The Statistical Scope of Multicalibration

Georgy Noarov [1] **Aaron Roth** [1]

Abstract

We make a connection between multicalibration and property elicitation and show that (under mild technical conditions) it is possible to produce a multicalibrated predictor for a continuous scalar property Γ if and only if Γ is *elicitable*. On the negative side, we show that for non-elicitable continuous properties there exist simple data distributions on which even the true distributional predictor is not calibrated. On the positive side, for elicitable Γ, we give simple canonical algorithms for the batch and the online adversarial setting, that learn a Γ-multicalibrated predictor. This generalizes past work on multicalibrated means and quantiles, and in fact strengthens existing online quantile multicalibration results. To further counter-weigh our negative result, we show that if a property Γ^1 is not elicitable by itself, but *is* elicitable *conditionally* on another elicitable property Γ^0, then there is a canonical algorithm that *jointly* multicalibrates Γ^1 and Γ^0; this generalizes past work on mean-moment multicalibration. Finally, as applications of our theory, we provide novel algorithmic and impossibility results for fair (multicalibrated) risk assessment.

1. Introduction

Consider a distribution D over a labeled data domain $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ of examples with observable features $x \in \mathcal{X}$ and labels $y \in \mathbb{R}$. A predictor $f : \mathcal{X} \to \mathbb{R}$ is *(mean) calibrated* if, informally, it correctly estimates the *mean label value* even conditional on its own predictions: i.e., $\mathbb{E}_{(x,y)\sim D}[y|f(x) = v] = v$ for all predictions v. Calibration is a desirable property, but a weak one, since it only refers to the *average* value of the label, averaged over all examples such that $f(x) = v$; it might be, for example, that there is a structured subset of examples $G \subset \mathcal{X}$ such that f systematically under-estimates label means for examples $x \in G$ — such a predictor can still be calibrated if it compensates by over-estimating the mean labels for $x \notin G$.

Multicalibration was introduced by Hébert-Johnson et al. (2018) to strengthen the notion of calibration. A multicalibrated predictor is parameterized by a collection of groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$, and is calibrated not just overall, but also conditional on membership in G for all groups $G \in \mathcal{G}$. That is, for all v, G, we must have: $\mathbb{E}_{(x,y)\sim D}[y|f(x) = v, x \in G] = v$.

Multicalibration was generalized from means to moments by Jung et al. (2021). By way of an explicit counterexample, Jung et al. (2021) showed that the variance (and other higher moments) cannot be multicalibrated by themselves but *can* be multicalibrated *jointly* with the mean, i.e., as part of a (mean, moment) pair. Later, Gupta et al. (2022) and Jung et al. (2023) showed how to obtain a *quantile* analogue of multicalibration, which requires that for any target coverage level $\tau \in [0, 1]$, for any v in the range of a predictor f and for any $G \in \mathcal{G}$: $\mathbb{E}[\mathbb{1}[y \leq f(x)]|f(x) = v, x \in G] = \tau$.

Thus, by now we have efficient batch (Hébert-Johnson et al., 2018; Jung et al., 2021; 2023) and online (Gupta et al., 2022; Bastani et al., 2022) multicalibration algorithms for several natural distributional properties (means and quantiles), an impossibility result for the variance and higher moments, and a result showing how to multicalibrate means and moments together despite moments not being multicalibratable on their own (Jung et al., 2021). But are these one-off results, or is there a more general theory of multicalibration for *distributional properties* — i.e., arbitrary functionals $\Gamma : \mathcal{P} \to \mathbb{R}$ mapping any distribution $P \in \mathcal{P}$ to a scalar statistic $\Gamma(P)$? To study this question, it is natural to define (cf. Dwork et al. (2022)) that a predictor f is $(\mathcal{G}, \Gamma)$-*multicalibrated* for property Γ if for all groups $G \in \mathcal{G}$ and values γ in f's range, it holds that: $\Gamma(D|(f(x) = \gamma, x \in G)) = \gamma$, where $D|(f(x) = \gamma, x \in G)$ is the data distribution conditioned on the event $\{x : f(x) = \gamma, x \in G\}$. (When $\mathcal{G} = \{\mathcal{X}\}$, we would simply say that f is Γ-calibrated.)

Our motivation In developing our theory of property multicalibration, we are guided by trying to answer several questions of special significance:

[1]Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, PA, USA. Correspondence to: Georgy Noarov <gnoarov@seas.upenn.edu>, Aaron Roth <aaroth@cis.upenn.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

(1) *Which* distributional properties of interest *are* possible to multicalibrate, and which ones are not?

(2) For those properties that *are* multicalibratable in the batch setting, do we always also have a solution in the online adversarial setting, or is there an online-offline separation?

(3) Both in the batch and in the online setting, can one formulate a natural *canonical* algorithm with simple and generic performance guarantees, which takes in a description (in some simple format) of any multicalibratable property of interest and outputs a multicalibrated predictor for it?

(4) For practically important properties that are not multicalibratable *per se*, are there any reasonably general techniques for us to achieve some modified notion of multicalibration? (Cf. the case of the variance, which becomes multicalibratable when paired with the mean as per Jung et al. (2021).)

1.1. Our Results

We give an (almost) complete answer to all these questions by connecting property multicalibration to the well-studied theory of *property elicitation*.

