Auditability and the Landscape of Distance to Multicalibration

Nathan Derhake
University of Southern California, Los Angeles, CA, USA

Siddartha Devic
University of Southern California, Los Angeles, CA, USA

Dutch Hansen
University of Washington, Seattle, WA, USA

Kuan Liu
University of Southern California, Los Angeles, CA, USA

Vatsal Sharan
University of Southern California, Los Angeles, CA, USA

Abstract

Calibration is a critical property for establishing the trustworthiness of predictors that provide uncertainty estimates. Multicalibration is a strengthening of calibration which requires that predictors be calibrated on a potentially overlapping collection of subsets of the domain. As multicalibration grows in popularity with practitioners, an essential question is: *how do we measure how multicalibrated a predictor is?* Błasiok et al. [4] considered this question for standard calibration by introducing the *distance to calibration* framework (dCE) to understand how calibration metrics relate to each other and the ground truth. Building on the dCE framework, we consider the auditability of the *distance to multicalibration* of a predictor f.

We begin by considering what are perhaps the two most natural generalizations of dCE to multiple subgroups: worst group dCE (wdMC), and distance to multicalibration (dMC). Using wdMC and dMC as a guiding path, we argue that there are two essential properties of any multicalibration error metric: 1) the metric should capture how much f would need to be modified in order to be perfectly multicalibrated; and 2) the metric should be auditable in an information theoretic sense (i.e., with some finite sample complexity). We show that wdMC and dMC each fail to satisfy one of these two properties, and that similar barriers arise when considering the auditability of general distance to multigroup fairness notions (*e.g.* multiaccuracy or low-degree multicalibration). We then propose two (equivalent) multicalibration metrics which do satisfy these requirements: 1) a continuized variant of dMC; and 2) a distance to *intersection* multicalibration, which leans on intersectional fairness desiderata.

Along the way, we shed light on the *loss-landscape* of distance to multicalibration and the geometry of the set of perfectly multicalibrated predictors. We also demonstrate that the loss surface of any metric which captures how much f would need to be modified to be perfectly multicalibrated often satisfies a *local minima are global minima* property. Our findings may have implications for the development of stronger multicalibration algorithms, as well as multicalibration auditing more generally.

2012 ACM Subject Classification Theory of computation → Machine learning theory

Keywords and phrases Multicalibration, Auditability, Fairness, Classification, Calibration

Digital Object Identifier 10.4230/LIPIcs.ITCS.2026.48

Related Version *Full Version*: `https://arxiv.org/abs/2509.16930` [10]

Funding *Vatsal Sharan*: Supported in part by an NSF CAREER Award CCF-2239265, an Amazon Research Award, a Google Research Scholar Award, and a Okawa Foundation Research Grant.

Acknowledgements We thank Parikshit Gopalan and Charlotte Peale for helpful discussions.

1 Introduction

Model calibration is an essential requirement in settings where trustworthy machine-learned predictions are used [9, 2]. Calibration requires that among all samples given score $p \in [0, 1]$ by a predictor f, exactly a p-fraction of those samples have positive label. However, an arbitrary predictor f – such as a neural network – may not output the exact same prediction p more than a handful of times, which makes directly measuring the true calibration error difficult. Instead, for much of the modern history of machine and deep learning, the (binned) *expected calibration error* (ECE) has been the most popular metric used to measure calibration [21, 47]. Nonetheless, numerous works have pointed to shortcomings of ECE [4, 32, 5, 35]. For example, a small perturbation to the predictor f can wildly change the ECE of f with respect to a ground truth distribution $\mathcal{D}$. In response to this, the calibration community has developed a dizzying array of alternative metrics such as kernel calibration error [39, 36, kCE], β-calibration error [34], Smooth ECE [6, smECE], and more [40, 41].

In order to provide firmer theoretical grounding for alternatives to ECE, [4] introduced an elegant unified framework for understanding the usefulness of calibration metrics. They posit that, in lieu of the ultimate goal of obtaining *better calibrated* predictors, the underlying desiderata for calibration error metrics should be: *low error implies a predictor can easily be post-processed to be calibrated, and high-error the opposite.* They do this by introducing the *distance to calibration error* (dCE) of a predictor f, which measures the minimal distance f would need to change (in a suitable ℓ_1 metric space of predictors) in order to obtain a calibrated predictor.[1] Different calibration metrics such as ECE or kCE are then understood as approximations to the dCE of f with varying degrees of usefulness.

Simultaneously and relatedly, model *multicalibration* has emerged as a recent quantity and tool from the algorithmic fairness literature [25]. Put simply, multicalibration posits that a predictor should be calibrated on a potentially rich, overlapping collection of subgroups of the data distribution. For example, a risk-predictor used by a bank to inform lending decisions [2] should be calibrated when conditioned on legally protected subgroups of the population such as particular races, residents from certain geographic area, individuals older than 67, etc. [46].

Although multicalibration originated as a tool within the fairness community, techniques and tools from the broader *multi-group* viewpoint have recently received attention in the theoretical computer science community more generally [20, 19, 43]. In particular, the multi-group literature has been used to obtain guarantees in *omniprediction* [18, 16, 42], robustness to distribution shift [29, 48], loss prediction [15], and more [12, 23]. On the empirical side, these algorithms have been found to be useful in improving the calibration of text and image classifiers [24, 3, 48], and even LLMs [11, 37].

As we start applying and measuring the effectiveness of multi-group post-processing notions in the wild, the very question of *measuring* multi-group calibration has become salient. Practitioners applying standard (non-*multi*) calibration metrics have a relatively good understanding of when different metrics such as ECE or smECE should be used based on the context and desired measurement properties [4, 6]. Practitioners of multicalibration, however, do not yet have much guidance on choice of measurement metrics. For example, prior proposed theory utilizes worst group, discretized ℓ_1 calibration errors [25, 23]. That is, they consider the subgroup of the domain with the highest discretized ℓ_1 calibration error as the multicalibration metric of interest. The empirically-minded works of [24] and [30] utilize worst-group unweighted smECE, while even others use kernel calibration error notions [38].

[1] This is non-trivial in part because the set of calibrated predictors is non-convex.

Given the smorgasbord of options directly extending from the standard calibration literature, our goal is to provide firm theoretical footing to practical advances in multicalibration. In particular, we ask:

(1) When can we measure the distance of a predictor to the nearest multicalibrated predictor?

(2) More broadly, when are distance to multi-group calibration notions efficiently auditable?

The first question (1) seeks to understand if the *distance* to calibration notion of [4] can directly be extended to multicalibration. A positive result could potentially provide grounded recommendations to practitioners measuring and utilizing multicalibration algorithms in practice. In addition, answering the question in any capacity also requires building a deeper understanding of the *geometry* of the set of multicalibrated predictors. This is because any distance to multicalibration notion is fundamentally a metric which measures the distance from a predictor f to the *set* of multicalibrated predictors for a ground truth distribution. Natural follow-up questions abound: What does this set look like? Is measuring the distance to this set feasible, and does the set have nice properties relative to the ground truth distribution?

In addition, answering (1) may help uncover properties of the *loss surface* of multicalibration in the metric space of all predictors. Many multicalibration post-processing algorithms happen to be boosting algorithms which combine and collect predictors into complicated sequences of updates [25, 23]. Can other types of algorithms which operate more like gradient descent on the loss surface of predictors exist? Boosting algorithms such as the one described in [25] can be viewed as a form of *functional* gradient descent. We are instead asking: are there algorithms which apply gradient updates directly on the predictor f (instead of some surrogate space), decreasing its multicalibration error at each step?

The second question (2) seeks to build upon three multi-group notions related to multicalibration: low-degree multicalibration [20], multiaccuracy [33, 38], and calibrated multiaccuracy [7, 17, 42]. We may be able to extend distance to multicalibration to these weaker multi-group fairness notions, unlocking many possible questions about the geometry and auditability of each of these special sets of predictors. Directions (1) and (2) are both important since they formally define the multi-group calibration evaluation problem, and may eventually allow for the development of more powerful algorithms for multi-group post-processing in a variety of domains.

1.1 Notation and Background

Fix a feature space $\mathcal{X}$ with finite (but large) cardinality $|\mathcal{X}| = n$, and define the label space as $\mathcal{Y} = \{0, 1\}$. Define a predictor as a function $f : S \to [0, 1]$ for an arbitrary set $S \subseteq \mathcal{X}$. Let $\mathcal{F}_{\mathcal{X}} = [0, 1]^{\mathcal{X}}$ denote the set of all predictors on $\mathcal{X}$. We sample example-outcome pairs $(\mathbf{x}, \mathbf{y})$ i.i.d. according to joint distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Equivalently, we can view this as first sampling $\mathbf{x}$ from a marginal distribution $\mathcal{D}_{\mathbf{x}}$ over the domain $\mathcal{X}$, and then sampling $\mathbf{y} \sim \text{Bernoulli}(p^*(\mathbf{x}))$ for some ground truth $p^* \in \mathcal{F}_{\mathcal{X}}$. We further assume that each $\mathcal{D}_{\mathcal{X}}$ is supported on all of $\mathcal{X}$. We will denote the set of all such distributions as $\Delta(\mathcal{X} \times \mathcal{Y})$.

▶ **Definition 1** (Calibration). *A predictor $f : \mathcal{X} \to [0, 1]$ is* perfectly calibrated *with respect to a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ if for every $v \in Im(f)$, we have:*

$$\mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\Big[\mathbf{y} \mid f(\mathbf{x}) = v\Big] = v.$$

We denote the set of all such predictors $\text{cal}(\mathcal{D})$.

Calibration is a well studied property of predictors which output uncertainty estimates in $[0, 1]$. Calibrated predictions are useful in a broad variety of settings spanning online learning [44], collaboration [8], conformal prediction [31], and classification with neural networks [21]. The task of *measuring* calibration has given rise to its own line of research (see [45] for a survey). In particular, a variety of calibration metrics now exist: kCE [39], smECE [6], ECE, etc. In order to derive an understanding of this veritable smorgasbord, [4] define the notion of *distance to calibration error* (dCE) of a predictor f.

Before proceeding, we establish notation. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and $f, g : \mathcal{X} \to \mathbb{R}$, we let

$$\|f\|_p := \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}_{\mathbf{x}}} \left[|f(\mathbf{x})|^p \right] \right)^{1/p} \quad \text{and} \quad \ell_p(f, g) := \|f - g\|_p.$$

Given $H \subseteq [0, 1]^{\mathcal{X}}$, we define the ℓ_p-distance to H as follows.

$$\ell_p(f, H) := \inf_{h \in H} \ell_p(f, h)$$

Notice that in our context, the ℓ_1 metric and norm differ slightly from the traditional norms and metrics on $\mathbb{R}^{|\mathcal{X}|}$ because we weigh each coordinate by the marginal distribution $\mathcal{D}_{\mathbf{x}}$.

▶ **Definition 2** (Distance to Calibration Error [4]). *For $f : \mathcal{X} \to [0, 1]$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we define*

$$\mathrm{dCE}_{\mathcal{D}}(f) := \ell_1(f, \mathrm{cal}(\mathcal{D})) = \inf_{g \in \mathrm{cal}(\mathcal{D})} \ell_1(f, g).$$

Note that $\mathrm{cal}(\mathcal{D})$ is finite [10] and hence closed under the topology induced by the ℓ_p metric for any $p \geq 1$; we can therefore replace the inf with a min in the definition of dCE.[2] As in [4], our presentation focuses on the ℓ_1 metric.

