X)$, which allows us to define the *Logit-Smoothed ECE (LS-ECE)* as follows:

$$\text{LS-ECE}_{\pi,\xi}(h) \triangleq \mathbb{E}_{X,\xi}\big[\big|\mathbb{E}[Y \mid \rho(h(X) + \xi)] - \rho(h(X) + \xi)\big|\big]. \tag{4.1}$$

Henceforth, we will always use $h(X)$ to denote the *logit* associated with $g(X)$, and $\rho : \mathbb{R} \to [0, 1]$ to denote a strictly increasing function with inverse ρ^{-1} differentiable everywhere on $(0, 1)$.

Comparison to Błasiok & Nakkiran (2023). While the main proposal of smECE in Błasiok & Nakkiran (2023) does not exactly match our notion of LS-ECE (as it corresponds to smoothing residuals of the predictions), we note that their notion of $\widetilde{\text{smECE}}$ shares more similarity to what we propose. Namely, $\widetilde{\text{smECE}}$ corresponds to smoothing the predictor g and then projecting back to $[0, 1]$. However, smoothing the logit function h directly avoids issues related to thresholding/projecting and leads to a cleaner development of the theory in this section, allowing us to prove continuity, consistency, and convergence to the true ECE under reasonable assumptions.

4.1. Continuity of LS-ECE

We now verify that unlike ECE_π, $\text{LS-ECE}_{\pi,\xi}(h)$ is continuous as a function of the logit h in the topology of L^∞ for *any* choice of π so long as ξ has a density with respect to the Lebesgue measure and satisfies very basic regularity conditions. The crux of our argument relies on analyzing how perturbations to the joint distribution of $(Y, \rho(h(X) + \xi))$ behave with respect to total variation, and then "pulling back" to continuity in the space of logit functions h.

We begin by showing in the next two propositions that the smoothed logits $h(X) + \xi$ are continuous in total variation with respect to the L^p norm on h.

Proposition 4.1. *Let Z_n denote a sequence of random variables converging to a random variable Z in L^p for $p \in [1, \infty]$, and let ξ be an independent, real-valued random variable with density p_ξ that is continuous Lebesgue almost everywhere. Suppose Z_n, Z are X-measurable for a random variable X. Then $(X, Z_n + \xi) \to (X, Z + \xi)$ in total variation.*

Remark 4.2. An example of a density that is not continuous Lebesgue almost everywhere is an indicator on a fat Cantor set, which is discontinuous on a set of positive Lebesgue measure. By the Lebesgue differentiation theorem, any equivalent density must agree almost everywhere, and hence still be discontinuous on a set of positive measure.

Proposition 4.3. *Let ξ be as in Proposition 4.1 and let $(X, Y) \sim \pi$. Suppose that $h_n(X) \to h(X)$ in L^p for some $p \in [1, \infty]$. Then $(Y, \rho(h_n(X) + \xi)) \overset{\text{TV}}{\to} (Y, \rho(h(X) + \xi))$.*

The final ingredient necessary for our proof of continuity is the connection between convergence in total variation of joint distributions and convergence of the associated ECEs. We provide this via the following general result, which may be of independent interest for future analysis of ECE.

Lemma 4.4. *Suppose that $(Y, T_n) \to (Y, T)$ in total variation, where T_n, T are random variables taking values in $[0, 1]$. Define*

$$\Delta_n = \big|\mathbb{E}_{T_n}\left[|\mathbb{E}[Y \mid T_n = t] - t|\right] - \mathbb{E}_T\left[|\mathbb{E}[Y \mid T = t] - t|\right]\big|. \tag{4.2}$$

Then $\lim_{n\to\infty} \Delta_n = 0$.

We can now prove the main result of this section.

Theorem 4.5. *Let ξ satisfy the conditions of Proposition 4.1. Then $\text{LS-ECE}_{\pi,\xi}(h)$ is a continuous functional of h in the topology of L^∞.*

Proof of Theorem 4.5. Let h_n denote a sequence of functions converging to h in L^∞. Let $Z_n = h_n(X)$, $Z = h(X)$ and $T_n = \rho(Z_n + \xi)$, $T = \rho(Z + \xi)$. By Lemma 4.3, we have that $(Y, T_n) \to (Y, T)$ in total variation. By Lemma 4.4, $|\text{LS-ECE}_{\pi,\xi}(h_n) - \text{LS-ECE}_{\pi,\xi}(h)| = \Delta_n \to 0$ as $n \to \infty$. □

5. Estimation of LS-ECE

Having established the continuity of $\text{LS-ECE}_{\pi,\xi}$, we turn to its estimation in practice. Let $\{(x_1, y_1), ..., (x_n, y_n)\}$ denote n points sampled from the data distribution π, and let $\hat{\pi}$ denote the empirical measure of the pairs (x_i, y_i). Then we can naturally approximate $\text{LS-ECE}_{\pi,\xi}$ by $\text{LS-ECE}_{\hat{\pi},\xi}$, and then estimate $\text{LS-ECE}_{\hat{\pi},\xi}$ by estimating the outer expectation in the definition of $\text{LS-ECE}_{\hat{\pi},\xi}$ via sampling.

We first explicitly derive the form of $\mathbb{E}[Y \mid \rho(h(X) + \xi)]$ in the population case of $(X, Y) \sim \pi$, as this will make clear the form of $\text{LS-ECE}_{\hat{\pi},\xi}$. This requires the following elementary proposition.

[2] We only run into issues if g takes the values 0 or 1 exactly, since then ρ^{-1} can be $-\infty$ and ∞ respectively. This is easily avoided in practice by adding/subtracting a small tolerance value to the predicted probabilities, if necessary.

Proposition 5.1. *Let Z be an arbitrary real-valued random variable and let ξ be a real-valued random variable with density p_ξ. Then $Z+\xi$ has the following density with respect to the Lebesgue measure:*

$$p_{Z+\xi}(t) = \mathbb{E}_Z[p_\xi(t - Z)]. \tag{5.1}$$

For brevity we now let $T = \rho(h(X) + \xi)$ with $(X, Y) \sim \pi$. Then we have from Proposition 5.1 that $\mathbb{E}[Y \mid T = t] = p_{T,Y=1}(t)/p_T(t)$, where the densities $p_{T,Y=1}$ and p_T are:

$$p_{T,Y=1}(t) = \pi_Y(1)(\rho^{-1})'(t)\mathbb{E}\big[p_\xi(\rho^{-1}(t) - h(X)) \mid Y = 1\big], \tag{5.2}$$

$$p_T(t) = (\rho^{-1})'(t)\mathbb{E}\left[p_\xi(\rho^{-1}(t) - h(X))\right]. \tag{5.3}$$

Now if we let $\hat{T}$ be analogous to T except with $(X, Y) \sim \hat{\pi}$, then we can similarly obtain the expression $\mathbb{E}[Y \mid \hat{T} = t] = p_{\hat{T},Y=1}(t)/p_{\hat{T}}(t)$, with the densities:

$$p_{\hat{T},Y=1}(t) = \frac{1}{n}\sum_{i=1}^{n}(\rho^{-1})'(t)p_\xi(\rho^{-1}(t) - h(x_i))\mathbb{1}_{y_i=1}, \tag{5.4}$$

$$p_{\hat{T}}(t) = \frac{1}{n}\sum_{i=1}^{n}(\rho^{-1})'(t)p_\xi(\rho^{-1}(t) - h(x_i)). \tag{5.5}$$

From here, it is straightforward to estimate $\text{LS-ECE}_{\hat{\pi},\xi}(h)$ by averaging over samples from $\hat{T}$, so long as we take p_ξ to be easy to sample from. We also see that $\mathbb{E}[Y \mid \hat{T} = t]$ is just the Nadaraya–Watson kernel regression estimator (Nadaraya, 1964; Watson, 1964) evaluated at t using h and a kernel corresponding to the density of ξ, as we would expect. The form of $\mathbb{E}[Y \mid \hat{T} = t]$ also makes it trivial to implement estimation of $\text{LS-ECE}_{\hat{\pi},\xi}(h)$ in practice, as we illustrate in Figure 1.

```
def gaussian_kernel(x, sigma):
    return 1/(sigma * np.sqrt(2 * np.pi)) * torch.exp(-torch.square(x) / (2 * sigma ** 2))

def kernel_reg(logits, labels, ts, sigma):
    total = gaussian_kernel(ts - logits, sigma)
    return (total * labels).sum(dim=0) / total.sum(dim=0)

def logit_smoothed_ece(logits, labels, n_t, sigma):
    # Expects logits to be shape (n, 1) and labels to be shape (n, 1).
    emp_sample = torch.randint(len(logits), (n_t,))
    ts = logits[emp_sample].squeeze(dim=1) + sigma * torch.randn(n_t,)
    ests = kernel_reg(logits, labels, ts, sigma)
    return torch.abs((ests - torch.nn.functional.sigmoid(ts))).mean()
```

Figure 1. Implementation of $\text{LS-ECE}_{\hat{\pi},\xi}(h)$ in 10 lines of PyTorch (Paszke et al., 2019) using broadcast semantics.