The modern formulation of property elicitation theory is due to Lambert et al. (2008), but it has been extensively developed in both earlier and later works. In a nutshell, properties are called *elicitable* if their value on any data distribution can always be directly learned as the minimizer of some loss function over the dataset. For example, means and quantiles are elicitable as they can be solved for, respectively, via least squares and quantile regression. But, for instance, variance is not elicitable. An equivalent (subject to mild assumptions) notion is that of *identifiable* properties: an identification function (typically a first-order condition on the property's "loss function") tells us if we over- or under-estimated the property value, in expectation over the dataset. For example, an *expected* identification function for a mean predictor f_m is simply $\mathbb{E}_{(x,y)}[f_m(x) - y]$, and an *expected* identification function for a τ-quantile predictor f_τ is $\Pr_{(x,y)}[y \leq f_\tau(x)] - \tau$, the average overcoverage of f_τ. We give the following collection of results.

A Feasibility Criterion: Γ-Multicalibration Is Possible If and Only If Γ Is Elicitable. We provide a very general *if-and-only-if* characterization that categorizes various distributional properties of interest as possible or not possible to (multi)calibrate. We show (under mild assumptions) that a property Γ is *sensible for calibration* (see Definition 3.2) if and only if Γ is elicitable, and if and only if Γ is identifiable. See Theorem 3.7 in Section 3. A crucial tool we use is a central result of Steinwart et al. (2014): (under mild conditions) a property Γ is elicitable $\iff$ it is identifiable $\iff$ its level sets are convex. Our key insight, which allows us to invoke this result, is a tight relationship between sensibility for Γ-calibration and the convexity of the level sets of Γ.

Canonical Batch and Online Algorithms. We identify two "canonical" Γ-multicalibration algorithms for bounded elicitable properties Γ — the batch Algorithm 1 and the online adversarial Algorithm 5 — and prove convergence guarantees for them. See Sections 4 and F, respectively. Our batch Algorithm 1 naturally extends the known methods for means and quantiles (Hébert-Johnson et al., 2018; Jung et al., 2023). Our online Algorithm 5 generalizes and improves existing online algorithms for means and quantiles (Gupta et al., 2022; Bastani et al., 2022; Lee et al., 2022).

***Joint* Multicalibration for Conditionally Elicitable Properties.** We show that if a property Γ^0 is elicitable, and Γ^1 is *conditionally* elicitable given Γ^0 (meaning, informally, that conditional on knowing the exact value of Γ^0, there *is* a regression procedure to learn Γ^1), then (under technical conditions) the pair (Γ^0, Γ^1) is *jointly* multicalibratable using a canonical Algorithm 2. This generalizes (mean, moment) multicalibration of Jung et al. (2021). See Section 5.

Positive and Negative Results on Fair Risk Assessment. Previously, nothing was known about multicalibrating any of the large collection of *risk measures* beyond quantiles and variances. In Section 6, we begin to fill this gap by applying our theory to derive results about a host of risk measures of central significance in financial risk assessment. We show a general negative result that the large family of *distortion risk measures* are not multicalibratable, except for means and quantiles (and two other technical quantile variants). On the positive side, we establish that so-called *Bayes risks* are multicalibratable jointly with the elicitable property whose risk they measure. This is exemplified by *Conditional Value at Risk* (CVaR), also known as *Expected Shortfall* (ES) — a risk assessment measure of central theoretical and practical significance — which, as we show, is not multicalibratable on its own but is multicalibratable jointly with quantiles.

We discuss further related work in Appendix A.

2. Preliminaries

We study prediction problems over labeled datapoints in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a space of *feature vectors* and $\mathcal{Y} \subseteq \mathbb{R}$ is a space of *labels*. We study both batch (offline) and sequential (online) prediction problems. The online setting will be discussed in Appendix F, and we defer sequential prediction definitions and preliminaries to that section.

In the *batch (offline)* setting, there is a distribution $D \in \Delta\mathcal{Z}$ over labeled *examples* $(x, y) \in \mathcal{Z}$. Such a distribution induces a distribution X over features and Y over labels. Keeping D implicit, we let $Y_x := (Y|\{X = x\})$ be the conditional label distribution given a feature vector $x \in \mathcal{X}$. Similarly, for any subset $G \subseteq \mathcal{X}$, we write $Y_G := (Y|\{x \in G\})$ for the conditional label distribution given $x \in G$. A *predictor* or *model* is a real-valued mapping $f : \mathcal{X} \to \mathbb{R}$.

2.1. Distributional Properties

A one-dimensional *(distributional) property* is a functional $\Gamma : \mathcal{P} \to \mathbb{R}$, where $\mathcal{P}$ is some space of probability distributions. We write $\text{Range}_\Gamma \subseteq \mathbb{R}$ to denote the range of Γ. For our algorithmic Γ-multicalibration results in this paper, we will assume (without further mention) that the property Γ is bounded, and w.l.o.g. rescale Γ so that $\text{Range}_\Gamma \subseteq [0, 1]$.

Examples of properties include the mean, the median, a τ-quantile, and the variance of a distribution (for further examples see Section 6). What all of these notions have in common is that each of them puts a single real number in correspondence with a given distribution.

Formally Defining $\mathcal{P}$ In this paper, the basic assumption we place on any distribution space $\mathcal{P}$ is that $\mathcal{P}$ is a *convex* subset of some vector space, so that taking convex combinations over distributions in $\mathcal{P}$ is a well-defined operation.