The definition of dCE is based on measuring how far a predictor f is from *perfect* calibration. This allows for [4] to understand and characterize calibration metrics like ECE which are used in practice. In particular, they show that ECE can be understood to give an *upper* bound on dCE, which implies that if ECE is small, the predictor is truly close to being perfectly calibrated (a soundness condition on the distance to the set of perfectly calibrated predictors). On the other hand, large ECE does *not* guarantee that the predictor is actually far from being calibrated (i.e., does not satisfy completeness). Similar completeness and soundness guarantees can be derived for other proposed calibration metrics.

Generalizing to Multicalibration

Let a collection $\mathcal{C} = \{S_1, \ldots, S_k\}$ with $S_i \subseteq \mathcal{X}$ be given. Multicalibration is a *strengthening* of calibration which requires a predictor to be simultaneously calibrated when conditioned on each $S_i \in \mathcal{C}$. To formalize this within the notation of dCE, we require restrictions of predictors, distributions, and metrics. It will be useful to generalize some of the objects already introduced. For $f : \mathcal{X} \to [0, 1]$ and $S \subset \mathcal{X}$, we let $\mathcal{D}|_S$ denote the distribution $\mathcal{D}$ conditioned on $\mathbf{x} \in S$; this is equivalent to conditioning $\mathcal{D}_{\mathbf{x}}$ on $\mathbf{x} \in S$ and drawing $\mathbf{y}$ according to ground truth $p^*|_S$. We will sometimes use a conditional version of the ℓ_p norm and metric:

$$\|f\|_{p,S} := \left(\mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}|_S} \left[|f(\mathbf{x})|^p \right] \right)^{1/p} \quad \text{and} \quad \ell_p|_S(f, g) := \|f - g\|_{p,S}.$$

[2] Finiteness of $\mathrm{cal}(\mathcal{D})$ was also used and pointed out in [4].

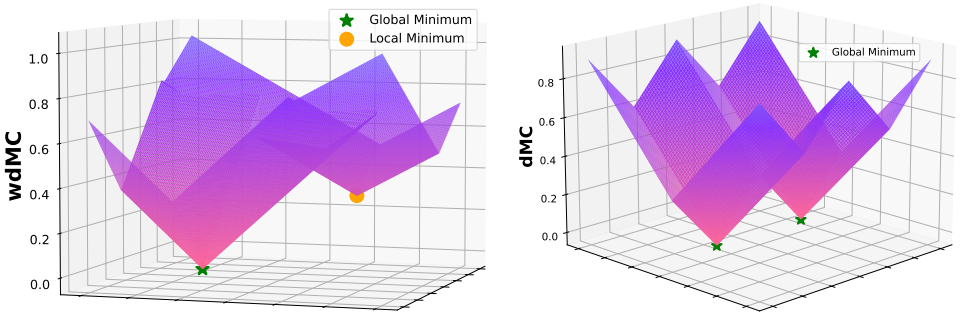

Figure 1 (**Left**): Proposition 5 demonstrates that the worst-group distance to calibration error function (wdMC) contains *local* minima which are not global minima in the space of all predictors. That is, there are predictors f which appear *locally not improvable* in terms of worst-group calibration error, but which can be improved with a large enough change. (**Right**): Conversely, Proposition 7 demonstrates that all local minima are global for dMC, even though the loss landscape may be non-convex.

We slightly abuse notation by writing $\ell_1|_S(f, g)$ when the domain of f and g contain S. We will sometimes say that a predictor $f \in \mathcal{F}_\mathcal{X}$ is calibrated on (or with respect to) S; formally, we mean that $f|_S$ is calibrated *w.r.t.* $\mathcal{D}|_S$.

▶ **Definition 3** (Multicalibration). *Given a collection $\mathcal{C} = \{S_1, \ldots, S_k\}$ with each $S_i \subseteq \mathcal{X}$, a predictor $f : \mathcal{X} \to [0, 1]$ is said to be* perfectly $\mathcal{C}$-multicalibrated *(or just multicalibrated) if f is perfectly calibrated with respect to S_i for all $i \in [k]$. We denote the set of such predictors as* $\mathrm{mcal}_\mathcal{C}(\mathcal{D})$.

We remark that p^* is always in the set $\mathrm{mcal}_\mathcal{C}(\mathcal{D})$ for the distribution $\mathcal{D}$ that it defines (the ground truth is perfectly calibrated on any subset). However, $\mathrm{mcal}_\mathcal{C}(\mathcal{D})$ is most interesting when it contains predictors *other* than p^*. Lastly, note that multicalibration says nothing about how f needs to behave on elements of $\mathcal{X}$ not covered by $\mathcal{C}$. Even even when $\mathcal{C}$ does cover $\mathcal{X}$, a predictor need not be calibrated over $\mathcal{X}$ to be perfectly $\mathcal{C}$-multicalibrated. It is therefore without loss of generality to assume that $\mathcal{C}$ is a cover of $\mathcal{X}$, *i.e.* $\bigcup_i S_i = \mathcal{X}$. We will nevertheless stipulate when this condition is required.

1.2 Summary of Contributions

1.2.1 A Practical Error Notion

Motivated by practitioners [24, 30], we begin by defining a straightforward ways to measure the distance of f to multicalibration. Let a subgroup collection $\mathcal{C}$ be given. In Section 2.1, we first introduce the worst-group distance to calibration error, denoted by $\mathrm{wdMC}_\mathcal{C}(f)$. This is defined as the maximum over all subgroups of how far f is from being calibrated in terms of dCE, when restricting f and the distribution $\mathcal{D}$ to only that subgroup. Importantly, $\mathrm{wdMC}_\mathcal{C}$ is also weighted by the "size" (mass) of each subgroup S_i in the marginal distribution $\mathcal{D}_\mathbf{x}$. This weighting ensures that a vanishingly small group will not have disproportionate impact on the multicalibration error, which intuitively would make it difficult to audit.[3]

[3] Consider a subgroup which appears with exponentially small probability in $\mathcal{D}_\mathbf{x}$, but introduces a high calibration error for f. Such a subgroup would not be detectable in a polynomial number of samples, yet would maximize an unweighted wdMC.

We then discuss the *loss-landscape* viewpoint of multicalibration error notions. The metric $\text{wdMC} : \mathcal{F}_{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ maps from the space of predictors to real numbers; in particular, it defines a *loss surface*. Multicalibration algorithms can be viewed as traversing this loss surface. In order for gradient-based multicalibration algorithms to be effective, a reasonable requirement on the loss surface is that local minima are global minima. If this were not the case, then any such algorithm could get stuck in a sub-optimal local minima and not be able to improve the multicalibration error. Existing boosting-based multicalibration algorithms do not have this issue, and can drive the objective functions as low as desired with enough samples [23, 25].

Unfortunately, in Proposition 5 we demonstrate that low wdMC *does not* imply that a predictor is close to being perfectly multicalibrated. In particular, we exhibit a multicalibration auditing instance – a distribution $\mathcal{D}$, subgroup collection $\mathcal{C}$, and predictor f to be audited – for which $\text{wdMC}(f) = \varepsilon$, but f is at a local minima which is not global. That is, the closest predictor $\tilde{f}$ which has $\text{wdMC}(\tilde{f}) < \varepsilon/2$ is $\Omega(1)$ distance away from f. This is illustrated in Figure 1.

1.2.2 Local and Global Minima

Using wdMC, we have developed one natural requirement for any multicalibration error metric: all local minima should be global minima. To directly address this, in Section 2.2 we introduce another multicalibration error metric: *distance to multicalibration* (denoted $\text{dMC}_{\mathcal{C}}(f)$). This metric is defined exactly as the distance from f to its projection onto the set of perfectly multicalibrated predictors defined in Definition 3. In Proposition 7, we show that dMC does indeed satisfy this property, since it is a function which computes the distance of f to a set. The set of perfectly multicalibrated predictors is actually a union of finitely many affine subspaces of $\mathcal{F}_{\mathcal{X}}$, and since dMC is defined with the ℓ_1 metric, the loss landscape itself turns out to be piecewise-linear.

Next, we turn to auditing or computing $\text{dMC}_{\mathcal{C}}(f)$. Throughout most of our presentation, we are concerned with *information-theoretic* auditability, which asks whether we can compute the metric or quantity of interest with any finite number of samples.[4] In Proposition 9, we demonstrate a pathology of dMC which makes it impossible to audit in this information theoretic sense. We construct a simple instance for which determining whether $\text{dMC}_{\mathcal{C}}(f) = 0$ or $\text{dMC}_{\mathcal{C}}(f) > C$ for a constant $C > 0.2$ requires finding the *exact* value of the ground truth predictor $p^*(x)$ on some point $x \in \mathcal{X}$. This is impossible with any (finite) number of samples, and hence, dMC is in general, inauditable. Near the end of Section 2.2 and in [10], we show that similar instances exist for distance to low-degree multicalibration [20] and calibrated multiaccuracy [7], two recently proposed relaxations of full multicalibration which may be of interest to the community.

1.2.3 Stepping Towards Auditability

The pathology exhibited in Proposition 9 arises precisely because a minuscule change in the ground truth $p^*(x)$ on one particular point x can cause the $\text{dMC}_{\mathcal{C}}(f)$ to rapidly change by a constant. That is, $\text{dMC}_{\mathcal{C}}(f)$ is not *Lipschitz* in the ground truth predictor p^*. Armed with this intuition, we seek new multicalibration error notions which can simultaneously satisfy two conditions:

[4] We also sometimes consider statistical auditability, (i.e., can we compute the metric in a number of samples polynomial in $|\mathcal{C}| = k$?) and will explicitly state when this is the case.

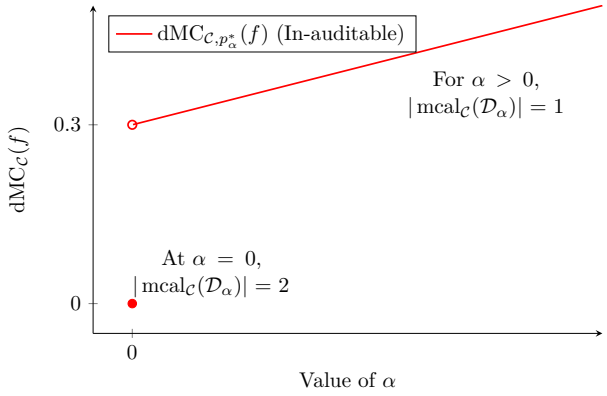

Figure 2 An illustration showing that the distance to multicalibration error (dMC) of a predictor f can be a non-continuous function of the ground truth predictor p^*. Here, p^* is parameterized by a parameter α. For small enough $\alpha > 0$, any finite number of samples is not enough to determine whether $\alpha = 0$ or $\alpha \neq 0$. Since $\mathrm{dMC}_{\mathcal{C}}(f) = 0$ if $\alpha = 0$, and $\mathrm{dMC}_{\mathcal{C}}(f) \geq 0.3$ otherwise, this demonstrates an information-theoretic barrier to auditing dMC. This happens because the number of multicalibrated predictors $|\mathrm{mcal}_{\mathcal{C}}(p^*)|$ grows to two at $\alpha = 0$, but for an $\alpha > 0$, there is only a single multicalibrated predictor (given by p^*). Since $\mathrm{dMC}_{\mathcal{C}}(f)$ measures the distance of f to $\mathrm{mcal}_{\mathcal{C}}(p^*)$, it experiences an induced discontinuity. See Proposition 9 for details.

(1) All local minima of the loss landscape induced by the metric are global minima;
(2) When holding f, $\mathcal{D}$, and $\mathcal{C}$ fixed, the metric is Lipschitz in the ground truth function p^*.