We need only now prove that $\text{LS-ECE}_{\hat{\pi},\xi}(h)$ converges in probability to $\text{LS-ECE}_{\pi,\xi}(h)$, i.e. that estimating $\text{LS-ECE}_{\hat{\pi},\xi}$ allows us to consistently estimate $\text{LS-ECE}_{\pi,\xi}(h)$. We note that here $\text{LS-ECE}_{\pi,\xi}(h)$ is a non-random scalar quantity depending on the logit function h while $\text{LS-ECE}_{\hat{\pi},\xi}(h)$ is a random variable depending on the pairs $(x_i, y_i) \sim \pi$.

It turns out that the only additional stipulation on ξ necessary to achieve consistency is that ξ be of the form $\xi = \sigma R$ for a random variable R with bounded density (for example, $\xi \sim \mathcal{N}(0, \sigma^2)$). So long as ξ is of this type, we can prove that $p_{\hat{T}} \to p_T$ and $p_{\hat{T},Y=1} \to p_{T,Y=1}$ in L^1, which is all we need for consistency. This is encapsulated in the following lemma, which once again may be of independent interest.

Lemma 5.2. *If $\xi = \sigma R$ for a random variable R with bounded density, then:*

$$\mathbb{E}_\pi\left[\int_0^1 \left|p_{T,Y=1}(t) - p_{\hat{T},Y=1}(t)\right| dt\right] = O\left(\frac{1}{\sqrt{n\sigma}}\right), \tag{5.6}$$

$$\mathbb{E}_\pi\left[\int_0^1 \left|p_T(t) - p_{\hat{T}}(t)\right| dt\right] = O\left(\frac{1}{\sqrt{n\sigma}}\right). \tag{5.7}$$

With Lemma 5.2, we can prove our main estimation result.

Theorem 5.3. *If $\xi = \sigma R$ for a random variable R with bounded density, we have that*

$$\mathbb{E}_\pi\left[\left|\text{LS-ECE}_{\pi,\xi}(h) - \text{LS-ECE}_{\hat{\pi},\xi}(h)\right|\right] = O\big(1/\sqrt{n\sigma}\big), \tag{5.8}$$

which implies $\text{LS-ECE}_{\hat{\pi},\xi}(h) \to \text{LS-ECE}_{\pi,\xi}(h)$ in probability.

Proof. Let T be as in Lemma 5.2, i.e. the population form of $\hat{T}$. Then we have by triangle inequality:

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left|\text{LS-ECE}_{\pi,\xi}(h) - \text{LS-ECE}_{\hat{\pi},\xi}(h)\right|\right] \\
&= \mathbb{E}_\pi\left[\left|\int_0^1 \left|\mathbb{E}[Y \mid T = t] - t\right| p_T(t)\, dt\right.\right. \\
&\qquad \left.\left. - \int_0^1 \left|\mathbb{E}[Y \mid \hat{T} = t] - t\right| p_{\hat{T}}(t)\, dt\right|\right] \\
&\le \mathbb{E}_\pi\left[\int_0^1 \left|p_{T,Y=1}(t) - p_{\hat{T},Y=1}(t)\right| dt\right. \\
&\qquad \left. + \int_0^1 t\left|p_T(t) - p_{\hat{T}}(t)\right| dt\right]
\end{aligned} \tag{5.9}$$

Noting that $t \le 1$ and applying Lemma 5.2 shows that Equation (5.9) is $O(1/\sqrt{n\sigma})$, which is the desired result. □

The quantitative bound in Theorem 5.3 shows that choosing σ in practice is — as intuition would suggest from the discussion of kernel regression — similar to choosing the kernel bandwidth. In general, σ can be thought of as a hyperparameter, but we will see in Section 6 that experiments are relatively insensitive to the choice of σ.

5.1. Consequences for Estimation of ECE

Nevertheless, for theoretical purposes, the scaling of σ is important. Theorem 5.3 suggests that σ should be at least $\omega(1/n)$ to prevent the estimation error from exploding.

It is then natural to ask what happens if we take $\sigma = \omega(1/n)$ but $\sigma \to 0$ as $n \to \infty$. By doing so — analogously, once again, to existing results in the kernel density estimation literature — we obtain that $\text{LS-ECE}_{\hat{\pi},\xi_n}(h)$ actually becomes a consistent estimator of the true ECE, under appropriate conditions on the logit distribution $h(X)$. This is a non-trivial estimation result that we effectively get for free as a result of our framework, as shown in the short proof below.

Theorem 5.4. *Suppose $\xi = \sigma R$ for a random variable R with bounded, almost-everywhere continuous density, and let $\xi_n = \sigma_n \xi$ with σ_n satisfying $\lim_{n\to\infty} \sigma_n = 0$ and $\sigma_n = \omega(1/n)$. If the distribution of $h(X)$ conditioned on $Y = y$ has an almost-everywhere continuous density for $y \in \{0, 1\}$, then $\text{LS-ECE}_{\hat{\pi},\xi_n}(h) \to \text{ECE}_{\pi}(\rho \circ h)$ in probability.*

Proof. First we claim that if Z has an a.e. continuous density p_Z, then $Z + \xi_n \to Z$ in total variation. This follows by noting that $p_{\sigma\xi}(x) = \frac{1}{\sigma} p_\xi\left(\frac{x}{\sigma}\right)$ is a sequence of good kernels (i.e. an approximation to the identity) (Stein & Shakarchi, 2011), so that

$$p_{Z+\xi_n}(x) = (p_Z * p_{\xi_n})(x) \to p_Z(x)$$

at any continuity point of p_Z. Then $Z + \xi_n \overset{\text{TV}}{\to} Z$ by Scheffé's Lemma.

Now for $y \in \{0, 1\}$, by assumption and the above claim, conditioned on $Y = y$, we have $h(X) + \xi_n \overset{\text{TV}}{\to} h(X)$. Hence $(Y, h(X) + \xi_n) \overset{\text{TV}}{\to} (Y, h(X))$.

By Lemma 4.4, $|\text{LS-ECE}_{\pi,\xi_n}(h) - \text{ECE}_{\pi}(\rho \circ h)| \to 0$ as $n \to \infty$. The result then follows from the triangle inequality and Theorem 5.3. □

Comparison to existing estimation results. To our knowledge, the only existing consistency results for ECE (or, more specifically, the L^1 ECE) are the works of Zhang et al. (2020) and Popordanoska et al. (2022) that show consistency via adapting corresponding results for kernel density estimation. These results thus require assumptions such as Hölder-smooth and bounded or Lipschitz continuity of the density[3] of $h(X)$ conditioned on $Y = y$, whereas we require only a.e. continuity.

[3]Technically these assumptions were stated in terms of the conditional density of the predictor value $g(X)$ given $Y = y$, but they imply similar constraints on h when considering $g = \rho \circ h$ with ρ being the sigmoid function.