If all distributions $P \in \mathcal{P}$ are defined over the label space $\mathcal{Y}$, we can impose extra structure on $\mathcal{P}$ in one of two ways, depending on whether $\mathcal{Y}$ is a finite set or not. If $|\mathcal{Y}| = d < \infty$, all $P \in \mathcal{P}$ can be viewed as elements of the d-dimensional simplex $\Delta(d) \subset \mathbb{R}^d$, so we view $\mathcal{P}$ as a subset of $\Delta(d)$ equipped with the Euclidean norm. If $\mathcal{Y}$ is infinite, we assume that $\mathcal{P} \subseteq W_{\text{TV}}$, where W_{TV} is a *Banach space* (i.e. a complete normed vector space; see Diestel & Uhl (1977) for an introduction to Banach spaces) of probability distributions over $\mathcal{Y}$ with almost everywhere bounded densities, equipped with the *TV norm* $||\cdot||_{\text{TV}}$. In particular, W_{TV} is a metric space where the distance between any $P_1, P_2 \in \mathcal{P}$ is the *total variation distance* $||P_1 - P_2||_{\text{TV}}$.

An assumption we will often place on a property Γ is continuity: that is, Γ cannot take drastically different values on very similar distributions. Formally, a *continuous property* is a functional $\Gamma : \mathcal{P} \to \mathbb{R}$ that is continuous relative to the metric topology on $\mathcal{P}$ and the standard topology on $\mathbb{R}$. (A great many properties, including means, variances, quantiles, entropies, etc. are indeed continuous.)

Batch Property Prediction In our batch setting, we consider datasets over $\mathcal{X} \times \mathcal{Y}$, and are interested in training predictors $f_\Gamma : \mathcal{X} \to \mathbb{R}$ for various properties Γ of the conditional label distributions $Y_x \in \Delta\mathcal{Y}$. Informally, a good predictor f_Γ would satisfy $f_\Gamma(x) \approx \Gamma(Y_x)$ for every $x \in \mathcal{X}$.

Since we study properties $\Gamma : \mathcal{P} \to \mathbb{R}$ of distributions over the dataset's label space $\mathcal{Y}$, we restrict attention to those dataset distributions $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ whose induced label distributions Y_x belong to $\mathcal{P}$ for all $x \in \mathcal{X}$, so that the property Γ is at least well-defined on the label distributions Y_x of all $x \in \mathcal{X}$. We formalize this as follows.

Definition 2.1 ($\mathcal{P}$-Compatible Dataset Distribution)**.** Given a family of distributions $\mathcal{P} \subseteq \Delta\mathcal{Y}$, we call a dataset distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ *$\mathcal{P}$-compatible* if $Y_x \in \mathcal{P}$ for all $x \in \mathcal{X}$: i.e. the induced label distribution given any x belongs to $\mathcal{P}$.

2.2. Property Elicitation and Identification

We are now ready to formally define three related concepts that are the subject of study in the property elicitation literature. These concepts are: elicitability, identifiability, and level set convexity (CxLS) of distributional properties.

Elicitability Simply put, a property defined on a family of distributions $\mathcal{P}$ is called *elicitable* if its value on any distribution $P \in \mathcal{P}$ can be obtained by minimizing some loss function in expectation over samples from P — or, said in the language of statistical learning, by solving a regression problem. As is customary in the elicitation literature, we refer to such loss functions as *scoring functions*: mathematically, a scoring function is just a function $S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$.

Definition 2.2 (Strictly Consistent Scoring Function)**.** Fix a space of probability distributions $\mathcal{P}$. A scoring function $S : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is *strictly $\mathcal{P}$-consistent* for property $\Gamma : \mathcal{P} \to \mathbb{R}$ if: $\Gamma(P) = \text{argmin}_{\gamma \in \mathbb{R}} \mathbb{E}_{y \sim P}[S(\gamma, y)]$ for all $P \in \mathcal{P}$. We also say that S *elicits* Γ.

For brevity, we denote: $S(\gamma, P) = \mathbb{E}_{\gamma \sim P}[S(\gamma, y)]$. S is called *$\mathcal{P}$-order sensitive* for Γ if for $P \in \mathcal{P}$ and $\gamma_1, \gamma_2 \in \mathbb{R}$, $|\gamma_1 - \Gamma(P)| < |\gamma_2 - \Gamma(P)|$ implies $S(\gamma_1, P) < S(\gamma_2, P)$.

Definition 2.3 (Elicitable Property)**.** Fix a space of probability distributions $\mathcal{P}$. A property $\Gamma : \mathcal{P} \to \mathbb{R}$ is said to be *elicitable* if it has a strictly $\mathcal{P}$-consistent scoring function.

As a basic example of the above definitions, the scoring function defined as $S(\gamma, y) = (\gamma - y)^2$ elicits distributional means; thus, means are an elicitable property.

Convexity of Level Sets (CxLS) A simple but deep *necessary* condition for elicitability due to Osband (1985) is that elicitable properties must have *convex level sets* (also referred to as *CxLS*). This will be key to our characterization of sensibility for calibration via elicitability.

Fact 1 (Osband (1985)). Let $\mathcal{P}$ be a convex space of probability distributions, and $\Gamma : \mathcal{P} \to \mathbb{R}$ be an elicitable property. Then for all $\gamma \in \text{Range}_\Gamma$, the *level set* $\{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ is *convex*: that is, for any P_1, P_2 with $\Gamma(P_1) = \Gamma(P_2) = \gamma$, it holds that $\Gamma(\lambda P_1 + (1 - \lambda)P_2) = \gamma$ for all $\lambda \in [0, 1]$.

Identifiability A concept related to the above two is *identifiability* (Osband, 1985; Gneiting, 2011). It requires that a property have a so-called *identification*, or *id*, *function*:

Definition 2.4 (Identification Function)**.** A function $V : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is a *$\mathcal{P}$-identification* (or *id*) *function* for property $\Gamma : \mathcal{P} \to \mathbb{R}$ if for $P \in \mathcal{P}$, $\mathbb{E}_{y \sim P}[V(\gamma, y)] = 0 \Leftrightarrow \Gamma(P) = \gamma$.