We note that (1) can be viewed as a form of a soundness condition relative to the distance of f to the set of perfectly multicalibrated predictors. On the other hand, (2) provides a *necessary* condition for information-theoretic auditability of the metric using samples from $\mathcal{D}$. If (2) was violated by non-Lipschitzness or C-Lipschitzness for some very large C, then arbitrarily small changes in the ground truth p^* could induce very large changes in the multicalibration metric. This would then require approximating p^* arbitrarily well in order to get any good estimate on the metric.

With these two basic requirements in hand, in Section 3.1 we examine why the non-Lipschitzness used in the instance of Proposition 9 occurs. Intuitively, there is a measure-zero set of "rogue" ground truth predictors which are the points of discontinuity of $\mathrm{dMC}_{\mathcal{C}}(f)$. This is illustrated in Figure 2, where a point-wise discontinuity causes dMC to change by a constant. To rid ourselves of this, we introduce a *continuized* distance to multicalibration error metric, which we denote by $\widetilde{\mathrm{dMC}}$ (Definition 11). At a high level, $\widetilde{\mathrm{dMC}}$ is defined by carefully *smoothing* the dMC over a sequence of local neighborhoods around the ground truth predictor p^*.

1.2.4 Suitable Metric(s)

In Theorem 13, we show that $\widetilde{\mathrm{dMC}}_{\mathcal{C}}(f)$ satisfies properties (1) and (2). Our proof of this is presented in Section 3.3 with lemmas developed in Section 3.2. Interestingly, the proof relies on yet another multicalibration error metric: the distance to *intersection* multicalibration error, denoted by $\mathrm{dIMC}_{\mathcal{C}}(f)$. This metric is simply defined as $\mathrm{dIMC}_{\mathcal{C}}(f) := \mathrm{dMC}_{\mathcal{I}(\mathcal{C})}(f)$, where $\mathcal{I}(\mathcal{C})$ is the set of all intersections of sets from $\mathcal{C}$. Measuring notions of fairness for *intersections* of subgroups is an important and difficult problem in the fairness community more broadly [14, 26], and hence, dIMC may be of independent interest.

In Proposition 20, we demonstrate that for all subgroup collections $\mathcal{C}$ (which are covers of the domain $\mathcal{X}$), distributions $\mathcal{D}$, and predictors f, we have that $\mathrm{dIMC}_{\mathcal{C}}(f) = \widetilde{\mathrm{dMC}}_{\mathcal{C}}(f)$. This reveals the surprising fact that *smoothing* dMC by removing the point-wise discontinuities is *equivalent* to measuring a strong, intersectional distance to multicalibration error notion. Using this fact, we can think of $\widetilde{\mathrm{dMC}}_{\mathcal{C}}(f)$ as exactly the ℓ_1 distance from f to the nearest predictor which is multicalibrated over $\mathcal{C}$ *and* all intersections of groups in $\mathcal{C}$.

In Lemma 19, we show that there is a *disjoint* partition $\mathcal{J}(\mathcal{C})$ of the domain $\mathcal{X}$ (defined in Definition 18) for which $\mathrm{dMC}_{\mathcal{J}(\mathcal{C})}(f) = \mathrm{dMC}_{\mathcal{I}(\mathcal{C})}(f) = \mathrm{dIMC}_{\mathcal{C}}(f)$. Why is such a disjoint partition helpful? In Lemma 17, we show that if the collection $\mathcal{C}$ is disjoint, then $\mathrm{dMC}_{\mathcal{C}}(f)$ reduces to a weighted sum of dCE (Definition 2) with respect to each subset in the collection $\mathcal{C}$. It follows that $\mathrm{dIMC}_{\mathcal{C}}(f)$ is a weighted sum of dCE with respect to each set in $\mathcal{J}(\mathcal{C})$. Since dCE satisfies property (2) in that it is Lipschitz with respect to the ground truth function p^* (Lemma 10), the disjointness of $\mathcal{J}(\mathcal{C})$ implies this carries through to dIMC. Using this interpretation, in Theorem 15 we show that $\mathrm{dIMC}_{\mathcal{C}}(f)$ – equivalently, $\widetilde{\mathrm{dMC}}_{\mathcal{C}}(f)$ – satisfies both desiderata (1) and (2) of a multicalibration error metric, completing the proof of Theorem 13.

1.2.5 (In)auditability of $\widetilde{\mathrm{dMC}}$ and dIMC

In Section 3.4, we discuss auditability of dIMC (equivalently, $\widetilde{\mathrm{dMC}}$) using the perspective developed in Section 3.2. In particular, we make use of the fact that dIMC is equivalent to a weighted average of dCE over the disjoint partition of $\mathcal{X}$ defined by $\mathcal{J}(\mathcal{C})$.

By *statistical* auditability of dCE ([4]), dIMC should be statistically feasible to audit. Nonetheless, the partition $\mathcal{J}(\mathcal{C})$ which we chose can turn out to have too many sets. In particular, it could be the case that $|\mathcal{J}(\mathcal{C})| = \Theta(2^{|\mathcal{C}|})$, meaning that auditing dIMC could require checking the dCE of exponentially many restrictions of f. Since a majority of the density in $\mathcal{D}_{\mathbf{x}}$ could be placed on sets that we never sample, in Proposition 23 we demonstrate an example where we cannot distinguish whether $\mathrm{dIMC}_{\mathcal{C}}(f) = 0$ or $\mathrm{dIMC}_{\mathcal{C}}(f) = 0.5$ using $\mathsf{poly}(|\mathcal{C}|)$ samples. On the flip side, Proposition 26 shows that when $\mathcal{J}(\mathcal{C})$ satisfies certain "nice" conditions – such as sufficient mass on a small number of sets in $\mathcal{J}(\mathcal{C})$, or $\mathcal{C}$ itself being relatively small or constant – there is hope for a statistical (and computational) auditability result of dIMC in $\mathsf{poly}(|\mathcal{C}|)$ samples. This is noteworthy because in the practice of fair machine learning, we often expect there to be a small number of subgroups relative to the size of the domain $\mathcal{X}$.

1.2.6 Distance to Multiaccuracy

Given the nature of the results presented in Section 3.4, in Section 4 of the full version [10] we ask: *Is there a natural relaxation of full multicalibration for which a "distance to" notion is statistically auditable?* To address this, we define and analyze the *weakest* relaxation of multicalibration: multiaccuracy [33]. Multiaccuracy asks that a predictor f be *unbiased* over a collection $\mathcal{C}$. We begin by introducing $\mathrm{wdMA}_{\mathcal{C}}(f)$ as a "worst-group bias" of f on a collection $\mathcal{C}$ (this is defined analogously to wdMC for multicalibration). We also define $\mathrm{dMA}_{\mathcal{C}}(f)$, which measures the distance of a predictor f to closest *perfectly multiaccurate* predictor (analogous to dMC).

We demonstrate that the loss landscapes of both functions are *convex*, which implies that all local minima are global minima. Interestingly, we then show that wdMA and dMA are different quantities, and $\mathrm{wdMA}_{\mathcal{C}}(f)$ fails to satisfy that "low error implies f is close to a multiaccurate predictor". In particular, we exhibit an instance for which a predictor f

has $\text{wdMA}_{\mathcal{C}}(f) = \varepsilon$, but the distance $\ell_1(f, \tilde{f})$ from the predictor f to *any* predictor $\tilde{f}$ with $\text{wdMA}_{\mathcal{C}}(\tilde{f}) \leq \varepsilon/6$ is $\Omega(\varepsilon \cdot d^{|\mathcal{C}|})$, where the constant $d > 1$ is the square root of the golden ratio. This emerges because the construction used in the proof uses a subgroup collection based on the Fibonacci recurrence. By highlighting this undesirable property of wdMA, this instance further motivates the definition of dMA.

Lastly, we close by providing a linear program for computing $\text{dMA}_{\mathcal{C}}(f)$ exactly given perfect information on the average value of p^* over each subgroup. The program has $2 \cdot |\mathcal{X}|$ variables (to model the closest multiaccurate predictor to f), and $O(n + k)$ constraints (to ensure that the modeled predictor is unbiased on each group). We leave open the question of statistically and computationally feasible auditing of dMA.

1.3 Related Work

We remark that we are not the first to study distance to multicalibration error. Given that it is a natural generalization of distance to *calibration* error (dCE, [4]), the recent work [13] shows that achieving low distance to multicalibration in an online setting may be intractable. This may translate to impossibility results in the offline setting via online-to-batch reductions. In general, we are interested in understanding whether we can even *measure* notions of distance to multicalibration error, let alone achieve them. This distinction between our work is discussed further in Remark 8 within Section 2.2.

There is also recent empirical work on measuring multicalibration in practice. In [22], the authors propose using a worst-group calibration error notion which weighs groups by the *signal-to-noise* ratio of the predictor restricted to that group. They demonstrate the superiority of their measurement method in a variety of practical settings. Their framing stands in parallel to how we define wdMC, which weighs groups by their size. Indeed, [22] emphasize the importance of not having multicalibration metrics being thrown off by a very small group for which f is poorly calibrated or inaccurate on. This further aligns with the results of [24], who found that the worst-group calibration error is often determined by small groups in the data distribution.

2 A Loss Landscape Approach to Multicalibration Measures

In Section 2.1, we formally define the worst-group distance to calibration error notion which has been used in prior work (wdMC). In Section 2.2, we introduce our *distance to multicalibration* notion, dMC, which satisfies two properties that we argue are important for any multicalibration metric. Finally, we close with an obstruction to auditing dMC, and how this relates to auditability of the standard (non-multi) calibration metric of dCE [4].

2.1 Measuring Multicalibration via Worst-Group Distance to Calibration

The most common way to measure multicalibration in both theory [25] and practice [24] is *worst-group* calibration error. To model this within the framework of [4], we consider an analogous definition: *worst-group* distance to calibration, denoted as wdMC.

▶ **Definition 4** (Worst-group Distance to Calibration Error). *Let a subgroup collection $\mathcal{C} = \{S_1, \ldots, S_k\}$ with $S_i \subseteq \mathcal{X}$ and a predictor $f : \mathcal{X} \to [0, 1]$ be given. We define the worst-group distance to calibration error* $\text{wdMC}_{\mathcal{C}}(f)$ *as:*

$$\text{wdMC}_{\mathcal{C}}(f) := \max_{S \in \mathcal{C}} \left(\mathbb{P}[S] \cdot \text{dCE}_{\mathcal{D}|_S}(f|_S) \right).$$

The metric wdMC is a natural way to generalize dCE to multiple subgroups. In particular, a small overall wdMC implies that on each subgroup $S \in \mathcal{C}$, there is a perfectly calibrated predictor $g : S \to [0, 1]$ for which $\ell_1|_S(f, g)$ is small.[5] Put differently, f has small (distance to) calibration error when restricted to each subgroup, which is precisely the optimization objective of most multicalibration post-processing algorithms. Nonetheless, these post-processing algorithms usually use notions like worst-group ECE or ℓ_p calibration error, instead of worst-group distance to calibration error [23, 25].

We proceed by viewing $\mathcal{F}_\mathcal{X}$ as a compact subset of a $|\mathcal{X}|$-dimensional vector space and $\text{wdMC} : \mathcal{F}_\mathcal{X} \to \mathbb{R}_{\geq 0}$ as a continuous hyper-surface – or *loss landscape* – over this space. One can view multicalibration post-processing algorithms as optimizing over this loss landscape (or some parameterized representation of it). Given the limited use of direct gradient-based methods in multicalibration algorithms, a natural question is whether this loss landscape presents any barriers to tried-and-true continuous optimization techniques.