6. Experiments

We now empirically verify that $\text{LS-ECE}_{\pi,\xi}$ behaves nicely even when ECE_{π} does not. We revisit the simple 2-point data distribution of Definition 3.1 in Section 6.1, and show that the discontinuity at $g(x) = 1/2$ leads to oscillatory behavior in $\text{ECE}_{\text{BIN},\pi}$ (defined above in (2.3)) as we change the number of bins, whereas $\text{LS-ECE}_{\hat{\pi},\xi}$ remains effectively constant irrespective of the choice of variance for ξ. On the other hand, we also show that for image classification using a wide range of models, $\text{ECE}_{\text{BIN},\pi}$ changes smoothly as we vary the number of bins, and the resulting estimates match up closely with both $\text{LS-ECE}_{\hat{\pi},\xi}$ as well as the smECE of Błasiok & Nakkiran (2023). For all experiments in this section, we take $\xi \sim \mathcal{N}(0, \sigma^2)$. We consider using uniform noise in Appendix D; the choice of ξ does not impact our conclusions. All of the code used to generate the plots in this section can be found at: `https://github.com/2014mchidamb/how-flawed-is-ece`.

6.1. Synthetic Data

We consider data drawn from a distribution π as in Definition 3.1, and the predictor $g(x) = \rho(\alpha x)$ where ρ is the sigmoid function and $\alpha = 10^{-3}$. We construct g in this way so that $g(-0.5) = 1/2 - \varepsilon$ and $g(0.5) = 1/2 + \varepsilon$, i.e. g matches our discussion in Section 3.1.

As we already know, ECE_{π} is discontinuous at the predictor which always predicts $1/2$. This immediately presents a problem for the estimation of $\text{ECE}_{\pi}(g)$ via $\text{ECE}_{\text{BIN},\pi}(g)$; indeed, one can readily compute (as done in (Błasiok et al., 2023)) that $\text{ECE}_{\text{BIN},\pi}(g)$ jumps between ≈ 0 and $\approx 1/2$ depending on the parity of the number of bins used. We visualize this behavior in Figure 2, where we plot $\text{ECE}_{\text{BIN},\pi}(g)$ evaluated on 1000 samples from π with the number of bins ranging from 1 to 100.

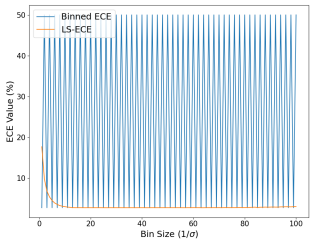


Figure 2. Comparison of $\text{ECE}_{\text{BIN},\pi}$ (blue) and $\text{LS-ECE}_{\pi,\xi}$ (orange) over bins (and correspondingly, inverse scalings for ξ) ranging from 1 to 100 on the model and data setup of Section 6.1.

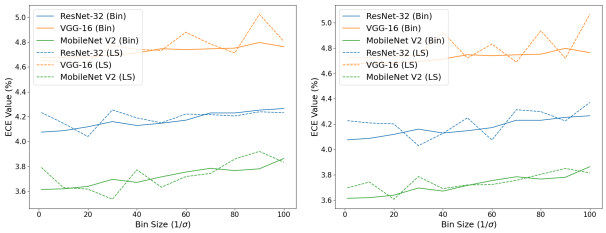
Alongside $\text{ECE}_{\text{BIN},\pi}(g)$, we also plot $\text{LS-ECE}_{\hat{\pi},\xi}(h)$, where $h(x) = \alpha x$ and $\xi \sim \mathcal{N}(0, \sigma^2)$. We let σ be the inverse of the number of bins used for $\text{ECE}_{\text{BIN},\pi}(g)$. The motivation for this choice of σ comes from considering a uniform kernel (i.e. $\xi \sim \sigma\text{Uni}([-1/2, 1/2])$), since in this

case σ corresponds to the bin size centered at each point. We estimate LS-ECE$_{\hat{\pi},\xi}(h)$ via 10000 independent samples drawn from the distribution of $h(X)+\xi$, as discussed in Section 5. We see that, outside the case of large σ (i.e. $\sigma \gtrsim 1$), LS-ECE$_{\hat{\pi},\xi}(h)$ remains effectively constant near zero and entirely avoids the oscillatory behavior of ECE$_{\text{BIN},\pi}$.

6.2. Image Classification

More importantly, we now check whether this disparity between ECE$_{\text{BIN},\pi}(g)$ and LS-ECE$_{\hat{\pi},\xi}(h)$ appears in settings of practical interest. We consider CIFAR-10, CIFAR-100 (Krizhevsky, 2009), and ImageNet (Deng et al., 2009) and compare ECE$_{\text{BIN},\pi}(g)$ using the bin numbers $\{1, 10, 20, ..., 100\}$ to both LS-ECE$_{\hat{\pi},\xi}(h)$ with inversely proportional σ and smECE. We point out that smECE uses a particular choice of kernel bandwidth which we do not vary, so the smECE results are constant with respect to σ for each model. As before, we estimate LS-ECE$_{\hat{\pi},\xi}(h)$ using 10000 independent samples drawn from the distribution of $h(X) + \xi$.

Since all of the experiments in this section deal with multi-class classification, we use the top-class (or confidence calibration) formulations of ECE, LS-ECE, and smECE (see (2.2)). For LS-ECE, we construct the logit function $h(x)$ as $h(x) = \rho^{-1}(\max_i g^i(x))$, i.e. we apply the inverse sigmoid function to the maximum predicted probability.



(a) CIFAR-10 (b) CIFAR-100

Figure 3. Comparison of ECE$_{\text{BIN},\pi}$ and LS-ECE$_{\pi,\xi}$ for different models on CIFAR datasets over bins/variance scalings ranging from 1 to 100. Solid lines correspond to ECE$_{\text{BIN},\pi}$ and dashed lines correspond to LS-ECE$_{\pi,\xi}$.

6.2.1. CIFAR Experiments

For our CIFAR experiments, we use pretrained versions (due to Yaofo Chen) of ResNet-32 (He et al., 2015), VGG-16 (Simonyan & Zisserman, 2015), and MobileNetV2 (Sandler et al., 2019) available on TorchHub. Results for evaluating these models on the CIFAR-10 and CIFAR-100 test data are shown in Figure 3.

As can be seen, ECE$_{\text{BIN},\pi}(g)$ stays nearly the same as we change the number of bins, and LS-ECE$_{\pi,\xi}(h)$ tracks it quite closely. (Although LS-ECE$_{\pi,\xi}(h)$ visually appears to exhibit more variance, the scale of this variance is small.) Furthermore, we note that the conclusions drawn from both ECE and LS-ECE about which model is best calibrated stay consistent across the choice of bin number/variance scaling.

6.2.2. ImageNet Experiments

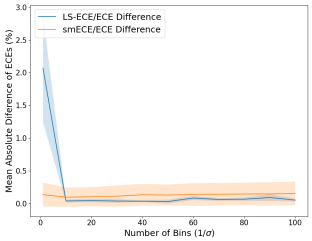


Figure 4. Mean absolute difference between ECE and LS-ECE, as well as ECE and smECE, on ImageNet-1K-val over all models considered in Section 6.2.2, with one standard deviation error bounds marked using the shaded region.

We repeat our CIFAR experimental setup for ImageNet, and consider a gamut of pretrained architectures: ResNet-18 and ResNet-50 (He et al., 2015; Wightman et al., 2021), EfficientNet (Tan & Le, 2019), MobileNetV3 (Howard et al., 2019), Vision Transformer (Dosovitskiy et al., 2021), and RegNet (Radosavovic et al., 2020). All of our models are obtained from the `timm` (Wightman, 2019) library and were pretrained using the techniques described by Wightman et al. (2021). We evaluate all models on the ImageNet-1K validation data, and in Figure 4 we report the mean absolute difference (over all models) between ECE and LS-ECE, as well as ECE and smECE, across bin numbers and choices of σ respectively.