For brevity, we denote: $V(\gamma, P) = \mathbb{E}_{\gamma \sim P}[V(\gamma, y)]$. An id function V for Γ is *oriented* if $V(\gamma, P) > 0 \Leftrightarrow \gamma > \Gamma(P)$.

In other words, identifiability makes it possible to compute the value of a property by setting to zero its expected id function over the distribution. For oriented identification functions, we further have that over- (resp. under-)estimating the property value leads to positive (resp. negative) expected identification function values.

Definition 2.5 (Identifiable Property)**.** Property $\Gamma : \mathcal{P} \to \mathbb{R}$ is *identifiable* if there is a $\mathcal{P}$-identification function for Γ.

For some intuition on how this relates to elicitability, note that if Γ has a differentiable convex scoring function, then its derivative (in the first argument) is an id function for Γ.

Connecting Elicitability, Identifiability and CxLS Under mild assumptions, elicitability, identifiability, and CxLS are in fact equivalent for continuous properties Γ, as shown by Steinwart et al. (2014). While Fact 1 shows, under no assumptions other than $\mathcal{P}$ being convex, that CxLS is a *necessary* condition for elicitability, the characterization of Steinwart et al. (2014) demonstrates that it is also *sufficient*, subject to some further assumptions. Formally, this characterization holds for $\mathcal{P}$ defined in one of two ways:

Definition 2.6. Let $\mathcal{P}_0$ be the subspace of the Banach space W_{TV} containing all probability distributions whose densities are upper-bounded almost everywhere. Let $\mathcal{P}_{>0}$ be the subspace of $\mathcal{P}_0$ containing all distributions $P \in \mathcal{P}_0$ whose densities are lower-bounded everywhere by some $\epsilon_P > 0$.

Theorem 2.7 (Steinwart et al. (2014))**.** *Consider a space of probability distributions $\mathcal{P} \in \{\mathcal{P}_0, \mathcal{P}_{>0}\}$. Let $\Gamma : \mathcal{P} \to \mathbb{R}$ be any continuous, strictly locally non-constant*[1] *property. Then the following statements are equivalent:*

1. *Γ is elicitable.*
2. *Γ has a bounded non-negative order-sensitive scoring function S.*
3. *Γ is identifiable and has a bounded, oriented identification function V.*
4. *Γ has convex level sets (CxLS): for any γ in the range of Γ, $\{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ is convex.*

2.3. (Multi)calibration for Property Predictors

We now give general definitions of calibration and multicalibration for (batch) predictors of any property Γ, extending the notions of mean and quantile calibration of Hébert-Johnson et al. (2018) and Jung et al. (2023). A variant of our definitions first appeared in Dwork et al. (2022) under the name *calibration consistency under mixtures*.

Fix any dataset over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, given by its data distribution $D \in \Delta\mathcal{Z}$. Suppose that, given any features $x \in \mathcal{X}$, we want to predict the value of property Γ on Y_x, the label distribution conditional on x. For this, we procure a Γ-*predictor* $f : \mathcal{X} \to \mathbb{R}$. We will call this predictor Γ-*calibrated* if for all $\gamma \in \text{Range}_f$, the conditional label distribution *given* the prediction $f(x) = \gamma$ indeed has property value γ.

Definition 2.8 (Calibrated Predictor for Property Γ)**.** A Γ-predictor $f : \mathcal{X} \to \mathbb{R}$ is Γ-*calibrated* on dataset distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ if for every $\gamma \in \text{Range}_f$: $\Gamma(Y_{f,\gamma}) = \gamma$, where $Y_{f,\gamma} = Y_{\{x : f(x) = \gamma\}}$ is the conditional label distribution induced by D conditional on $f(x) = \gamma$.

Now, we extend this definition to that of *multicalibration*, which offers calibration guarantees for arbitrary collections of subsets ('groups') of the feature space $\mathcal{X}$.

Definition 2.9 ($(\mathcal{G}, \Gamma)$-Multicalibrated Γ-Predictor)**.** Fix a collection of groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$. A Γ-predictor $f : \mathcal{X} \to \mathbb{R}$ is $(\mathcal{G}, \Gamma)$-*multicalibrated* on dataset distribution $D \in \Delta(\mathcal{X} \times \mathcal{Y})$ if for every $\gamma \in \text{Range}_f$ and $G \in \mathcal{G}$: $\Gamma(Y_{f,\gamma,G}) = \gamma$, where $Y_{f,\gamma,G} = Y_{\{x : f(x) = \gamma, x \in G\}}$ is the conditional label distribution induced by D given $f(x) = \gamma$ and $x \in G$. In other words, a $(\mathcal{G}, \Gamma)$-multicalibrated predictor f satisfies that, conditional on *both* the prediction being $f(x) = \gamma$ and on x being a member of group G, the conditional label distribution indeed has property value γ.

We will later need to work with a definition of *approximate* multicalibration for predictors with *finite range*. We adopt an ℓ_2-notion of calibration error, generalizing approximate quantile multicalibration as defined in Jung et al. (2023).