A very basic property of this loss-landscape would be that there are no local minima which are not global. Such a property would guarantee that if we could perform gradient steps, we would not stumble upon a sub-optimal solution. Unfortunately, via a simple example, we first demonstrate that $\text{wdMC}_\mathcal{C}(f)$ does not satisfy this property. In fact, it's possible for a local minimum to be *far* from any predictor improving on this local minimum by a constant fraction. The proof is deferred to [10].

▶ **Proposition 5.** *Take $\varepsilon > 0$ sufficiently small. There exists a distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, subgroups $\mathcal{C} = \{S_1, S_2\}$, and predictor f such that f locally minimizes $\text{wdMC}_\mathcal{C}$ and $\text{wdMC}_\mathcal{C}(f) \leq \varepsilon$, but for any $\tilde{f}$ satisfying $\text{wdMC}_\mathcal{C}(\tilde{f}) \leq \varepsilon/2$, it holds that $\ell_1(f, \tilde{f}) \in \Omega(1)$.*

The intuition for Proposition 5 is conveyed in Figure 3. In particular, Proposition 5 exhibits a distribution $\mathcal{D}$, collection $\mathcal{C} = \{S_1, S_2\}$, and predictor to be audited f which is ε-close to being calibrated on both groups S_1 and S_2 (and further, f is a local minimum of $\text{wdMC}_\mathcal{C}(f)$). However, the only place where the sets $\text{cal}_{S_1}(\mathcal{D})$ and $\text{cal}_{S_1}(\mathcal{D})$ intersect is actually very far from f. As a bottom line, Proposition 5 demonstrates that measuring the worst-group calibration error does not tell us anything about the *minimum* a predictor f may be required to change – in the geometry of the ℓ_p metric space – in order to be converted into a predictor with lower error.

Proposition 5 may seem somewhat counter-intuitive given the results we have on achieving multicalibration via post-processing algorithms. Works such as [25, 23] provide theorems which can drive down the worst-group calibration error of a predictor f to ε with $O(\frac{1}{\text{poly}(\varepsilon)})$ examples for arbitrarily small ε.[6] Proposition 5 does not contradict these results precisely because multicalibration post-processing algorithms do *not* give guarantees on how much the input predictor f may be required to be changed at different error thresholds. Importantly, it instead demonstrates that there exist certain thresholds ε at which the predictor f may be required to change *drastically* in order to achieve a worst-group calibration error below a constant fraction of ε.

[5] For technical reasons, we define wdMC with a normalization term based on the subgroup size. We note that our main result for wdMC, presented in Proposition 5, can be extended to a variant of wdMC which is simply the non-weighted, absolute worst group dCE. However, this non-weighted version has little hope of being auditable, since subgroups with very small mass in $\mathcal{D}_\mathbf{x}$ may drive up the worst-group dCE in unpredictable ways.

[6] Note that these algorithms objective functions are often a discretized, ℓ_1 worst-group calibration error notion, *not* the worst-group distance to calibration error wdMC that we consider.

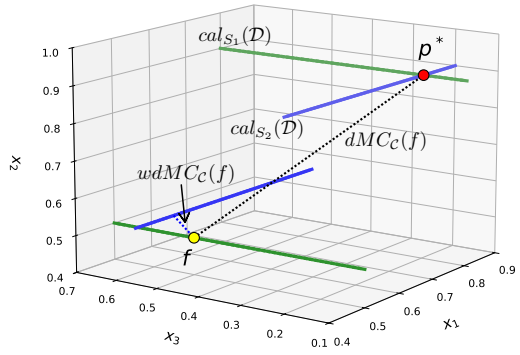

Figure 3 A visual illustration of the difference between wdMC and dMC. The example pictured is detailed in the proof of Proposition 9, setting $\alpha = 0.1$. The green lines represent the set of predictors $\text{cal}_{S_1}(\mathcal{D})$ calibrated with respect to S_1, and the blue lines are predictors from the set $\text{cal}_{S_2}(\mathcal{D})$ calibrated with respect to S_2. In this example, the predictor f is calibrated with respect to S_1, and has small worst-group calibration error wdMC, represented by the dotted distance of f to the closest predictor calibrated on S_2. However, f is far from p^*, which is the only predictor which is simultaneously calibrated on S_1 *and* S_2. Hence, f has low wdMC but high distance to the nearest multicalibrated predictor, which is measured by dMC.

2.2 A Nicely Behaved Multicalibration Error Loss Surface

Proposition 5 directly motivates our next question: Is there a multicalibration error metric whose loss surface has the property that *any local minima are global minima?* If such a metric existed, it would suggest the possibility of gradient-based algorithms that optimized this metric directly to find multicalibrated predictors.

We will argue that such a metric exists, but at a cost. To start, we consider another multi-group analog of dCE.

▶ **Definition 6** (Distance to Multicalibration Error). *Let a predictor f, distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, and subgroup collection $\mathcal{C}$ be given. We define the distance to multicalibration* $\text{dMC}_{\mathcal{C}}(f)$ *as*

$$\text{dMC}_{\mathcal{C},\mathcal{D}}(f) = \inf_{g \in \text{mcal}_{\mathcal{C}}(\mathcal{D})} \ell_1(f, g).$$

When clear from context, we omit the distribution subscript and write $\text{dMC}_{\mathcal{C}}(f)$.

The main difference between wdMC and dMC is that the latter considers "perfect" predictors to be those that are calibrated on all subgroups in $\mathcal{C}$ simultaneously. This is illustrated with an example in Figure 3. Furthermore, in [10] we show that wdMC is always upper bounded by dMC.

Just as $\text{cal}(\mathcal{D})$, $\text{mcal}_{\mathcal{C}}(\mathcal{D})$ is generally not convex. Neither is it true that $\text{dMC}(f)$ is convex in f. However, we demonstrate that dMC does satisfy the property that all local minima are global minima.

▶ **Proposition 7.** *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, $\mathcal{C} \subset 2^{\mathcal{X}}$, and $f \in \mathcal{F}_{\mathcal{X}}$. All local minima of* $\text{dMC}_{\mathcal{C}}$ *are global minima.*

We note that this claim is a special case of a much more general fact about distances to sets in normed vector spaces; we leave this generalization to the appendix of the full version [10].

2.2.1 On the Geometry of dMC and Perfectly Multicalibrated Predictors

In addition to nice minima, dMC behaves more nicely than wdMC with respect to its decomposition into convex functions.

In particular, take $\mathcal{C} = \{S_1, \ldots, S_k\}$. Given a subset $S \subseteq X$, and a predictor $g : S \to [0,1]$, let $A_g \subseteq [0,1]^{\mathcal{X}}$ be a subset of predictors such that for all $g' \in A_g$, we have $g'|_S = g$. Note A_g is an affine subspace of $\mathbb{R}^{\mathcal{X}}$ intersected with $[0,1]^{\mathcal{X}}$. In particular, A_g is compact and convex. Hence, wdMC can be seen as the maximum over finitely many minimums of convex functions.

$$\begin{aligned}\mathrm{wdMC}_{\mathcal{C}}(f) &= \max_{S_i \in \mathcal{C}} \inf_{g \in \mathrm{cal}(\mathcal{D}|_{S_i})} \mathbb{P}[S_i] \cdot \ell_1|_{S_i}(f,g) \\ &= \max_{S_i \in \mathcal{C}} \min_{g \in \mathrm{cal}(\mathcal{D}|_{S_i})} \ell_1(f, A_g) \end{aligned} \tag{1}$$

In contrast, dMC is the minimum of finitely many convex functions from $[0,1]^{\mathcal{X}} \to \mathbb{R}$. We know $\mathrm{mcal}_{\mathcal{C}}(\mathcal{D})$ is a finite set (see full version [10]), and therefore we can write dMC as follows:

$$\mathrm{dMC}_{\mathcal{C},\mathcal{D}}(f) = \min_{g \in \mathrm{mcal}_{\mathcal{C}}(\mathcal{D})} \ell_1(f,g). \tag{2}$$

When using ℓ_1 norm, both wdMC and dMC are piecewise linear functions. The decompositions seen in Equation (1) and Equation (2) provide intuition for some of the results that follow.

▶ Remark 8 (Auditability of dMC). While Proposition 7 and the decomposition in Equation (2) suggests that we can optimize dMC to obtain good solutions, it has been shown in restricted contexts that achieving low dMC may be difficult. [13] show that minimizing dMC – which they refer to as "strong" distance to multicalibration error – may be impossible in a general online setting, and online-to-batch results suggest that this may hold even in the offline setting. The hard instance that they present uses a number of groups k on the order of the size of the domain $|\mathcal{X}|$. We find reason to be even more pessimistic; we demonstrate an instance on which dMC cannot even be *audited* in a property testing sense, as studied in the calibration context by [28]. Even with only two groups, each with a size given as a constant fraction of the domain, we find an information-theoretic barrier: discontinuities w.r.t. p^*.

▶ **Proposition 9.** *There exists a domain $\mathcal{X}$, subgroups $\mathcal{C} \subset 2^{\mathcal{X}}$, a predictor $f \in \mathcal{F}_{\mathcal{X}}$, and family of distributions $\mathcal{D}_\alpha \in \Delta(\mathcal{X} \times \mathcal{Y})$ for $\alpha \in [0, .2]$ such that $\mathrm{dMC}_{\mathcal{C},\mathcal{D}_\alpha}(f) > .3$ if $\alpha > 0$, and $\mathrm{dMC}_{\mathcal{C},\mathcal{D}_0}(f) = 0$, but $\mathrm{TV}(\mathcal{D}_0, \mathcal{D}_\alpha) \in O(\alpha)$.*

Proof. Let $\mathcal{X} = \{x_1, x_2, x_3\}$ and $\mathcal{C} = \{S_1, S_2\}$, where $S_1 = \{x_1, x_2\}$, and $S_2 = \{x_2, x_3\}$. Let the marginal distribution $\mathcal{D}_{\mathbf{x}}$ be the uniform distribution. For a fixed $\alpha \in [0, 0.2]$, we define the ground truth distribution p^* as:

$$p^*(x_1) = 0.8,\ p^*(x_2) = 0.2,\ p^*(x_3) = 0.8 + \alpha.$$

Define $f : \mathcal{X} \to [0,1]$ by $f(x) = 0.5$ for all $x \in \mathcal{X}$.

There are two predictors (up to restriction by S_1) which are perfectly calibrated with respect to S_1. The first one being $p^*|_{S_1}$, and the second one being the predictor g such that $g(x_1) = g(x_2) = 0.5$. Similarly, there are two predictors (up to restriction by S_2) which are perfectly calibrated with respect to S_2: $p^*|_{S_2}$ and h where $h(x_2) = h(x_3) = 0.5 + \alpha/2$.

Case 1. If $\alpha > 0$, then $g(x_2) \neq h(x_2)$, thus p^* is the only predictor which is both calibrated with respect to S_1 when restricted to S_1 and calibrated with respect to S_2 when restricted to S_2. It follows that $\mathrm{dMC}_{\mathcal{C}}(f) = \ell_1(f, p^*) > C$ for some constant $C > 0.3$.

Case 2. If $\alpha = 0$, then we have that $f|_{S_1} = g$ and $f|_{S_2} = h$. In this case, there are two multicalibrated predictors, p^* and f, and therefore, $\mathrm{dMC}_{\mathcal{C}}(f) = 0$.