The ImageNet results further corroborate our CIFAR findings: ECE, LS-ECE, and smECE take near-identical values for all models considered, across the range of possible bin numbers and variances. Although this is by no means a comprehensive evaluation, the fact that the *continuous* LS-ECE so closely tracks ECE in these experiments suggests that the theoretical pathologies of ECE may not pose a problem for assessing the calibration of real-world models in practice.

7. Conclusion

In summary, we have entirely characterized the discontinuities of ECE in a very general setting. We further used these continuity results to motivate the construction of LS-ECE, a continuous analogue of ECE that tracks it closely, and which in fact can be used to obtain a consistent estimator of ECE. As the results in this work are largely theoretical, a natural direction for future work would be a large-scale empirical validation of ECE results in the literature.

Acknowledgments

The work of M.C. is supported (via Rong Ge) by the National Science Foundation under grant numbers DMS-2031849 and CCF-1845171 (CAREER). The work of C.M. is supported by the National Science Foundation under grant number DMS-2103170 and by a grant from the Simons Foundation. The work of S.R. is supported by the National Science Foundation under grant number DMS-2202959. Part of this work was conducted while the authors were attending the Random Theory 2023 workshop.

Impact Statement

This paper presents work whose goal is to advance the understanding of a popular metric used for calibration. As our work is largely theoretical, we do not anticipate any negative uses or harmful societal impacts.

References

Arrieta-Ibarra, I., Gujral, P., Tannen, J., Tygert, M., and Xu, C. Metrics of calibration for probabilistic predictions. *arXiv preprint arXiv:2205.09680*, 2022.

Bröcker, J. Some remarks on the reliability of categorical probability forecasts. *Monthly weather review*, 136(11): 4488–4502, 2008.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ..., and Amodei, D. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Błasiok, J. and Nakkiran, P. Smooth ece: Principled reliability diagrams via kernel smoothing, 2023.

Błasiok, J., Gopalan, P., Hu, L., and Nakkiran, P. A unifying theory of distance from calibration, 2023.

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

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dimitriadis, T., Gneiting, T., and Jordan, A. I. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8):e2016191118, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880): 348–352, 2021.

Foster, D. P. and Hart, S. Smooth calibration, leaky forecasts, finite recall, and Nash dynamics. *Games Econ. Behav.*, 109:271–293, 2018. URL https://doi.org/10.1016/j.geb.2017.12.022.

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

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (2):243–268, 2007.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, Q., Dai, J., Qiao, Y., and Li, H. Planning-oriented autonomous driving, 2023.

Kakade, S. and Foster, D. Deterministic calibration and Nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

Krizhevsky, A. Learning multiple layers of features from tiny images. pp. 32–33, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Kudô, H. A note on the strong convergence of σ-algebras. *Annals of Probability*, 2:76–83, 1974.

Kull, M. and Flach, P. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 68–85. Springer, 2015.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Kumar, A., Liang, P. S., and Ma, T. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pp. 3792–3803, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Lee, D., Huang, X., Hassani, H., and Dobriban, E. T-Cal: An optimal test for the calibration of predictive models. *arXiv preprint arXiv:2203.01850*, 2022.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Müller, R., Kornblith, S., and Hinton, G. When does label smoothing help?, 2020.

Nadaraya, E. A. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020. URL https://doi.org/10.1137/1109020.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. Binary classifier calibration: Non-parametric approach. *arXiv preprint arXiv:1401.3390*, 2014.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *CVPR Workshops*, volume 2, 2019.

Nogales, A., Álvaro J. García-Tejedor, Monge, D., Vara, J. S., and Antón, C. A survey of deep learning models in medical therapeutic areas. *Artificial Intelligence in Medicine*, 112:102020, 2021. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2021.102020. URL https://www.sciencedirect.com/science/article/pii/S0933365721000130.

OpenAI. Gpt-4 technical report, 2023.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. URL http://arxiv.org/abs/1912.01703.

Popordanoska, T., Sayer, R., and Blaschko, M. A consistent and differentiable L_p canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35: 7933–7946, 2022.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Doll'ar, P. Designing network design spaces. In *CVPR*, 2020.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Roelofs, R., Cain, N., Shlens, J., and Mozer, M. C. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4036–4054. PMLR, 2022.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. MobileNetV2: Inverted residuals and linear bottlenecks, 2019.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015.

Stein, E. M. and Shakarchi, R. *Fourier Analysis: An Introduction*, volume 1. Princeton University Press, 2011.

Tan, M. and Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32:13888–13899, 2019.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and efficient foundation language models, 2023.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3459–3467. PMLR, 2019.

Wang, C. and Golebiowski, J. Meta-calibration regularized neural networks, 2023.

Wang, D.-B., Feng, L., and Zhang, M.-L. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 11809–11820. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/61f3a6dbc9120ea78ef75544826c814e-Paper.pdf.

Watson, G. S. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 26(4): 359–372, 1964. ISSN 0581572X. URL http://www.jstor.org/stable/25049340.

Wightman, R. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wightman, R., Touvron, H., and Jégou, H. ResNet strikes back: An improved training procedure in timm. *CoRR*, abs/2110.00476, 2021. URL https://arxiv.org/abs/2110.00476.

Zhang, J., Kailkhura, B., and Han, T. Y.-J. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*, pp. 11117–11128. PMLR, 2020.

A. Proofs for Section 3

A.1. Proofs in the Discrete Setting

Theorem 3.2. *[Discontinuities for Discrete ECE] Let π be any distribution such that $\mathrm{supp}(\pi_X) = [n]$ for an arbitrary positive integer n, and let $g^*(x) = P(Y = 1 \mid X = x)$ denote the ground truth conditional distribution. Then the set of discontinuities of ECE_π (in the space of predictors $g : [n] \to [0, 1]$ endowed with the ℓ^∞ norm) is exactly the set of g such that there exists $m \in [n]$ with $P(X = m) \neq 0$ and*

$$\big|g^*(m) - g(m)\big| \neq \big|\mathbb{E}[Y \mid g(X) = g(m)] - g(m)\big|. \tag{3.1}$$

Proof of Theorem 3.2. Write $p^* : [n] \to [0, 1]$ for the probability mass function of π_X. Then we have:

$$\mathrm{ECE}_\pi(g) = \sum_{i=1}^{n} p_i^* \left| \frac{\sum_{j \in S(g,g(i))} p_j^* g^*(j)}{\sum_{j \in S(g,g(i))} p_j^*} - g(i) \right|. \tag{A.1}$$

Suppose now that there exists an $m \in [n]$ such that $P(X = m) \neq 0$ and (3.1) holds. Consider a new predictor $\tilde{g}$ such that $\tilde{g}(i) = g(i)$ for $i \neq m$, and $\tilde{g}(m) = g(m) + \delta$ for $|\delta|$ small enough that $\tilde{g}(m) \in [0, 1]$ and $|S(\tilde{g}, \tilde{g}(m))| = 1$. Then it follows that

$$\mathrm{ECE}_\pi(\tilde{g}) = \mathrm{ECE}_\pi(g) - p_m^* \left| \frac{\sum_{j \in S(g,g(m))} p_j^* g^*(j)}{\sum_{j \in S(g,g(m))} p_j^*} - g(m) \right| + p_m^* |g^*(m) - \tilde{g}(m)|, \tag{A.2}$$

which implies:

$$\begin{aligned} |\mathrm{ECE}_\pi(g) - \mathrm{ECE}_\pi(\tilde{g})| &= p_m^* \left| \left| \frac{\sum_{j \in S(g,g(m))} p_j^* g^*(j)}{\sum_{j \in S(g,g(m))} p_j^*} - g(m) \right| - |g^*(m) - \tilde{g}(m)| \right| \\ &= p_m^* \Big| \big|\mathbb{E}[Y \mid g(X) = g(m)] - g(m)\big| - |g^*(m) - \tilde{g}(m)| \Big|. \end{aligned} \tag{A.3}$$

Thus we have $\lim_{\delta \to 0} \|g - \tilde{g}\|_\infty = 0$, whereas

$$\lim_{\delta \to 0} |\mathrm{ECE}_\pi(g) - \mathrm{ECE}_\pi(\tilde{g})| = p_m^* \Big| \big|\mathbb{E}[Y \mid g(X) = g(m)] - g(m)\big| - |g^*(m) - g(m)| \Big|,$$

which is positive by (3.1). Therefore ECE_π is discontinuous at g.