Definition 2.10 (α-Approximately $(\mathcal{G}, \Gamma)$-Multicalibrated Γ-Predictor)**.** Fix a distribution $D \in \Delta\mathcal{Z}$ and a collection of groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$. For each $G \in \mathcal{G}$, let $\mu(G) = \Pr_{(x,y) \sim D}[x \in G]$ be the probability mass on group G. A finite-range predictor $f : \mathcal{X} \to \text{Range}_f$ is α-*approximately* $(\mathcal{G}, \Gamma)$-*multicalibrated* on D if for all $G \in \mathcal{G}$:

$$\sum_{\gamma \in \text{Range}_f} \Pr_{(x,y) \sim D}[f(x) = \gamma | x \in G] \left(\gamma - \Gamma(Y_{f,\gamma,G})\right)^2 \leq \frac{\alpha}{\mu(G)}.$$

Note that 0-approximate $(\mathcal{G}, \Gamma)$-multicalibration is equivalent to the (exact) Definition 2.9 of $(\mathcal{G}, \Gamma)$-multicalibration.

3. Sensibility for Calibration and Elicitability

We are now ready to make a connection between property elicitation and (multi)calibration. First we define the notion of a property Γ being *sensible* for calibration.

Definition 3.1 (True Distributional Predictor for a Property)**.** Fix a distributional property $\Gamma : \mathcal{P} \to \mathbb{R}$ and a $\mathcal{P}$-compatible

[1] 'Strictly locally non-constant' is a (weak) requirement that for every P in the interior of $\mathcal{P}$ with $\Gamma(P) = \gamma$, and any ϵ-neighborhood U_ϵ of P in the metric topology on $\mathcal{P}$, there are distributions $P', P'' \in U_\epsilon$ such that $\Gamma(P'') < \gamma < \Gamma(P')$.

dataset distribution D. The *true distributional predictor* f_Γ^D for Γ on D is defined as $f_\Gamma^D(x) = \Gamma(Y_x)$ for $x \in \mathcal{X}$ — i.e. the predictor that for every $x \in \mathcal{X}$ gives the correct value of property Γ on the conditional label distribution given x.

Definition 3.2 (Property Sensible for Calibration)**.** Fix a property $\Gamma : \mathcal{P} \to \mathbb{R}$, and a collection $\mathcal{D}$ of $\mathcal{P}$-compatible dataset distributions. We say that Γ is *sensible for calibration over* $\mathcal{D}$ if the true distributional predictor f_Γ^D is Γ-calibrated on D for all $D \in \mathcal{D}$.

A key motivation for multicalibration (elaborated on in Dwork et al. (2021)) is that we want to produce a predictor f that is *indistinguishable* from f_Γ^D with respect to a class of *calibration tests* parameterized by $\mathcal{G}$—which only makes sense if Γ is sensible for calibration. In general, for properties that are not sensible for calibration, there need not exist calibrated predictors at all (even beyond f_Γ^D).

Jung et al. (2021) observed that (in our terminology) *variance* is not sensible for calibration. We now significantly generalize and tighten this observation into a characterization saying that (under mild assumptions) a property is sensible for calibration *if and only if it is elicitable* (or, equivalently, is identifiable/has convex level sets).

We begin by showing that if a property $\Gamma : \mathcal{P} \to \mathbb{R}$ does not have convex level sets on $\mathcal{P}$, then it is not sensible for calibration over any family of $\mathcal{P}$-compatible datasets that includes all possible datasets supported on two points in $\mathcal{X}$ whose respective label distributions belong to $\mathcal{P}$.

Definition 3.3 (2-Point Dataset Distribution)**.** A dataset distribution D over feature-label pairs $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ is called *2-point* if there exist two feature vectors $x_1 \neq x_2 \in \mathcal{X}$ such that $\Pr_D[X \notin \{x_1, x_2\}] = 0$ and $\Pr_D[X = x_1] \neq 0, \Pr_D[X = x_2] \neq 0$ — i.e., exactly two feature vectors have nonzero probability of occurring under distribution D.

Theorem 3.4 (No CxLS $\implies$ Not Sensible for Calibration)**.** *Fix a property $\Gamma : \mathcal{P} \to \mathbb{R}$ with $\mathcal{P}$ convex, and any family $\mathcal{D}$ of $\mathcal{P}$-compatible dataset distributions containing all the $\mathcal{P}$-compatible 2-point dataset distributions. Then, if Γ violates CxLS on $\mathcal{P}$, it is not sensible for calibration over $\mathcal{D}$.*

Proof. Suppose Γ violates CxLS on $\mathcal{P}$. Then, $\{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ is nonconvex for some $\gamma \in \text{Range}_\Gamma$. Thus, there exist distributions $Y_1, Y_2 \in \mathcal{P}$ such that $\Gamma(Y_1) = \Gamma(Y_2) = \gamma$ but $\Gamma(\lambda Y_1 + (1-\lambda)Y_2) \neq \gamma$ for some $\lambda \in [0, 1]$. Now construct a 2-point dataset distribution $D \in \mathcal{D}$ as follows: have it supported on any two feature vectors $x_1 \neq x_2 \in \mathcal{X}$, and set $Y_{x_1} = Y_1$, $Y_{x_2} = Y_2$, and $\Pr[X = x_1] = \lambda = 1 - \Pr[X = x_2]$. Then, its true distributional predictor f_Γ^D is not Γ-calibrated, as $\Gamma(Y_{f_\Gamma^D, \gamma}) \neq \gamma$. □

To prove the converse, we impose a weak and natural regularity assumption on the dataset distribution D: we require that the mapping from features x to the corresponding Y_x induced by D be just well-behaved enough that the label distributions Y_G over any $G \subseteq \mathcal{X}$ are well-defined as mixtures over the individual label distributions Y_x for $x \in G$.