Finally, note that

$$\mathrm{TV}(\mathcal{D}_0, \mathcal{D}_\alpha) = \frac{1}{2}\left|\frac{1}{3}\mathop{\mathbb{P}}_{\mathcal{D}_0}[\mathbf{y} = 1|x_3] - \frac{1}{3}\mathop{\mathbb{P}}_{\mathcal{D}_\alpha}[\mathbf{y} = 1|x_3]\right| + \frac{1}{2}\left|\frac{1}{3}\mathop{\mathbb{P}}_{\mathcal{D}_0}[\mathbf{y} = 0|x_3] - \frac{1}{3}\mathop{\mathbb{P}}_{\mathcal{D}_\alpha}[\mathbf{y} = 0|x_3]\right| = \frac{\alpha}{3}.$$

The discontinuity which emerges due to this example is illustrated in Figure 2. ◀

The result above has immediate no-free-lunch implications. Let $\varepsilon \in (0, 0.3]$, and suppose one has a hypothesis test which, given a finite i.i.d. sequence $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)$ for $\mathbf{x}_i \sim \mathcal{D}$ and $\mathbf{y}_i \sim \mathrm{Ber}(p^*(\mathbf{x}_i))$ and unlimited oracle-query access to f, can determine whether $\mathrm{dMC}_{\mathcal{C}}(f) = 0$ or $\mathrm{dMC}_{\mathcal{C}}(f) \geq \varepsilon$ with probability at least $1 - \delta > 1/2$, for any $\mathcal{D}$ chosen from $\{\mathcal{D}_\alpha : \alpha \in [0, 0.2]\}$. By Proposition 9, this implies that for any $\alpha \in (0, 0.2]$, one can distinguish between $\mathcal{D}_0$ and $\mathcal{D}_\alpha$ with probability at least $1 - \delta$. By Le Cam, however, we have that

$$2\delta > 1 - \mathrm{TV}(\mathcal{D}_0^{\otimes m}, \mathcal{D}_\alpha^{\otimes m}) \geq (1 - O(\alpha))^m.$$

Since we can take any $\alpha > 0$, no such test exists. Crucial in this setup is the requirement that a multicalibration test must succeed with a finite number of samples dependent on the family of distributions from which $\mathcal{D}$ is chosen. Fundamentally, this barrier arises from discontinuities that occur within such a family.

We remark that a similar counterexample exists to Proposition 9 when we place additional natural requirements on $\mathcal{C}$ which at first glance may appear get around the counterexample. In particular, in Corollary 21 we will demonstrate an instance where: (1) $\mathcal{X} \setminus S \in \mathcal{C}$ of each $S \in \mathcal{C}$, (2) $\mathcal{X} \in \mathcal{C}$, and (3) $|\mathcal{X}|$ is arbitrarily large but Proposition 9 still holds.

Proposition 9 showcases that an arbitrarily small change α in the ground truth labeling function p^* can both (1) alter the size of the set $\mathrm{mcal}_{\mathcal{C}}(\mathcal{D})$ of $\mathcal{C}$-multicalibrated predictors for the distribution $\mathcal{D}$; and (2) add a new multicalibrated predictor in a totally different region of the prediction space $\mathcal{F}_{\mathcal{X}}$ (see Figure 2). This is unfortunate news, since it shows that even when the marginal $\mathcal{D}_{\mathbf{x}}$ is held fixed, there are *discontinuities* in dMC with respect to p^*. This discontinuity highlights another key desirable property of any "distance to" notion: when holding the marginal $\mathcal{D}_{\mathbf{x}}$ fixed, the metric should be *continuous* with respect to changes in the ground truth p^*. This is a core requirement for auditing to be possible, since it implies that we do not have to obtain the *exact* ground truth p^* to audit for the calibration notion. It is relatively straightforward to see that the original dCE definition of [4] satisfies this property.[7]

It will be useful to establish additional notation. Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. For a predictor $p \in \mathcal{F}_{\mathcal{X}}$, we let $\mathcal{D}_p$ denote the distribution induced by drawing $\mathbf{x} \sim \mathcal{D}_x$ and $\mathbf{y} \sim \mathrm{Ber}(p^*(\mathbf{x}))$, and define $\mathrm{dCE}_{\mathcal{D},p}(f) := \mathrm{dCE}_{\mathcal{D}_p}(f)$. We may now consider continuity notions with respect to different ground truths.

▶ **Lemma 10.** *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ and $f \in \mathcal{F}_{\mathcal{X}}$. Then $\mathrm{dCE}_{\mathcal{D}}(f)$ is 1-Lipschitz with respect to the ground truth p^*. That is, for $p_1, p_2 \in \mathcal{F}_{\mathcal{X}}$, we have*

$$|\,\mathrm{dCE}_{\mathcal{D},p_1}(f) - \mathrm{dCE}_{\mathcal{D},p_2}(f)| \leq \ell_1(p_1, p_2).$$

We leave the proof to the appendix in the full version [10].

[7] Indeed, this also follows from the fact that dCE is approximately auditable in the "prediction-access" model of [4].

2.2.2 Extension of Proposition 9 to Other Multi-group Notions

The works [20, 7] showcase the utility of weakening the full multicalibration definition by relaxing the conditioning on $f(x)$ in the definition of calibration. This is done by, *e.g.*, restricting the dependence to various degree polynomials, and combining it with other notions like global calibration. A natural question is whether we can *circumvent* the pathology exhibited in Proposition 9 by weakening the notion of multicalibration considered. In the appendix of the full version [10], we show that similar counter-examples exist for two weaker distance to multi-group notions: distance to *low-degree multicalibration* and distance to *calibrated multiaccuracy*. Both remain impossible to audit due to discontinuities with respect to the ground truth.

2.2.3 Lipschitzness of wdMC

It is not difficult to see that wdMC is also 1-Lipschitz in p^*. By definition, of wdMC, we have $\mathrm{wdMC}_{\mathcal{C}}(f) = \max_i \mathbb{P}[S_i] \cdot \mathrm{dCE}_{\mathcal{D}|_{S_i}}(f)$. The observation follows from that fact that $\mathrm{dCE}_{\mathcal{D}|_{S_i}}(f)$ is 1-Lipschitz and the maximum function is 1-Lipschitz for the supremum norm.

3 Continuized Distance to Multicalibration Error

The example in Proposition 9 demonstrates that there are circumstances where dMC is impossible to audit or even approximate. In Section 3.1, we introduce a variant of dMC, the *continuized* distance to multicalibration error, $\widetilde{\mathrm{dMC}}$). Carefully *smoothing* the dMC over a sequence of local neighborhoods, $\widetilde{\mathrm{dMC}}$ is 1-Lipschitz in the ground truth labeling function p^*, satisfying a key property for the information-theoretic tractability discussed in Section 2.2. We examine this notion over the course of Sections 3.2 and 3.3, via an equivalence to a new distance notion, the distance to *intersection* multicalibration error. Unfortunately, in Proposition 23, we show that auditing $\widetilde{\mathrm{dMC}}$ (and dIMC) remains *statistically* hard with only a number of samples from $\mathcal{D}$ which is polynomial in the domain size $n = |\mathcal{X}|$ and subgroup collection size $k = |\mathcal{C}|$. However, for the common scenario in practice with a *constant* number of groups k, we show that $\widetilde{\mathrm{dMC}}$ (and dIMC) are both statistically and computationally feasible to audit, and provide a recipe for doing so via access to a dCE auditing oracle.

We defer most proofs to the full version [10].

3.1 Smoothing Away Discontinuities of dMC

The reason that dMC is not auditable is that the set $\mathrm{mcal}_{\mathcal{C}}(\mathcal{D})$ does not continuously vary with p^*. In particular, recall the example given in Proposition 9 where p^* is parameterized by α. Whenever $\alpha > 0$, p^* ended up being the *only* multicalibrated predictor, and $\mathrm{dMC}_{\mathcal{C}}(f) > 0.3$. However, when we change α to be *exactly* zero, f suddenly also becomes a multicalibrated predictor, and $\mathrm{dMC}_{\mathcal{C}}(f)$ drops to zero by definition. This also demonstrates that even the cardinality of the set $\mathrm{mcal}_{\mathcal{C}}(\mathcal{D})$ is not constant with slightly changing p^* (see Figure 2).

At a high level, we argue that $\alpha = 0$ is a special case for the ground truth p^* that induces *agreement* between groups. This is because at $\alpha = 0$, p^* looks identical when conditioning on the groups S_1 and S_2. This in turns adds another predictor f to the multicalibrated predictor set $\mathrm{mcal}_{\mathcal{C}}(\mathcal{D})$, which drops the distance dMC from 0.3 to 0. We will eventually show that these discontinuities, for example at $\alpha = 0$, can only ever *lower* the dMC. Intuitively, this occurs because the $\mathrm{mcal}_{\mathcal{C}}(\mathcal{D})$ momentarily expands, which makes the ℓ_1 distance of any predictor to the set shrink.

Working with the fact that all discontinuities have their image below the local neighborhood of dMC, it is natural to introduce a *continuized* version of dMC, which we refer to as $\widetilde{\text{dMC}}$. In this continuized version $\widetilde{\text{dMC}}$, we remove all discontinuities of dMC by taking the *supremum* of $\text{dMC}_{\mathcal{C}}(f)$ over a sequence of neighborhoods around the ground truth p^*, effectively removing these lower discontinuities.

As above, we will need specify metrics with respect to varying ground truths and subgroup collections. Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. For a predictor $p \in \mathcal{F}_{\mathcal{X}}$, recall that $\mathcal{D}_p$ denotes the distribution induced by drawing $\mathbf{x} \sim \mathcal{D}_x$ and $\mathbf{y} \sim \text{Ber}(p^*(\mathbf{x}))$. We let $\text{dMC}_{\mathcal{C},p}(f) := \text{dMC}_{\mathcal{C},\mathcal{D}_p}(f)$. When p^* is clear from context, we omit the associated subscript. We are now ready to formally introduce the continuized metric.

▶ **Definition 11** (Continuized Distance to Multicalibration Error). *Let $B_\varepsilon(p)$ be the closed ball of ℓ_1-radius ε around the predictor $p \in [0,1]^{\mathcal{X}}$.*[8] *We define* continuized distance to multicalibration error *as follows:*

$$\widetilde{\text{dMC}}_{\mathcal{C},p^*}(f) := \lim_{\varepsilon \to 0^+} \sup_{p \in B_\varepsilon(p^*)} \text{dMC}_{\mathcal{C},p}(f).$$

There are a few important things to note about the definition of $\widetilde{\text{dMC}}$. First, notice that since $p^* \in B_\varepsilon(p^*)$ for all $\varepsilon > 0$, it is always the case that $\text{dMC}_{\mathcal{C},p^*}(f) \leq \widetilde{\text{dMC}}_{\mathcal{C},p^*}(f)$. Second, as a sanity check, we note that $\widetilde{\text{dMC}}$ is not simply measuring distance to the ground truth p^*. Rather, we wish for $\widetilde{\text{dMC}}$ to rule out "bad" multicalibrated predictors which induce discontinuities in the metric, and account for the remaining "good" multicalibrated predictors. The following proposition demonstrates that $\widetilde{\text{dMC}}$ is not pathological in this sense, measuring distances to multicalibrated predictors other than p^*.

▶ **Proposition 12.** *There exists distribution $\mathcal{D}$ with ground truth p^* and subgroup collection $\mathcal{C}$, and predictor $f \neq p^*$ such that $\ell_1(f, p^*) = \Omega(1)$ and $\widetilde{\text{dMC}}_{\mathcal{C},p^*}(f) = 0$.*

Finally, and most importantly, $\widetilde{\text{dMC}}$ seems impossible to compute a priori as its definition relies heavily on knowledge of p^*. Surprisingly, in Proposition 20, we show that this is actually not the case: computing $\widetilde{\text{dMC}}$ is equivalent to computing distance to the nearest predictor which is calibrated with respect to all *intersections* of groups in $\mathcal{C}$. This provides a path towards auditability of $\widetilde{\text{dMC}}$ in some common scenarios discussed further in Section 3.4.