For the other direction, we show that if (3.1) does not hold for any $m \in [n]$ with $P(X = m) \neq 0$, then ECE_π is continuous at g. For any such g and π, we have:

$$\mathrm{ECE}_\pi(g) = \sum_{i=1}^{n} p_i^* |g^*(i) - g(i)|. \tag{A.4}$$

Now set $\delta = \min |g(i) - g(j)|/2$, where the minimum runs over pairs of distinct $i \neq j \in [n]$ with $P(X = i), P(x = j) > 0$. By Lemma 3.3 we must have $|S(g, g(i))| = |S(g, g(j))| = 1$ for any such i, j, so that $\delta > 0$. Then any $\tilde{g}$ satisfying $\|\tilde{g} - g\|_\infty < \delta$ must also satisfy the property $\tilde{g}(i) \neq \tilde{g}(j)$ for any $i \neq j \in [n]$ with $P(X = i), P(x = j) > 0$. Therefore, again by Lemma 3.3, (3.1) cannot hold with $\tilde{g}$ in the place of g for any $m \in [n]$ with $P(X = m) \neq 0$, so that $\mathrm{ECE}_\pi(\tilde{g})$ has the same form as A.4. We thus find

$$|\mathrm{ECE}_\pi(g) - \mathrm{ECE}_\pi(\tilde{g})| = \sum_{i=1}^{n} p_i^* \Big| |g^*(i) - g(i)| - |g^*(i) - \tilde{g}(i)| \Big| \leq \|\tilde{g} - g\|_\infty,$$

which shows that ECE_π is continuous at g. □

A.2. Proofs in the General Setting

Proposition 3.5. *Take π with $\operatorname{supp}(\pi_X) = [0,1]^2$ such that $\pi_{X_1} = \operatorname{Uni}([0,1])$,*

$$\pi_{X_2|X_1=x_1} = x_1 \operatorname{Uni}([0.5,1]) + (1-x_1)\operatorname{Uni}([0,0.5)),$$

and $\mathbb{P}(Y=1 \mid X_2 = x_2) = \mathbb{1}_{x_2 \geq 0.5}$. Then ECE_π is discontinuous at the predictor $g(x) = x_1$, despite the fact that g has no level sets of positive measure.

Proof. First, observe that the second assumption on π implies

$$\begin{aligned}\mathbb{P}(Y=1 \mid X_1 = x_1) &= \int_0^1 \mathbb{P}(Y=1 \mid X_2 = x_2, X_1 = x_1)\mathbb{P}(X_2 = x_2 \mid X_1 = x_1)\, dx_2 \\ &= \int_{0.5}^1 2x_1 \, dx_2 = x_1, \end{aligned} \tag{A.5}$$

from which it follows that $\mathbb{E}[Y \mid g(X)] = g(X)$ and therefore $\mathrm{ECE}_\pi(g) = 0$. Now the idea is to apply an L^∞ perturbation of size δ to g such that we can separate the points for which x_1 is the same but $x_2 \geq 0.5$ or $x_2 < 0.5$, and in doing so obtain that the conditional probability (given x_1) of the label being 1 is either 0 or 1.

Taking $\delta = 1/n$ for $n \in \mathbb{N}$ sufficiently large, we define the following two functions $g_{\delta,0}$ and $g_{\delta,1}$:

$$g_{\delta,0}(z) = \left(\left\lfloor \frac{z}{\delta} \right\rfloor + 1 - \frac{\delta}{4} \right) \delta, \tag{A.6}$$

$$g_{\delta,1}(z) = \left(\left\lfloor \frac{z}{\delta} \right\rfloor + \frac{\delta}{4} \right) \delta. \tag{A.7}$$

Clearly we have that $g_{\delta,0}(z) \neq g_{\delta,1}(z)$ for $z \in [0,1]$ and that both $|g_{\delta,0}(z) - z| < \delta$ and $|g_{\delta,1}(z) - z| < \delta$. Additionally, $1 \geq g_{\delta,0} > g_{\delta,1}$ for $z \in [0,1)$. Now we can define a perturbation of $g(x) = x_1$ by:

$$g_\delta(x) = \begin{cases} g_{\delta,0}(x_1) & \text{if } x_2 < 0.5 \\ g_{\delta,1}(x_1) & \text{if } x_2 \geq 0.5. \end{cases} \tag{A.8}$$

We can then compute $\mathrm{ECE}_\pi(g_\delta)$ as follows:

$$\begin{aligned}\mathrm{ECE}_\pi(g_\delta) &= \int |\mathbb{E}[Y \mid g_\delta(x)] - g_\delta(x)| \, d\pi_X(x) \\ &= \int |\mathbb{E}[Y \mid g_\delta(x)] - g_\delta(x)| \mathbb{1}_{x_2<0.5} \, d\pi_X(x) + \int |\mathbb{E}[Y \mid g_\delta(x)] - g_\delta(x)| \mathbb{1}_{x_2 \geq 0.5} \, d\pi_X(x) \\ &\geq \int x_1 \mathbb{1}_{x_2<0.5} \, d\pi_X(x) + \int (1-x_1)\mathbb{1}_{x_2 \geq 0.5} \, d\pi_X(x) - 2\delta \\ &= \int_0^1 \int_0^{0.5} 2x_1(1-x_1) \, dx_2 \, dx_1 + \int_0^1 \int_{0.5}^1 2x_1(1-x_1) \, dx_2 \, dx_1 \\ &= \frac{1}{3} - 2\delta. \end{aligned} \tag{A.9}$$

Therefore $\mathrm{ECE}_\pi(g_\delta) \not\to 0$ as $\delta \to 0$, and thus ECE_π is discontinuous at g. □

Lemma 3.6. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, and let $\mathcal{F} \subseteq \mathcal{G} \subseteq \Sigma$ be sub-σ-algebras. Then for any $f \in L^1(\Omega; \mathbb{P})$, $\|\mathbb{E}[f|\mathcal{F}]\|_1 \leq \|\mathbb{E}[f|\mathcal{G}]\|_1$.*

Proof. By the conditional Jensen's inequality and the tower property of conditional expectation,

$$\big\|\mathbb{E}[f|\mathcal{F}]\big\|_1 = \int_\Omega \big|\mathbb{E}[f|\mathcal{F}]\big| \, d\mathbb{P} = \int_\Omega \big|\mathbb{E}[\,\mathbb{E}[f|\mathcal{G}] \,|\, \mathcal{F}]\big| \, d\mathbb{P} \leq \int_\Omega \mathbb{E}\big[\,|\mathbb{E}[f|\mathcal{G}]| \,\big|\, \mathcal{F}\big] \, d\mathbb{P} = \big\|\mathbb{E}[f|\mathcal{G}]\big\|_1. \qquad \square$$

Lemma 3.7. *Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, let $f \in L^1(\Omega; \mathbb{P})$, and let $g, g_1, g_2, \ldots$ be real-valued random variables such that $g_n \to g$ in probability. Then* $\liminf_{n\to\infty} ||\mathbb{E}[f|g_n]||_1 \geq ||\mathbb{E}[f|g]||_1$.