In the case of $|\mathcal{Y}| < \infty$, the well-behaved nature of the mapping $x \to Y_x$ can be formalized by requiring it to be Lebesgue measurable. In the case when $|\mathcal{Y}| = \infty$, the space W_{TV} that each Y_x belongs to is a Banach space — and in this setting, the notions of Lebesgue measurability and integrability (which are only defined in finite-dimensional Euclidean spaces) are replaced by the analogous concepts of Bochner measurability and integrability (see e.g. Diestel & Uhl (1977) for formal definitions). Thus, when $|\mathcal{Y}| = \infty$, we assume the map $x \to Y_x$ is Bochner measurable.

Definition 3.5 ($\mathcal{P}$-Regular Dataset Distribution)**.** Fix feature space $\mathcal{X}$, label space $\mathcal{Y}$, and a family of probability distributions $\mathcal{P}$ over $\mathcal{Y}$. Consider a $\mathcal{P}$-compatible dataset distribution D over $\mathcal{X} \times \mathcal{Y}$. Let $\xi_D : \mathcal{X} \to \mathcal{P}$ be defined by $\xi_D(x) = Y_x$, i.e. ξ_D is the mapping $x \to Y_x$ from feature vectors to their label distributions induced by D. Then, D is called *$\mathcal{P}$-regular* if any of the following is true:

1. $\mathcal{X}$ is finite;
2. $\mathcal{Y}$ is finite ($|\mathcal{Y}| < \infty$) and ξ_D is Lebesgue measurable when $\mathcal{P}$ is viewed as a subset of $\mathbb{R}^{|\mathcal{Y}|}$;
3. $\mathcal{X}, \mathcal{Y}$ are infinite and ξ_D is Bochner measurable when $\mathcal{P}$ is viewed as a subset of W_{TV}.

For $\mathcal{P}$-regular datasets D, we now show that for any continuous property Γ that has CxLS on $\mathcal{P}$, the true distributional predictor f_Γ^D is Γ-calibrated. (The proof is in Appendix B.)

Theorem 3.6 (CxLS $\implies$ Sensible for Calibration)**.** *Consider a continuous property $\Gamma : \mathcal{P} \to \mathbb{R}$, and any family $\mathcal{D}$ of $\mathcal{P}$-regular dataset distributions. Then, if Γ has convex level sets on $\mathcal{P}$, it is sensible for calibration over $\mathcal{D}$.*

Together, Theorems 3.4 and 3.6 show under mild conditions that for continuous properties, sensibility for calibration is equivalent to having convex level sets. To finally link sensibility for calibration to elicitability and identifiability, we can now invoke the above stated Theorem 2.7 of Steinwart et al. (2014), with its extra assumptions on $\mathcal{P}$ (see Definition 2.6) and Γ (the nowhere-locally-constant assumption, see Footnote 1), to obtain our final characterization result:

Theorem 3.7 (Sensibility for Calibration $\iff$ Elicitability $\iff$ Identifiability $\iff$ CxLS)**.** *Let $\Gamma : \mathcal{P} \to \mathbb{R}$ be a continuous strictly locally non-constant property on a convex space of distributions $\mathcal{P} \in \{\mathcal{P}_0, \mathcal{P}_{>0}\}$. Let $\mathcal{D}$ be a family of $\mathcal{P}$-regular dataset distributions over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, that includes all the $\mathcal{P}$-compatible 2-point dataset distributions.*

Then Γ is sensible for calibration over $\mathcal{D}$ $\iff$ Γ is elicitable $\iff$ is identifiable $\iff$ has convex level sets.

Proof. Our Theorems 3.4 and 3.6 establish that Γ is sensible for calibration if and only if it has convex level sets on $\mathcal{P}$, which by Theorem 2.7 (Steinwart et al., 2014) is equivalent to Γ being elicitable, and to Γ being identifiable. □

4. Batch Multicalibration

In this section we give a generic batch Γ-multicalibration algorithm for elicitable properties Γ. It generalizes (and is similar to) past multicalibration algorithms designed for specific properties, like means (Hébert-Johnson et al., 2018; Gopalan et al., 2022a) and quantiles (Jung et al., 2023). These algorithms differ in their specifics; we most closely mirror the multicalibration algorithm of Jung et al. (2023).

We henceforth focus on continuous, strictly locally non-constant, bounded properties Γ, with $\text{Range}_\Gamma = [0,1]$. Our canonical batch Algorithm 1, described below, will produce *finite-range* multicalibrated Γ-predictors f: Namely, for any integer $m \geq 1$, denoting $[1/m] := \{\frac{1}{m+1}, \frac{2}{m+1}, \ldots, \frac{m}{m+1}\}$, Algorithm 1 finds an $O(\frac{1}{m})$-approximately multicalibrated Γ-predictor f with $\text{Range}_f = [1/m]$.

Any elicitable Γ, under the just stated assumptions, has a strictly consistent scoring function and an id function (by Theorem 2.7). We will now need a further assumption:

Assumption 4.1. Assume $\Gamma : \mathcal{P} \to \text{Range}_\Gamma$ has an identification function V such that $V(\cdot, Y)$ is strictly increasing and L-Lipschitz for each label distribution $Y \in \mathcal{P}$: $|V(\gamma, Y) - V(\gamma', Y)| \leq L|\gamma - \gamma'|$ for all γ, γ'.