We are now prepared to discuss the properties of $\widetilde{\text{dMC}}$. In Sections 2.1 and 2.2, we argued that there were two essential requirements for multicalibration error metric: local minima are global, and lipschitzness in the ground truth labeling function p^*. The main result in this section is that $\widetilde{\text{dMC}}$ indeed satisfies both.

▶ **Theorem 13.** *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, subgroup collection $\mathcal{C}$, and $f \in \mathcal{F}_{\mathcal{X}}$ be given. Assume that $\mathcal{C}$ covers the domain $\mathcal{X}$. Then, $\widetilde{\text{dMC}}_{\mathcal{C},p^*}(f)$ satisfies the following:*

1. *All local minima are global minima.*
2. *$\widetilde{\text{dMC}}_{\mathcal{C},p^*}(f)$ is 1-Lipschitz with respect to the ground truth p^*.*

The proof of Theorem 13 is presented in Section 3.3 and crucially relies on the equivalence between the continuized and *intersection* distance notions, which we examine in Section 3.2.

[8] We remark that the open ball can work as well, at the expense of longer proofs.

3.2 Distance to Intersection Multicalibration Error

As in the statement of Theorem 13, we will assume that $\mathcal{C}$ covers $\mathcal{X}$ throughout this section.

▶ **Definition 14** (Distance to Intersection Multicalibration Error). *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ with ground truth p^*, subgroup collection $\mathcal{C}$, and predictor $f \in \mathcal{F}_\mathcal{X}$ be given. Let $\mathcal{I}(\mathcal{C})$ be the intersection closure of the set $\mathcal{C}$. Explicitly,*

$$\mathcal{I}(\mathcal{C}) = \left\{ \bigcap_{S \in \mathcal{A}} S : \mathcal{A} \subseteq \mathcal{C} \ \wedge \ \mathcal{A} \neq \emptyset \right\}.$$

Then, we define the distance to intersection multicalibration error*,* dIMC*, as follows:*

$$\mathrm{dIMC}_{\mathcal{C},p^*}(f) := \mathrm{dMC}_{\mathcal{I}(\mathcal{C}),p^*}(f).$$

Before proceeding, we note that dIMC in fact satisfies the two desiderata of distance to multicalibration metrics.

▶ **Theorem 15.** *Let the distribution $\mathcal{D}$, subgroup collection $\mathcal{C}$, and predictor to audit f be given. Assume that $\mathcal{C}$ covers the domain $\mathcal{X}$. Then,* $\mathrm{dIMC}_{\mathcal{C},p^*}(f)$ *satisfies the following:*

1. *All local minima are global minima.*
2. $\mathrm{dIMC}_{\mathcal{C},p^*}(f)$ *is 1-Lipschitz with respect to the ground truth p^*.*

Since dIMC is a special case of dMC, the first criteria of Theorem 15 comes directly from dMC. The second comes from the nontrivial fact that multicalibration on $\mathcal{I}(\mathcal{C})$ is equivalent to multicalibration on a disjoint cover of $\mathcal{X}$, defined in Definition 18; then by Lemma 17.

As a first step to proving Theorem 13, we will show that surprisingly, dMC and dIMC are equivalent *almost everywhere* in the space of ground truth predictors (Proposition 16). That is, for most practical scenarios, we can treat them as equivalent quantities.

▶ **Proposition 16.** *Let a subgroup collection $\mathcal{C} = \{S_1, \ldots, S_k\}$ be given. Let $P \subset \mathcal{F}_\mathcal{X}$ be the set of ground truth predictors p^* such that* $\mathrm{dMC}_{\mathcal{C},p^*}(f) \neq \mathrm{dIMC}_{\mathcal{C},p^*}(f)$ *for some $f \in \mathcal{F}_\mathcal{X}$.*

Then P is contained in a set of Lebesgue-measure zero.

As results that follow will demonstrate, intuitively, the set of points where dIMC and dMC differ is exactly the "bad" ground truths p^* that block auditability in Proposition 9. Viewed another way, Proposition 16 says that these "bad" ground truths constitute a measure-zero subset in the space of all possible ground truths. In other words, most of the time the multicalibrated predictor $g \in \mathrm{mcal}_\mathcal{C}(\mathcal{D})$ for which $\mathrm{dMC}_\mathcal{C}(f) = \ell_1(f, g)$ is also multicalibrated on the *intersections* of subgroups in the collection $\mathcal{C}$.

Proposition 16 also implies that any properties of dMC will also, for most ground truths p^*, be true for dIMC. In particular, our next result shows that dMC is 1-Lipschitz when restricted to *disjoint covers* of $\mathcal{X}$. Combining this with a particular choice of disjoint cover then allows us to demonstrate that dIMC is 1-Lipschitz.

▶ **Lemma 17.** *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, subgroups $\mathcal{C}$, and predictor $f \in \mathcal{F}_\mathcal{X}$ be given. If $\mathcal{C}$ is a disjoint cover of $\mathcal{X}$, then* $\mathrm{dMC}_{\mathcal{C},p^*}$ *is 1-Lipschitz with respect to p^*. That is, for any two ground truth labeling functions $p_1, p_2 \in \mathcal{F}_\mathcal{X}$:*

$$|\mathrm{dMC}_{\mathcal{C},p_2}(f) - \mathrm{dMC}_{\mathcal{C},p_1}(f)| \leq \ell_1(p_1, p_2).$$

We still require two additional ingredients to show that dIMC is Lipschitz. First, we need to have a way of converting $\mathcal{I}(\mathcal{C})$ to a disjoint cover. Then, we need to show that dMC on the chosen disjoint cover preserves its value even though we have changed the subgroup collection. The following choice of partition is sufficient.

▶ **Definition 18.** *Let subgroup collection* $\mathcal{C} = \{S_1, \ldots, S_k\}$ *be given. Define the* partition generated by $\mathcal{C}$ *as:*

$$\mathcal{J}(\mathcal{C}) = \left\{ \left(\bigcap_{S \in \mathcal{A}} S \right) - \left(\bigcup_{S \in \mathcal{C} - \mathcal{A}} S \right) : \mathcal{A} \subseteq \mathcal{C} \ \wedge \ \mathcal{A} \neq \emptyset \right\}.$$

We know that $\mathcal{J}(\mathcal{C})$ is in fact a partition of $\mathcal{X}$ because for all $x \in \mathcal{X}$, there is a unique $\mathcal{A} \subseteq \mathcal{C}$ such that for all $i \in [k]$ we have $x \in S_i$ if and only if $S_i \in \mathcal{A}$. The following lemma shows that we can convert between $\mathcal{I}(\mathcal{C})$ and $\mathcal{J}(\mathcal{C})$ without losing any information on the set of multicalibrated predictors. This is key, since $\mathcal{I}(\mathcal{C})$ is *not* in general a partition.

▶ **Lemma 19.** *Let* $\mathcal{C}$ *be a disjoint cover of* $\mathcal{X}$*. For any distribution* $\mathcal{D}$ *with ground truth labeling distribution* p^**, we have that* $\text{mcal}_{\mathcal{I}(\mathcal{C})}(\mathcal{D}) = \text{mcal}_{\mathcal{J}(\mathcal{C})}(\mathcal{D})$*. Furthermore,* $\text{dIMC}_{\mathcal{C},p^*}(f) = \text{dMC}_{\mathcal{I}(\mathcal{C}),p^*}(f) = \text{dMC}_{\mathcal{J}(\mathcal{C}),p^*}(f)$.

We may now prove Theorem 15.

Proof of Theorem 15. To see property (1), notice that dIMC is simply defined as dMC for the particular subgroup collection $\mathcal{I}(\mathcal{C})$. Therefore, (1) is true via Proposition 7.

To prove Lipschitzness of dIMC (property (2) of Theorem 15), notice that by Lemma 19, we can define $\text{dIMC}_{\mathcal{C},p^*}(f)$ as $\text{dMC}_{\mathcal{J}(\mathcal{C}),p^*}(f)$. Since $\mathcal{J}(\mathcal{C})$ is a collection of disjoint sets which cover $\mathcal{X}$, by Lemma 17, we know dIMC is 1-Lipschitz in p^*. ◀

3.3 Completing the Proof of Theorem 13

In Theorem 15, we showed that dIMC satisfies both 1-Lipschitzness and local minima are global minima. The following proposition demonstrates that the two notions dIMC and $\widetilde{\text{dMC}}$ are equivalent, which therefore implies Theorem 13 as a corollary.

▶ **Proposition 20.** *For* $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ *and* $f, p^* \in [0,1]^{\mathcal{X}}$*, we have* $\text{dIMC}_{\mathcal{C},p^*}(f) = \widetilde{\text{dMC}}_{\mathcal{C},p^*}(f)$.

Proof of Theorem 13. This follows from the equivalence of $\widetilde{\text{dMC}}$ and dIMC in Proposition 20, and the properties of dIMC given in Theorem 15. ◀

Discussion and Implications

The equivalence between $\widetilde{\text{dMC}}$ and dIMC may have implications that go beyond our purposes. To start, because $\mathcal{J}(\mathcal{C})$ is a disjoint cover of $\mathcal{X}$, $\text{mcal}_{\mathcal{J}(\mathcal{C})}(\mathcal{D})$ is precisely the set of predictors f defined by $f|_{S_i} \in \text{cal}(\mathcal{D}|_{S_i})$ for each i. Applying total expectation, one obtains a decomposition of the continuized metric.

$$\begin{aligned}
\widetilde{\text{dMC}}_{\mathcal{C}}(f) &= \text{dMC}_{\mathcal{J}(\mathcal{C})}(f) \\
&= \min_{g \in \text{mcal}_{\mathcal{J}(\mathcal{C})}(\mathcal{D})} \ell_1(f, g) \\
&= \min_{g \in \text{mcal}_{\mathcal{J}(\mathcal{C})}(\mathcal{D})} \sum_{S \in \mathcal{J}(\mathcal{C})} \mathbb{P}[S] \cdot \ell_1|_S(f, g) \\
&= \sum_{S \in \mathcal{J}(\mathcal{C})} \mathbb{P}[S] \min_{g \in \text{cal}(\mathcal{D}|_S)} \ell_1|_S(f, g) \\
&= \sum_{S \in \mathcal{J}(\mathcal{C})} \mathbb{P}[S] \cdot \text{dCE}_{\mathcal{D}|_S}(f).
\end{aligned}$$

This property will be particularly useful for the purpose of auditing. This decomposition comes at a cost, however: $|\mathcal{J}(\mathcal{C})|$ may be exponential in $|\mathcal{C}|$.

We also can use the notion of intersection multicalibration to show a strengthened version of Proposition 9, which extends to settings where $\mathcal{C}$ is closed under set compliments, $\mathcal{X} \in \mathcal{C}$, and $|\mathcal{X}|$ is arbitrarily large. As in Proposition 9, all groups are a constant fraction of the domain.

▶ **Corollary 21.** *Let $N \in \mathbb{N}$. There exists a domain $\mathcal{X}$ such that $|\mathcal{X}| \geq N$, subgroups $\mathcal{C}$, predictor $f \in \mathcal{F}_{\mathcal{X}}$, and family of distributions $\mathcal{D}_\alpha \in \Delta(\mathcal{X} \times \mathcal{Y})$ for $\alpha \in [0, 0.1]$ such that $\mathrm{dMC}_{\mathcal{C},p^*}(f) = 0$, $\mathrm{dMC}_{\mathcal{C},p_\alpha}(f) \geq 0.2$, and $\mathrm{TV}(\mathcal{D}_0, \mathcal{D}_\alpha) \leq \alpha$.*

Corollary 21 demonstrates a discontinuity in dMC analogous to that of Proposition 9, confirming that the information-theoretic barrier to auditing this measure do not disappear asymptotically, even with additional structural assumptions on $\mathcal{C}$.