Proof of Lemma 3.7. We write $\sigma(g)$ for the σ-algebra generated by preimages of Borel sets under g. The *lower limit* of the sequence of σ-algebras $\sigma(g_1), \sigma(g_2), \ldots$, defined by (Kudô, 1974), is the σ-algebra

$$\mathcal{G} = \Big\{ A \in \Sigma \;\Big|\; \lim_{n\to\infty} \inf_{B \in \sigma(g_n)} \mathbb{P}(A \Delta B) = 0 \Big\},$$

where Δ indicates the symmetric difference of sets. The lower limit satisfies the property that

$$\liminf_{n\to\infty} \int_\Omega \big|\mathbb{E}[h|g_n]\big| \, d\mathbb{P} \geq \int_\Omega \big|\mathbb{E}[h|\mathcal{G}]\big| \, d\mathbb{P} \tag{A.10}$$

for any bounded Σ-measurable function h, and if $\mathcal{F}$ is any σ-algebra such that (A.10) holds with $\mathcal{F}$ in the place of $\mathcal{G}$, then $\mathcal{F} \subset \mathcal{G}$.

By (A.10) and Lemma 3.6, to prove the desired result it suffices to show that $\sigma(g) \subset \mathcal{G}$, and thus it suffices to show that some generating set of $\sigma(g)$ is contained in $\mathcal{G}$. Let C be the set of atoms of the pushforward distribution $g_*\mathbb{P}$ of g on $\mathbb{R}$. Then C is at most countable, and thus sets of the form

$$A = \{\omega \in \Omega \mid g(\omega) < x\}, \tag{A.11}$$

for $x \in \mathbb{R} \setminus C$, generate $\sigma(g)$. We will show that $\lim_{n\to\infty} \inf_{B\in\sigma(g_n)} \mathbb{P}(A \Delta B) = 0$, so that $A \in \mathcal{G}$.

Fix a set A of the form (A.11) and $\varepsilon > 0$, and let $B_{n,\varepsilon} = \{\omega \in \Omega \mid g_n(\omega) < x + \varepsilon\}$. We then have

$$\mathbb{P}(|g_n - g| > \varepsilon) \geq \mathbb{P}(A \Delta B_{n,\varepsilon}) - \mathbb{P}(x \leq g \leq x + 2\varepsilon),$$

so that

$$\inf_{B\in\sigma(g_n)} \mathbb{P}(A \Delta B) \leq \mathbb{P}(|g_n - g| > \varepsilon) + \mathbb{P}(x \leq g \leq x + 2\varepsilon),$$

and since $g_n \to g$ in probability, we obtain

$$\lim_{n\to\infty} \inf_{B\in\sigma(g_n)} \mathbb{P}(A \Delta B) \leq \mathbb{P}(x \leq g \leq x + 2\varepsilon).$$

Since x is not an atom of $g_*\mathbb{P}$, the right-hand side above can be made arbitrarily small. Therefore $\lim_{n\to\infty} \inf_{B\in\sigma(g_n)} \mathbb{P}(A \Delta B) = 0$, which completes the proof. □

Lemma 3.9. *Let $(\Omega, \mathcal{B}(\Omega), \mathbb{P})$ be a probability space, where Ω is a Polish space and $\mathcal{B}(\Omega)$ is its Borel σ-algebra. For $1 \leq p \leq \infty$ define*

$$L^p_{\text{inj}}(\Omega; \mathbb{P}) = \big\{ f \in L^p(\Omega; \mathbb{P}) \;\big|\; f \text{ is almost surely equal to a bijection onto a standard Borel space} \big\}.$$

Then $L^p_{\text{inj}}(\Omega; \mathbb{P})$ is dense in $L^p(\Omega; \mathbb{P})$.

Proof of Lemma 3.9. Although the lemma holds for the full L^p space of complex-valued functions, it is sufficient to show the result for the subspace of real-valued functions in L^p, and we will only need to use this case below. Moreover, by the Kuratowski isomorphism theorem, it is sufficient to consider the cases $\Omega = \mathbb{R}$, $\mathbb{Z}$, or $[n]$. We take $\Omega = \mathbb{R}$; the cases $\Omega = \mathbb{Z}$ or $[n]$ can be treated by a simpler version of the same argument.

Given a real-valued $f \in L^p(\Omega; \mathbb{P})$, we can choose simple functions

$$f_n = \sum_{j=1}^{n} a_j^{(n)} \mathbb{1}_{A_j^{(n)}},$$

where $A_1^{(n)}, \ldots, A_n^{(n)} \subset \Omega$ are pairwise disjoint open or half-open intervals and $a_n^{(n)} \geq \ldots \geq a_1^{(n)} \in \mathbb{R}$, such that $f_n \to f$ in L^p. Again by the Kuratowski isomorphism theorem, there exist measurable bijections $\varphi_j^{(n)} : A_j^{(n)} \to (0,1)$. For $\varepsilon > 0$, define

$$f_{n,\varepsilon} = \sum_{j=1}^{n} \left[a_j^{(n)} + \varepsilon(\varphi_j^{(n)} + 2j - 1)\right] \mathbb{1}_{A_j^{(n)}}.$$

Observe that $f_{n,\varepsilon}$ maps each set $A_j^{(n)}$ bijectively to the interval $(a_j^{(n)} + (2j-1)\varepsilon,\ a_j^{(n)} + 2j\varepsilon)$, and for ε sufficiently small, these latter intervals have pairwise disjoint closures for different values of j. Therefore, for all small enough $\varepsilon_n > 0$, f_{n,ε_n} is injective, and thus a measurable bijection onto its image. This image is a union of finitely many open intervals with pairwise disjoint closures, which is a Polish space, and thus f_{n,ε_n} is an isomorphism from Ω to a standard Borel space. Moreover $\|f_n - f_{n,\varepsilon}\|_p \leq 2n\varepsilon$, so that additionally taking ε_n small enough that $n\varepsilon_n \to 0$, we have that f_{n,ε_n} is a sequence in $L^p_{\mathrm{inj}}(\Omega; \mathbb{P})$ that converges to f in L^p. □

B. Proofs for Section 4

Proposition 4.1. *Let Z_n denote a sequence of random variables converging to a random variable Z in L^p for $p \in [1, \infty]$, and let ξ be an independent, real-valued random variable with density p_ξ that is continuous Lebesgue almost everywhere. Suppose Z_n, Z are X-measurable for a random variable X. Then $(X, Z_n + \xi) \to (X, Z + \xi)$ in total variation.*

Proof. By assumption, $p_\xi(x - \epsilon) \to p_\xi(x)$ as $\epsilon \to 0$, almost everywhere. By Scheffé's Lemma, $d_{\mathrm{TV}}(\xi, \xi + \epsilon) \to 0$ as $\epsilon \to 0$. Hence, for all $\epsilon > 0$, there exists $\delta > 0$ such that $d_{\mathrm{TV}}(\xi, \xi + \delta') < \epsilon$ for all $|\delta'| < \delta$.

Note we may assume $p = 1$. Choosing δ based on ϵ, we have

$$\begin{aligned}
d_{\mathrm{TV}}((X, Z_n + \xi), (X, Z + \xi)) &\leq \iint d_{\mathrm{TV}}(z_n + \xi, z + \xi)\, d\mathbb{P}_{Z_n, Z|X=x}(z_n, z)\, d\mathbb{P}_X(x) \\
&\leq \iint d_{\mathrm{TV}}(\xi, z - z_n + \xi)\, d\mathbb{P}_{Z_n, Z|X=x}(z_n, z)\, d\mathbb{P}_X(x) \\
&\leq \iint_{|z_n - z| < \delta} d_{\mathrm{TV}}(\xi, z - z_n + \xi)\, d\mathbb{P}_{Z_n, Z|X=x}(z_n, z)\, d\mathbb{P}_X(x) + \mathbb{P}(|Z_n - Z| \geq \delta) \\
&\leq \epsilon + \mathbb{P}(|Z_n - Z| \geq \delta) \\
&\leq \epsilon + \frac{\mathbb{E}|Z_n - Z|}{\delta}
\end{aligned} \tag{B.1}$$

where the last step follows from Markov's inequality. Choose N such that for $n \geq N$, $\mathbb{E}|Z_n - Z| \leq \delta\epsilon$. Then for $n \geq 2N$, we have $d_{\mathrm{TV}}((X, Z_n + \xi), (X, Z + \xi)) \leq 2\epsilon$. □