This assumption is arguably mild. Let S be an antiderivative of V, so that $V(\gamma, y) = \frac{\partial S(\gamma, y)}{\partial \gamma}$. Then, V being strictly increasing is equivalent to S being a *convex* strictly consistent scoring function for Γ. Such a convexity assumption is quite natural in the context of optimization; furthermore, Finocchiaro & Frongillo (2018) show that for elicitable properties over finite label spaces $|\mathcal{Y}| < \infty$, this is without loss of generality. The extra Lipschitz assumption is what will allow us to quantify our algorithm's convergence rate.

To state Algorithm 1, it is convenient for us to re-parameterize our Definition 2.10 of Γ-multicalibration in terms of an id function V for Γ. (By properties of V, as this updated notion of calibration error goes to 0, so will the one in Definition 2.10.) Below, let $Y_{(G,\gamma)}$ be the label distribution conditional on the event $\{x \in \mathcal{X} : f(x) = \gamma, x \in G\}$.

Definition 4.2 (Approximate $(\mathcal{G}, V)$-Multicalibration)**.** Fix groups $\mathcal{G}$, a distribution D, and an id function V for a property Γ. A finite-range Γ-predictor $f : \mathcal{X} \to [0,1]$ is *α-approximately $(\mathcal{G}, V)$-multicalibrated* if for all $G \in \mathcal{G}$:

$$\sum_{\gamma \in \text{Range}_f} \Pr_{x \sim X}[f(x) = \gamma | x \in G] \cdot (V(\gamma, Y_{(G,\gamma)}))^2 \leq \frac{\alpha}{\Pr_{x \in X}[x \in G]}.$$

Algorithm 1 is quite natural. While it can, it finds an intersection Q_t of a group $G \in \mathcal{G}$ and a level set of the current predictor f, such that f's prediction on Q_t is too far from the truth (as measured by the magnitude of the expected id function value over Q_t) — and fixes the situation by shifting f's value on Q_t to the best grid point $\gamma \in [1/m]$.

Algorithm 1 BatchMulticalibration$(\Gamma, \mathcal{G}, m, f, L)$

Initialize $t = 1$ **and** $f_1 = f$**.**
Let $\alpha = \frac{4L^2}{m}$**, and let** $V : \text{Range}_\Gamma \times \mathcal{Y} \to \mathbb{R}$ be an L-Lipschitz id function for Γ satisfying Assumption 4.1.
while f_t not α-approximately $(\mathcal{G}, V)$-multicalibrated **do**
  Let $Q_t = \{x : f_t(x) = \gamma, x \in G\}$**, where**

$$(\gamma, G) \in \underset{(\gamma', G') \in [\frac{1}{m}] \times \mathcal{G}}{\text{argmax}} \Pr_x[f_t(x) = \gamma', x \in G'] (V(\gamma', Y_{(\gamma', G')}))^2.$$

  Let: $\gamma' = \text{argmin}_{\gamma'' \in [\frac{1}{m}]} |V(\gamma'', Y_{Q_t})|$
  Update: $f_{t+1}(x) := \mathbb{1}[x \notin Q_t] \cdot f_t(x) + \mathbb{1}[x \in Q_t] \cdot \gamma'$
  for all $x \in \mathcal{X}$**, and** $t \leftarrow t + 1$**.**
end while
Output f_t.

Theorem 4.3 (Guarantees of Algorithm 1)**.** *Fix data distribution $D \in \Delta\mathcal{Z}$ and groups $\mathcal{G} \subseteq 2^\mathcal{X}$. Fix an elicitable property Γ with its scoring function S and id function $V = \frac{\partial S}{\partial \gamma}$ satisfying Assumption 4.1, so that $V(\cdot, Y_Q)$ is L-Lipschitz on all label distributions Y_Q (for $Q \subseteq \mathcal{X}$) induced by D. Set discretization $m \geq 1$. If Algorithm 1 is initialized with predictor $f_1 : \mathcal{X} \to \mathbb{R}$ with score $\mathbb{E}_{(x,y)\sim D}[S(f_1(x), y)] = C_{init}$, and $C_{opt} = \mathbb{E}_{(x,y)\sim D}[S(f^D_\Gamma(x), y)]$ is the score of the true distributional predictor f^D_Γ, then Algorithm 1 produces a $\frac{4L^2}{m}$-approximately $(\mathcal{G}, V)$-multicalibrated Γ-predictor f after at most $(C_{init} - C_{opt})\frac{m^2}{L}$ updates.*

The proof of Theorem 4.3 is given in Appendix C, and is similar to the analyses of several existing multicalibration algorithms (Hébert-Johnson et al., 2018; Jung et al., 2023; Deng et al., 2023). We show that the expected score $\mathbb{E}_D[S]$ (where $V = \frac{\partial S}{\partial \gamma}$) decreases at every step of Algorithm 1, making it a potential function for the algorithm. Convergence rates follow by lower-bounding the per-iteration decrease in $\mathbb{E}[S]$. This naturally generalizes, to arbitrary elicitable properties, the existing analyses of mean (Hébert-Johnson et al., 2018) and quantile multicalibration (Jung et al., 2023), which respectively use squared loss (consistent for means) and pinball loss (consistent for quantiles) as potential functions.

We have described Algorithm 1 as able to directly query the expected identification function V on the true data distribution D. In practice, we would instead run it on the empirical distribution over an i.i.d. *sample* $\hat{D} \sim D^n$ of n points from D. Appendix D gives finite sample guarantees for this case.

5. Joint Multicalibration

Now we take a step towards understanding two interrelated issues: (1) how to multicalibrate vector-valued properties, and (2) how to appropriately extend the notion of multicalibration to some practically important scalar properties that have non-convex level sets and are thus neither elicitable nor sensible for calibration by our Theorem 3.4 (e.g. variance).