We close this section by relating all of the notions thus far introduced.

▶ **Proposition 22.** *For $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, $\mathcal{C} \subset 2^{\mathcal{X}}$, and $f \in \mathcal{F}_{\mathcal{X}}$, we have the following hierarchy:*

$$\mathrm{wdMC}_{\mathcal{D},\mathcal{C}}(f) \leq \mathrm{dMC}_{\mathcal{D},\mathcal{C}}(f) \leq \widetilde{\mathrm{dMC}}_{\mathcal{D},\mathcal{C}}(f) = \mathrm{dIMC}_{\mathcal{D},\mathcal{C}}(f) = \mathrm{dMC}_{\mathcal{D},\mathcal{J}(\mathcal{C})}\,.$$

3.4 Auditability of $\widetilde{\mathrm{dMC}}$ and dIMC

Though Theorems 13 and 15 show that testing $\widetilde{\mathrm{dMC}}$ may be tractable in restricted contexts, it is not hard to construct instances on which computing this notion – or even distinguishing between the case where $\widetilde{\mathrm{dMC}} = 0$ or $\widetilde{\mathrm{dMC}} \geq 0.5$ – requires $\mathsf{exp}(k)$ samples. Such instances arise due to the same mechanisms that make *intersectionality* guarantees in algorithmic fairness difficult to audit [27]. Consider, for example, a case in which $\mathcal{J}(\mathcal{C})$ is exponentially large in $k = |\mathcal{C}|$. On such an instance, auditing $\widetilde{\mathrm{dMC}}$ may require more knowledge of p^* than is statistically tractable. We demonstrate this fact in the following result.

▶ **Proposition 23.** *Let $k \geq 2$. There exists a domain $\mathcal{X}$, subgroup collection $\mathcal{C} = \{S_1, \ldots, S_k\}$, predictor $f \in \mathcal{F}_{\mathcal{X}}$, and distributions $\{\mathcal{D}_i\}_0^N \subset \Delta(\mathcal{X} \times \mathcal{Y})$ such that:*

1. *$\widetilde{\mathrm{dMC}}_{\mathcal{C},\mathcal{D}_0}(f) = 0$ and $\widetilde{\mathrm{dMC}}_{\mathcal{C},\mathcal{D}_i}(f) = 0.5$ for all $i \geq 1$;*
2. *For $m < 2^{k-2}$, we have $\mathrm{TV}(\mathcal{D}_0^{\otimes m}, \sum_{i=1}^{N} \frac{1}{N}\mathcal{D}_i^{\otimes m}) \in O(m^2/2^k)$.*

Proof. For chosen k, let $\mathcal{X} = \{0,1\}^{k-1}$ and $\mathcal{C} = \{S_i\}_1^{k-1} \cup \{\{0\}\}$ for $S_i := \{(x_1, \cdots, x_{k-1}) \in \mathcal{X} : x_i = 1\}$. Let $f(x) = 0.5$ everywhere. Let $\mathcal{D}_{\mathbf{x}}$ be uniform over $\mathcal{X}$. Define $\mathcal{D}_0 = \mathcal{D}_{\mathbf{x}} \otimes \mathrm{Ber}(0.5)$. Next, for each $T \subset \mathcal{X}$ of cardinality 2^{k-2}, take $p_T(x) = \mathbb{1}[x \in T]$ and define $\mathcal{D}_S$ by drawing $\mathbf{y} \sim \mathrm{Ber}(p_T(\mathbf{x}))$. We index these distributions $\mathcal{D}_i$ for $i \in [N]$ where $N := \binom{n}{n/2}$, and $n := 2^{k-1}$. One observes that $\mathcal{J}(\mathcal{C})$ is the set of all 2^{k-1} singletons, and hence, $\mathrm{mcal}_{\mathcal{C}}(\mathcal{D})$ consists of only the ground truth p^*. It follows that $\widetilde{\mathrm{dMC}}_{\mathcal{C},\mathcal{D}_0}(f) = \ell_1(f, 0.5) = 0$ and $\widetilde{\mathrm{dMC}}_{\mathcal{C},\mathcal{D}_i}(f) = 0.5$ for all $i \geq 1$.

Next, note that by birthday paradox, a sequence of samples $((\mathbf{x}_j, \mathbf{y}_j))_1^m$ admits repeats with probability at most $m(m-1)/2n$. Hence,

$$\begin{aligned}
\mathrm{TV}\left(\mathcal{D}_0^{\otimes m}, \sum_i \frac{1}{N}\mathcal{D}_i^{\otimes m}\right) &\leq \frac{m(m-1)}{2n} + \frac{1}{2} \sum_{\substack{((x_j,y_j))_1^m \\ x_j \text{ distinct}}} \left| \prod_{j=1}^{m} \mathbb{P}_{\mathcal{D}_0}[x_j, y_j] - \sum_{i=1}^{N} \frac{1}{N} \prod_{j=1}^{m} \mathbb{P}_{\mathcal{D}_i}[x_j, y_j] \right| \\
&= \frac{m(m-1)}{2n} + \frac{1}{2} \sum_{\substack{((x_j,y_j))_1^m \\ x_j \text{ distinct}}} \left(\prod_{j=1}^{m} \mathbb{P}_{\mathcal{D}_{\mathbf{x}}}[x_j] \right) \left| \prod_{j=1}^{m} \mathbb{P}_{\mathcal{D}_0}[y_j|x_j] - \sum_{i=1}^{N} \frac{1}{N} \prod_{j=1}^{m} \mathbb{P}_{\mathcal{D}_i}[y_j|x_j] \right|.
\end{aligned}$$

Note that for a sequence $((x_j, y_j))_1^m$ with all distinct x_j, there are exactly $\binom{n-m}{n/2-m}$ sets $S \subset \mathcal{X}$ of cardinality $n/2$ that contain all x_j. Hence, by definition of the $\mathcal{D}_i$,

$$\sum_{i=1}^{N} \frac{1}{N} \prod_{j=1}^{m} \mathop{\mathbb{P}}_{\mathcal{D}_i}[x_j, y_j] = \frac{\binom{n-m}{n/2-m}}{\binom{n}{n/2}} = \prod_{\ell=0}^{m-1} \frac{\frac{n}{2} - \ell}{n - \ell}.$$

Moreover,

$$\prod_{\ell=0}^{m-1} \frac{\frac{n}{2} - \ell}{n - \ell} = 2^{-m} \prod_{\ell=0}^{m-1} \left(\frac{1 - 2\ell/n}{1 - \ell/n} \right) \geq 2^{-m} \left(1 - \frac{2(m-1)}{n} \right)^m.$$

Fixing some $((x_j, y_j))_1^m$ such that each x_j is distinct, we have the following inequality:

$$\frac{1}{2} \sum_{\substack{((x_j,y_j))_1^m \\ x_j \text{ distinct}}} \left(\prod_{j=1}^{m} \mathop{\mathbb{P}}_{\mathcal{D}_\mathbf{x}}[x_j] \right) \left| \prod_{j=1}^{m} \mathop{\mathbb{P}}_{\mathcal{D}_0}[y_j|x_j] - \sum_{i=1}^{N} \frac{1}{N} \prod_{j=1}^{m} \mathop{\mathbb{P}}_{\mathcal{D}_i}[y_j|x_j] \right| \leq \left| \prod_{j=1}^{m} \mathop{\mathbb{P}}_{\mathcal{D}_0}[y_j|x_j] - \sum_{i=1}^{N} \frac{1}{N} \prod_{j=1}^{m} \mathop{\mathbb{P}}_{\mathcal{D}_i}[y_j|x_j] \right|.$$

Since $\mathop{\mathbb{P}}_{\mathcal{D}_0}[\mathbf{y} = 1|\mathbf{x}] = 0.5$,

$$\begin{aligned}
\mathrm{TV}\left(\mathcal{D}_0^{\otimes m}, \sum_i \frac{1}{N} \mathcal{D}_i^{\otimes m} \right) &\leq \frac{m(m-1)}{2n} + \left| 2^{-m} - \prod_{\ell=0}^{m-1} \frac{\frac{n}{2} - \ell}{n - \ell} \right| \\
&\leq \frac{m(m-1)}{2n} + 2^{-m} \left| 1 - \left(1 - \frac{2(m-1)}{n} \right)^m \right| \\
&\leq \frac{m(m-1)}{2n} + 2^{-m+1} \frac{m(m-1)}{n}. \qquad \text{(Bernoulli)}
\end{aligned}$$

To conclude, there is an absolute constant $C > 0$ such that $\mathrm{TV}(\mathcal{D}_0^{\otimes m}, \sum_i \frac{1}{N} \mathcal{D}_i^{\otimes m}) \leq Cm^2/2^k$. ◀

Proposition 23 exhibits a statistical barrier to auditing $\widetilde{\mathrm{dMC}}$ when limited assumptions are made on how the ground truth p^* is chosen. Suppose $\mathcal{D}$ is chosen with probability $1/2$ to be $\mathcal{D}_0$, and with probability $1/2$ is chosen uniformly from $\{\mathcal{D}_i\}_1^n$. The TV bound in Proposition 23 implies that auditing in this setting with constant advantage requires a number of samples exponential in k. We leave it to future work to examine reasonable constraints on p^* under which auditing can be done with efficient sample complexity (*e.g.* restriction to classes of bounded VC dimension).

▶ Remark 24. We remark that in the above proof, the difference between $\widetilde{\mathrm{dMC}}_{\mathcal{C},\mathcal{D}_0}(f) = 0$ and $\widetilde{\mathrm{dMC}}_{\mathcal{C},\mathcal{D}_i}(f) = 0.5$ corresponds to distinguishing between the cases where the constant predictor $f(x) = 0.5$ correctly estimates that the ground truth p^* is 0.5 on every x, or f is incorrect and p^* is always 0 or 1 on each x. In turn, these two cases model all uncertainty in f being *aleatoric* vs. *epistemic* respectively. Aleatoric uncertainty is *irreducible* uncertainty in the data. On the other-hand, epistemic uncertainty is – also known as "model uncertainty" or reducible uncertainty – captures deficiencies in the modeling capabilities of the predictor. As discussed in [1], predictors trained from only samples of the form $(\mathbf{x}, \mathbf{y})$ may not be able to distinguish between these two types of uncertainties. [1] therefore propose *higher order predictors*, which *are* able to distinguish between aleatoric and epistemic uncertainty since they are trained with "k-snapshots": samples of the form $(\mathbf{x}, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k)$ constructed by first sampling $\mathbf{x} \sim \mathcal{D}_\mathbf{x}$, then sampling $\mathbf{y}_i \overset{\text{i.i.d}}{\sim} p^*(x)$ from the ground truth labeling function p^*.

This represents, for example, asking multiple doctors to label a patient X-Ray as containing a broken bone or not. Such an approach may provide workarounds of roadblocks within multicalibration auditing. We leave exploration to future work.

Next, we exhibit a loose sample complexity upper bound for estimating $\widetilde{\mathrm{dMC}}$ when $\mathcal{J}(\mathcal{C})$ is not too large and none of its elements have too small of a probability mass. We start by estimating a weighted version of dCE for conditionals $\mathcal{D}|_S$. The following lemma uses tools developed in [4].