Proposition 4.3. *Let ξ be as in Proposition 4.1 and let $(X, Y) \sim \pi$. Suppose that $h_n(X) \to h(X)$ in L^p for some $p \in [1, \infty]$. Then $(Y, \rho(h_n(X) + \xi)) \overset{\mathrm{TV}}{\to} (Y, \rho(h(X) + \xi))$.*

Proof. Let $Z_n = h_n(X)$ and $Z = h(X)$. By Proposition 4.1, $(X, Z_n + \xi) \overset{\mathrm{TV}}{\to} (X, Z + \xi)$. By the data processing inequality, applying the kernel $\pi_{Y|X}$ to the first argument and ρ to the second argument, $(Y, \rho(Z_n + \xi)) \overset{\mathrm{TV}}{\to} (Y, \rho(Z + \xi))$. □

Lemma 4.4. *Suppose that $(Y, T_n) \to (Y, T)$ in total variation, where T_n, T are random variables taking values in $[0, 1]$. Define*

$$\Delta_n = \left|\mathbb{E}_{T_n}\left[|\mathbb{E}[Y \mid T_n = t] - t|\right] - \mathbb{E}_T\left[|\mathbb{E}[Y \mid T = t] - t|\right]\right|. \tag{4.2}$$

Then $\lim_{n\to\infty} \Delta_n = 0$.

Proof of Lemma 4.4. Let $\varepsilon_1 > 0$ be arbitrary. As a first step, we will apply a change of measure to write both expectations in (4.2) in terms of a single random variable. Let S_n denote a random variable that is distributed as T_n with probability

$\frac{1}{2}$ and T with probability $\frac{1}{2}$. Note that $\mathbb{P}_{S_n}$ is mutually absolutely continuous with both $\mathbb{P}_{T_n}$ and $\mathbb{P}_T$. Then by Markov's inequality, for $U \in \{T_n, T\}$,

$$\mathbb{P}_{S_n}\left(\left|\frac{d\mathbb{P}_U}{d\mathbb{P}_{S_n}} - 1\right| \geq \epsilon_1\right) \leq \frac{\mathbb{E}_{S_n}\left|\frac{d\mathbb{P}_U}{d\mathbb{P}_{S_n}} - 1\right|}{\epsilon_1} = \frac{d_{\mathrm{TV}}(T_n, T)}{\epsilon_1}.$$

By Lemma 4.3, for any $\delta_1 > 0$, we can choose n sufficiently large so that $d_{\mathrm{TV}}(T_n, T) \leq \epsilon_1\delta_1$, so that

$$\mathbb{P}_{S_n}\left(\left|\frac{d\mathbb{P}_X}{d\mathbb{P}_{S_n}} - 1\right| \geq \epsilon_1\right) \leq \delta_1. \tag{B.2}$$

We change the measure using $\mathbb{P}_{T_n}, \mathbb{P}_T \ll \mathbb{P}_{S_n}$:

$$\Delta_n \leq \left|\mathbb{E}_{S_n}\left[\frac{d\mathbb{P}_{T_n}}{d\mathbb{P}_{S_n}}\left|\mathbb{E}[Y \mid T_n = t] - t\right|\right]\right| - \left|\mathbb{E}_{S_n}\left[\frac{d\mathbb{P}_T}{d\mathbb{P}_{S_n}}\left|\mathbb{E}[Y \mid T = t] - t\right|\right]\right|. \tag{B.3}$$

We compare each term to $\mathbb{E}_{S_n}\left[\left|\mathbb{E}[Y|U=t] - t\right|\right]$ for $U = T_n, T$, respectively. By the triangle inequality,

$$\begin{aligned}\left|\mathbb{E}_{S_n}\left[\frac{d\mathbb{P}_U}{d\mathbb{P}_{S_n}}\left|\mathbb{E}[Y \mid U = t] - t\right|\right]\right| - \left|\mathbb{E}_{S_n}\left[\left|\mathbb{E}[Y \mid U = t] - t\right|\right]\right| &\leq \left|\mathbb{E}_{S_n}\left[\left|\frac{d\mathbb{P}_U}{d\mathbb{P}_{S_n}} - 1\right|\left|\mathbb{E}[Y \mid U = t] - t\right|\right]\right| \\ &\leq \mathbb{P}\left(\left|\frac{d\mathbb{P}_U}{d\mathbb{P}_{S_n}} - 1\right| \geq \epsilon_1\right) + \epsilon_1 \leq \delta_1 + \epsilon_1 \end{aligned} \tag{B.4}$$

where we split the expectation depending on whether $\left|\frac{d\mathbb{P}_{T_n}}{d\mathbb{P}_{S_n}} - \frac{d\mathbb{P}_T}{d\mathbb{P}_{S_n}}\right| \geq \epsilon_1$, note that $\frac{d\mathbb{P}_U}{d\mathbb{P}_{S_n}} \leq 2$ and $|\mathbb{E}[Y|U=t] - t| \leq 1$, and use Lemma B.2. From (B.3) and (B.4),

$$\begin{aligned}\Delta_n &\leq 2(\epsilon_1 + \delta_1) + \mathbb{E}_{S_n}\left[\left|\mathbb{E}[Y \mid T_n = t] - t\right|\right] - \mathbb{E}_{S_n}\left[\left|\mathbb{E}[Y \mid T = t] - t\right|\right] \\ &\leq 2(\epsilon_1 + \delta_1) + \mathbb{E}_{S_n}\left[\left|\mathbb{E}[Y \mid T_n = t] - \mathbb{E}[Y \mid T = t]\right|\right], \end{aligned} \tag{B.5}$$

where we used the triangle inequality. To simplify notation moving forward, we will use $d\mathbb{P}_{1,U}$ to denote the density of $(Y = 1, U)$. Finally,

$$\begin{aligned}\mathbb{E}_{S_n}\left[\left|\mathbb{E}[Y \mid T_n = t] - \mathbb{E}[Y \mid T = t]\right|\right] &= \mathbb{E}_{S_n}\left[\left|\frac{d\mathbb{P}_{1,T_n}}{d\mathbb{P}_{T_n}} - \frac{d\mathbb{P}_{1,T}}{d\mathbb{P}_T}\right|\right] \\ &= \mathbb{E}_{S_n}\left[\left|\frac{d\mathbb{P}_{1,T_n}}{d\mathbb{P}_{S_n}}\frac{d\mathbb{P}_{S_n}}{d\mathbb{P}_{T_n}} - \frac{d\mathbb{P}_{1,T}}{d\mathbb{P}_{S_n}}\frac{d\mathbb{P}_{S_n}}{d\mathbb{P}_T}\right|\right] \end{aligned} \tag{B.6}$$

For $U \in \{T_n, T\}$, we know $\frac{d\mathbb{P}_{S_n}}{d\mathbb{P}_U} \in [1 - O(\epsilon_1), 1 + O(\epsilon_1)]$ with probability $\geq 1 - \delta_1$ by (B.2). For the first factors, we have $\frac{d\mathbb{P}_{1,U}}{d\mathbb{P}_{S_n}} \approx \frac{d\mathbb{P}_{1,S_n}}{d\mathbb{P}_{S_n}}$ from the same logic as (B.2):

$$\mathbb{P}_{S_n}\left(\left|\frac{d\mathbb{P}_{1,U}}{d\mathbb{P}_{S_n}} - \frac{d\mathbb{P}_{1,S_n}}{d\mathbb{P}_{S_n}}\right| \geq \epsilon_1\right) \leq \frac{\mathbb{E}_{S_n}\left[\left|\frac{d\mathbb{P}_{1,U}}{d\mathbb{P}_{S_n}} - \frac{d\mathbb{P}_{1,S_n}}{d\mathbb{P}_{S_n}}\right|\right]}{\epsilon_1} = \frac{d_{\mathrm{TV}}((Y, T_n), (Y, T))}{\epsilon_1} \leq \delta_1$$

where we now use $(Y, T_n) \overset{\mathrm{TV}}{\to} (Y, T)$ and choose n sufficiently large so that $d_{\mathrm{TV}}((Y, T_n), (Y, T)) \leq \epsilon_1\delta_1$. By comparing both terms in (B.6) to $\frac{d\mathbb{P}_{1,S_n}}{d\mathbb{P}_{S_n}}$, we then get that (B.6) is $O(\epsilon_1 + \delta_1)$. Together with (B.5), we have $\Delta_n = O(\delta_1 + \epsilon_1)$, and since δ_1, ϵ_1 were arbitrary we have that $\lim_{n\to\infty} \Delta_n = 0$, as desired. □