▶ **Lemma 25.** *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, $S \in 2^{\mathcal{X}} \setminus \emptyset$, $f \in \mathcal{F}_{\mathcal{X}}$, and $\varepsilon, \delta > 0$. There is an estimator $\hat{\mu}$ such that, given a sequence of $m = O(\varepsilon^{-2} log(1/\delta))$ samples $((f(\mathbf{x}_j), \mathbf{y}_j))_{j=1}^m$ for i.i.d. $(\mathbf{x}_j, \mathbf{y}_j) \sim \mathcal{D}|_S$, it holds with failure probability at most δ that*

$$\hat{\mu} - \varepsilon \leq \mathrm{dCE}_{\mathcal{D}|_S}(f) \leq 4\sqrt{\hat{\mu} + \varepsilon}. \tag{3}$$

▶ **Proposition 26.** *Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, $f \in \mathcal{F}_{\mathcal{X}}$, and $\mathcal{C} \subset 2^{\mathcal{X}} \setminus \emptyset$. Suppose $\mathcal{J}(\mathcal{C}) = \{S_1, \ldots, S_\ell\}$ and $\gamma = \min_{S \in \mathcal{J}(\mathcal{C})} \mathbb{P}[S] > 0$. Let $\varepsilon \in (0, \gamma]$ and $\delta > 0$. There is an estimator $\hat{\theta}$ such that, given a sequence of $m = O(\ell^4 \varepsilon^{-2} \gamma^{-1} \log(\ell/\delta))$ samples $(f(\mathbf{x}_j), \mathbf{y}_j, (\mathbb{1}[\mathbf{x}_j \in S_i])_{i=1}^{\ell})_{j=1}^m$ for i.i.d. $(\mathbf{x}_j, \mathbf{y}_j) \sim \mathcal{D}$, it holds with failure probability at most δ that*

$$\hat{\theta} - \varepsilon \leq \widetilde{\mathrm{dMC}}_{\mathcal{C}}(f) \leq 4\sqrt{\ell \hat{\theta}} + \sqrt{\varepsilon}.$$

4 Conclusions and Future Work

Our work broadly demonstrates barriers to auditing the distance to multicalibration notions. However, it also reveals that in certain restricted settings which may be useful to practitioners, distance to multicalibration *can* be audited (Section 3.4). Indeed, the fact that the continuized distance to multicalibration error $\widetilde{\mathrm{dMC}}$ introduced and developed throughout Section 3 is exactly measuring an *intersectional* [26] analogue of distance to multicalibration error is surprising and useful. However, this also means that for any setting with many subgroups in the collection $\mathcal{C}$, auditing the continuized distance is difficult.

The instance used to exhibit in-auditability of $\widetilde{\mathrm{dMC}}$ in Proposition 23 seems wholly due to the fact that the ground-truth has, in some sense, a "maximal" complexity. In particular, one can ask if "simpler" ground truths p^* give rise to positive auditability results. One can imagine restricting the allowable ground truth distributions to, for example, those with low VC dimension. This complexity restriction may demonstrate other settings where positive auditability results emerge.

Our preliminary foray into distance to *multiaccuracy*, available in Section 4 of the full version [10], also reveals interesting subtleties. In particular, we show that the worst-group bias quantity wdMA exhibits the following interesting property: Even if $\mathrm{wdMA}_{\mathcal{C}}(f) = \varepsilon$, the predictor f may need to change by some amount *exponential* in the collection size k in order to reach some $\tilde{f}$ with $\mathrm{wdMA}_{\mathcal{C}}(\tilde{f}) \leq \varepsilon/6$. This holds in spite of all local minima of wdMA being global minima; this suggests that there can be large "basins" with low slope near the minima of the loss surface of wdMA. Taken together, this demonstrates that dMA and wdMA are indeed different quantities, and suggests further motivation for finding a statistically and computationally efficient way to audit dMA.

More generally, auditing for dMA may require a better understanding the geometry of the space of multiaccurate predictors. It can be shown that the set $\mathrm{macc}(\mathcal{D})$ is affine subspace of $[0,1]^{\mathcal{X}}$, which lends some nice properties. For example, if f is perfectly multiaccurate with respect to the ground truth p^*, then p^* is perfectly multiaccurate with respect to the distribution given by a ground truth with f. The same cannot be said for multicalibration.

Lastly, extending to settings beyond finite $\mathcal{X}$ is especially interesting since the sets of calibrated or multicalibrated predictors would no longer be finite, an important assumption used throughout our proofs.

References

1 Gustaf Ahdritz, Aravind Gollakota, Parikshit Gopalan, Charlotte Peale, and Udi Wieder. Provable uncertainty decomposition via higher-order calibration. In *The Thirteenth International Conference on Learning Representations*, 2025.

2 Artem Bequé, Kristof Coussement, Ross Gayler, and Stefan Lessmann. Approaches for credit scorecard calibration: An empirical analysis. *Knowledge-Based Systems*, 134:213–227, 2017. doi:10.1016/J.KNOSYS.2017.07.034.

3 Beepul Bharti, Mary Versa Clemens-Sewall, Paul H Yi, and Jeremias Sulam. Multiaccuracy and multicalibration via proxy groups. In *Forty-second International Conference on Machine Learning*, 2025.

4 Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023.

5 Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009.

6 Jarosław Błasiok and Preetum Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*, 2024.

7 Sílvia Casacuberta, Parikshit Gopalan, Varun Kanade, and Omer Reingold. How global calibration strengthens multiaccuracy. *arXiv preprint arXiv:2504.15206*, 2025. doi:10.48550/arXiv.2504.15206.

8 Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. Collaborative prediction: Tractable information aggregation via agreement. *arXiv preprint arXiv:2504.06075*, 2025. doi:10.48550/arXiv.2504.06075.

9 Issa J Dahabreh, Jeffrey A Chan, Amy Earley, Denish Moorthy, Esther E Avendano, Thomas A Trikalinos, Ethan M Balk, and John B Wong. A review of validation and calibration methods for health care modeling and simulation. *Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment [Internet]*, 2017.

10 Nathan Derhake, Siddartha Devic, Dutch Hansen, Kuan Liu, and Vatsal Sharan. Auditability and the landscape of distance to multicalibration. *arXiv preprint arXiv:2509.16930*, 2025. doi:10.48550/arXiv.2509.16930.

11 Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms. In *International Conference on Machine Learning*, 2024.

12 Siddartha Devic, Aleksandra Korolova, David Kempe, and Vatsal Sharan. Stability and multigroup fairness in ranking with uncertain predictions. In *International Conference on Machine Learning*, 2024.

13 Cynthia Dwork, Chris Hays, Nicole Immorlica, Juan C Perdomo, and Pranay Tankala. From fairness to infinity: Outcome-indistinguishable (omni) prediction in evolving graphs. In *The Thirty Eighth Annual Conference on Learning Theory*, 2025.

14 Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6619–6627, 2023. doi:10.24963/IJCAI.2023/742.

15 Aravind Gollakota, Parikshit Gopalan, Aayush Karan, Charlotte Peale, and Udi Wieder. When does a predictor know its own loss? In Mark Bun, editor, *6th Symposium on Foundations of Responsible Computing, FORC 2025, June 4-6, 2025, Stanford University, CA, USA*, volume 329 of *LIPIcs*, pages 22:1–22:22, 2025. doi:10.4230/LIPIcs.FORC.2025.22.

16 Aravind Gollakota, Parikshit Gopalan, Adam Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

17 Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA*, 2023.

18 Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *Innovations in Theoretical Computer Science*, 2022.

19 Parikshit Gopalan, Michael P. Kim, and Omer Reingold. Characterizing notions of omniprediction via multicalibration. In *Advances in Neural Information Processing Systems*, 2023.

20 Parikshit Gopalan, Michael P Kim, Mihir A Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022. URL: https://proceedings.mlr.press/v178/gopalan22a.html.

21 Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. URL: http://proceedings.mlr.press/v70/guo17a.html.

22 Ido Guy, Daniel Haimovich, Fridolin Linder, Nastaran Okati, Lorenzo Perini, Niek Tax, and Mark Tygert. Measuring multi-calibration. *arXiv preprint arXiv:2506.11251*, 2025. doi:10.48550/arXiv.2506.11251.

23 Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

24 Dutch Hansen, Siddartha Devic, Preetum Nakkiran, and Vatsal Sharan. When is multicalibration post-processing necessary? In *Advances in Neural Information Processing Systems*, 2024.

25 Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

26 Johannes Himmelreich, Arbie Hsu, Kristian Lum, and Ellen Veomett. The intersectionality problem for algorithmic fairness. *arXiv preprint arXiv:2411.02569*, 2024. doi:10.48550/arXiv.2411.02569.

27 Johannes Himmelreich, Arbie Hsu, Ellen Veomett, and Kristian Lum. The intersectionality problem for algorithmic fairness. In *Workshop on Algorithmic Fairness Through the Lens of Metrics and Evaluation*, pages 68–95. PMLR, 2025.

28 Lunjia Hu, Arun Jambulapati, Kevin Tian, and Chutong Yang. Testing calibration in nearly-linear time. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL: http://papers.nips.cc/paper_files/paper/2024/hash/d20e3c35bedb17a9f6f01fc434a30fa3-Abstract-Conference.html.

29 Lunjia Hu, Charlotte Peale, and Judy Hanwen Shen. Multigroup robustness. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

30 Hongyi Henry Jin, Zijun Ding, Dung Daniel Ngo, and Zhiwei Steven Wu. Discretization-free multicalibration through loss minimization over tree ensembles. *arXiv preprint arXiv:2505.17435*, 2025. doi:10.48550/arXiv.2505.17435.

31 Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023.

32 Sham M Kakade and Dean P Foster. Deterministic calibration and nash equilibrium. In *International Conference on Computational Learning Theory*, pages 33–48. Springer, 2004. doi:10.1007/978-3-540-27819-1_3.

33 Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019. doi:10.1145/3306618.3314287.

34 Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631. PMLR, 20–22 April 2017. URL: http://proceedings.mlr.press/v54/kull17a.html.

35 Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, pages 68–85. Springer, 2015. doi:10.1007/978-3-319-23528-8_5.

36 Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR, 2018.

37 Terrance Liu and Steven Wu. Multi-group uncertainty quantification for long-form text generation. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

38 Carol Xuan Long, Wael Alghamdi, Alexander Glynn, Yixuan Wu, and Flavio P Calmon. Kernel multiaccuracy. In *6th Symposium on Foundations of Responsible Computing (FORC 2025)*, pages 7:1–7:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025. doi:10.4230/LIPIcs.FORC.2025.7.

39 Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: Trainable kernel calibration metrics. *Advances in Neural Information Processing Systems*, 36:25910–25928, 2023.

40 Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

41 Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, pages 625–632, 2005. doi:10.1145/1102351.1102430.

42 Princewill Okoroafor, Robert Kleinberg, and Michael P Kim. Near-optimal algorithms for omniprediction. In *66th Annual Symposium on Foundations of Computer Science*, 2025.

43 Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. In *Advances in Neural Information Processing Systems*, volume 33, pages 13331–13340, 2020.

44 Judy Hanwen Shen, Ellen Vitercik, and Anders Wikum. Algorithms with calibrated machine learning predictions. In *Forty-second International Conference on Machine Learning*, 2025.

45 Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023. doi:10.1007/S10994-023-06336-7.

46 USA. *Equal Credit Opportunity Act (15 U.S.C. 1691 et seq.)*. Federal Trade Commission, 1976.

47 Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *arXiv preprint arXiv:2308.01222*, 2023. doi:10.48550/arXiv.2308.01222.

48 Jiayun Wu, Jiashuo Liu, Peng Cui, and Steven Z Wu. Bridging multicalibration and out-of-distribution generalization beyond covariate shift. *Advances in Neural Information Processing Systems*, 37:73036–73078, 2024.