C. Proofs for Section 5

Proposition 5.1. *Let Z be an arbitrary real-valued random variable and let ξ be a real-valued random variable with density p_ξ. Then $Z + \xi$ has the following density with respect to the Lebesgue measure:*

$$p_{Z+\xi}(t) = \mathbb{E}_Z[p_\xi(t - Z)]. \tag{5.1}$$

Proof. We have that:

$$\begin{aligned}\mathbb{P}(Z+\xi\leq t) &= \int_{-\infty}^{\infty} \mathbb{P}(\xi\leq t-z)\, d\mathbb{P}_Z(z)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{t-z} p_\xi(u)\, du\, d\mathbb{P}_Z(z)\\ &= \int_{-\infty}^{t}\int_{-\infty}^{\infty} p_\xi(u-z)\, d\mathbb{P}_Z(z)\, du \end{aligned} \tag{C.1}$$

from Fubini's theorem and translation invariance of the Lebesgue measure. □

Lemma 5.2. *If $\xi = \sigma R$ for a random variable R with bounded density, then:*

$$\mathbb{E}_\pi\left[\int_0^1 \left|p_{T,Y=1}(t) - p_{\hat{T},Y=1}(t)\right| dt\right] = O\left(\frac{1}{\sqrt{n\sigma}}\right), \tag{5.6}$$

$$\mathbb{E}_\pi\left[\int_0^1 \left|p_{T}(t) - p_{\hat{T}}(t)\right| dt\right] = O\left(\frac{1}{\sqrt{n\sigma}}\right). \tag{5.7}$$

Proof of Lemma 5.2. We recall that $p_{T,Y=1}(t)$ has the following form, which is a result of Proposition 5.1:

$$p_{T,Y=1}(t) = (\rho^{-1})'(t)\,\pi_Y(1)\,\mathbb{E}\left[p_\xi(\rho^{-1}(t)-h(X)) \mid Y=1\right]. \tag{C.2}$$

Now from Cauchy–Schwarz, Jensen's inequality, Fubini's theorem, and a change of variable (in that order), we obtain:

$$\begin{aligned}\mathbb{E}_\pi\left[\int_0^1 \left|p_{T,Y=1}(t) - p_{\hat{T},Y=1}(t)\right| dt\right] &\leq \mathbb{E}_\pi\left[\sqrt{\int_0^1 \left|p_{T,Y=1}(t) - p_{\hat{T},Y=1}(t)\right|^2 dt}\right]\\ &\leq \sqrt{\int_0^1 \mathbb{E}_\pi\left[\left|p_{T,Y=1}(t) - p_{\hat{T},Y=1}(t)\right|^2\right] dt}\\ &= \left(\int_0^1 \mathbb{E}_\pi\left[\left|(\rho^{-1})'(t)\,\pi_Y(1)\,\mathbb{E}\left[p_\xi(\rho^{-1}(t)-h(X)) \mid Y=1\right]\right.\right.\right.\\ &\qquad \left.\left.\left. - \frac{1}{n}\sum_{i=1}^n (\rho^{-1})'(t)\, p_\xi(\rho^{-1}(t)-h(x_i))\,\mathbb{1}_{y_i=1}\right|^2\right] dt\right)^{1/2}\\ &= \left(\int_{-\infty}^{\infty} \mathbb{E}_\pi\left[\left|\pi_Y(1)\,\mathbb{E}\left[p_\xi(u-h(X)) \mid Y=1\right]\right.\right.\right.\\ &\qquad \left.\left.\left. - \frac{1}{n}\sum_{i=1}^n p_\xi(u-h(x_i))\,\mathbb{1}_{y_i=1}\right|^2\right] du\right)^{1/2}. \end{aligned} \tag{C.3}$$

For simplicity, let us make the following definitions before moving forward:

$$f(u) = \pi_Y(1)\,\mathbb{E}\left[p_\xi(u-h(X)) \mid Y=1\right], \tag{C.4}$$

$$\hat{f}(u) = \frac{1}{n}\sum_{i=1}^n p_\xi(u-h(x_i))\,\mathbb{1}_{y_i=1}. \tag{C.5}$$

We observe that $\mathbb{E}_\pi[\hat{f}(u)] = f(u)$ for every u, since the (x_i, y_i) are i.i.d. according to π. Now we can continue from (C.3):

$$\begin{aligned}
\mathbb{E}_\pi\left[\int_0^1 \left|p_{T,Y=1}(t) - p_{\hat{T},Y=1}(t)\right| dt\right] &\leq \sqrt{\int_{-\infty}^{\infty} \mathbb{E}_\pi\left[\left|f(u) - \hat{f}(u)\right|^2\right] du} \\
&= \sqrt{\int_{-\infty}^{\infty} \mathrm{Var}_\pi(\hat{f}(u))\, du} \\
&\leq \sqrt{\frac{\pi_Y(1)^2}{n} \int_{-\infty}^{\infty} \mathbb{E}\left[p_\xi(u - h(X))^2 \mid Y = 1\right] du} \\
&= \sqrt{\frac{\pi_Y(1)^2}{n\sigma} \mathbb{E}\left[\int_{-\infty}^{\infty} \frac{M}{\sigma} p_R\left(\frac{u - h(X)}{\sigma}\right) du \mid Y = 1\right]} \\
&= O\left(\frac{1}{\sqrt{n\sigma}}\right) \qquad \text{(C.6)}
\end{aligned}$$

where $M = \sup p_R$. The result for $\left|p_T(t) - p_{\hat{T}}(t)\right|$ follows identically. □

D. Impact of Noise Distribution on LS-ECE

Here we examine the effect of changing the noise distribution of ξ on $\text{LS-ECE}_{\hat{\pi},\xi}$. In particular, we contrast the choice of Gaussian noise to compactly supported noise, and consider instead $\xi \sim \sigma\mathrm{Uni}([-1/2, 1/2])$. Figure 5 shows the results of redoing the experiments of Section 6.2.2 with this choice of ξ. As can be seen, the results are essentially the same as Figure 4.

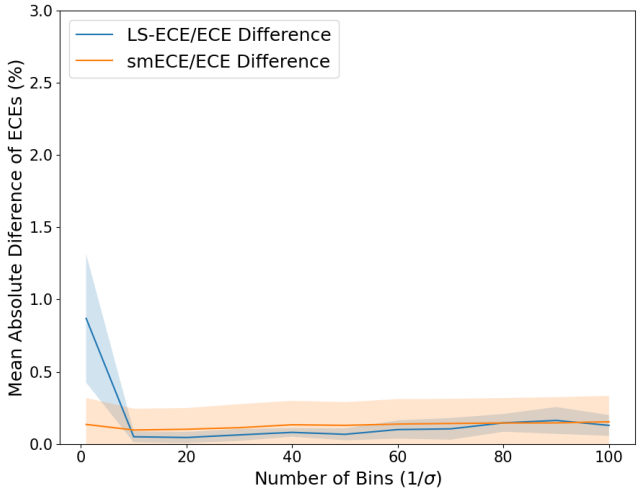


Figure 5. Mean absolute difference between ECE and LS-ECE (using uniform noise instead of Gaussian), as well as ECE and smECE, on ImageNet-1K-val over all models considered in Section 6.2.2, with one standard deviation error bounds marked using the shaded region.