Specifically, we study the important case of two-dimensional properties $\Gamma = (\Gamma^0, \Gamma^1)$, where Γ^0 is elicitable whereas Γ^1 is *not* elicitable *per se*, but *is* elicitable *conditional* on any fixed value of Γ^0. We give an algorithm that can produce *jointly multicalibrated* estimators for such Γ.

Definition 5.1 (Conditional Elicitability)**.** We say that a property $\Gamma^1 : \mathcal{P} \to \mathbb{R}$ is *elicitable conditionally on a property* $\Gamma^0 : \mathcal{P} \to \mathbb{R}$, if the restriction of Γ^1 to each level set of Γ^0 (i.e. to each distribution family $\mathcal{P}_{\gamma^0} = \{P \in \mathcal{P} : \Gamma^0(P) = \gamma^0\}$ for $\gamma^0 \in \text{Range}_{\Gamma^0}$) is elicitable.

For the elicitable component Γ^0, we denote its scoring and identification functions by S^0, V^0. For property Γ^1 that is elicitable conditionally on Γ^0, for each $\gamma^0 \in \text{Range}_{\Gamma^0}$ we denote by $V^1_{\gamma^0} : \text{Range}_{\Gamma^1} \times \mathcal{Y} \to \mathbb{R}$ a function that identifies Γ^1 on every distribution P such that $\Gamma^0(P) = \gamma^0$, and by $S^1_{\gamma^0} : \text{Range}_{\Gamma^1} \times \mathcal{Y} \to \mathbb{R}$ a score that is strictly consistent for Γ^1 on every distribution P such that $\Gamma^0(P) = \gamma^0$.

We assume that the elicitable Γ^0 satisfies Assumption 4.1, with $V^0(\gamma^0, y)$ strictly increasing and L^0-Lipschitz in γ^0. We will also need the opposite (similarly mild) assumption:

Assumption 5.2. $V^0(\cdot, Y)$ is L^0_a-anti-Lipschitz at $\Gamma^0(Y)$ for all $Y \in \mathcal{P}$: $|\gamma^0 - \Gamma^0(Y)| \le L^0_a |V^0(\gamma^0, Y)|$ for all γ^0.

The situation with Γ^1 is more complex: it has different id functions $V^1_{\gamma^0}$ for different level sets of Γ^0, instead of a single function for all $P \in \mathcal{P}$. In general, nothing prevents these functions V_{γ^0} from being completely unrelated to each other for different values of γ^0 (and even undefined on each other's level sets). However, for most properties of interest we can expect $V^1_{\gamma^0}(\gamma^1, P)$ to vary continuously in γ^0 and be well-defined even for distributions P' for which $\Gamma^0(P') \neq \gamma^0$. To reflect this, and enable our joint multicalibration algorithm's guarantees, we make the following (mildly stronger) assumption:

Assumption 5.3. Assume for all $P \in \mathcal{P}$ that $V^1_{\gamma^0}(\gamma^1, P)$ is defined and is L_c-Lipschitz as a function of γ^0: that is, for any $\gamma^1, \gamma^0_1, \gamma^1_1$, $|V^1_{\gamma^0_0}(\gamma^1, P) - V^1_{\gamma^1_0}(\gamma^1, P)| \le L_c |\gamma^0_0 - \gamma^0_1|$.

Further, we assume that the conditional identification functions $V^1_{\gamma^0}$ for Γ^1 on Γ^0's level sets $\{\Gamma^0 = \gamma^0\}$ retain their "shape", i.e. remain strictly increasing and Lipschitz, even for distributions from other level sets of Γ^0. While this is a nontrivial assumption to make, in Section 6 we confirm that it holds e.g. when (Γ^0, Γ^1) is a so-called *Bayes pair*. This will let us establish Bayes pairs, an important and general class of properties (Embrechts et al., 2021), as a major use case for our theory of joint multicalibration.

Assumption 5.4. For all[2] $\gamma^0 \in \text{Range}_{\Gamma^0}$, assume $V^1_{\gamma^0}(\cdot, P)$ is L^1-Lipschitz and strictly increasing for all $P \in \mathcal{P}$.

We now define the central notion of jointly multicalibrated predictors for properties $\Gamma = (\Gamma^0, \Gamma^1) : \mathcal{P} \to \mathbb{R}^2$. Analogously to Definitions 2.10, 4.2, we define this concept in two versions: one that is parameterized by the property Γ itself, and another one that involves its id functions (V^0, V^1). As in Section 4, the latter version serves to simplify notation in our algorithm analysis (and as this notion of multicalibration error goes to 0, so does the one parameterized by (Γ^0, Γ^1)). We will use some shorthands (for $i = 0, 1$): $\mu_f(\gamma^i | G, \gamma^{1-i}) := \Pr_x[f^i(x) = \gamma^i | x \in G, f^{1-i}(x) = \gamma^{1-i}]$ and $\mu_f(G, \gamma^i) := \Pr_x[x \in G, f^i(x) = \gamma^i]$. Also, we let $Y_{(G,\gamma^0,\gamma^1)} := (Y | \{x \in G, f(x) = (\gamma^0, \gamma^1)\})$.

Definition 5.5 (Approximate Joint Multicalibration)**.** Fix distribution $D \in \Delta\mathcal{Z}$ and group family $\mathcal{G}$. Given a property $\Gamma = (\Gamma^0, \Gamma^1)$, a finite-range predictor $f = (f^0, f^1) : \mathcal{X} \to \mathbb{R}^2$ is (α^0, α^1)-*approximately* $(\mathcal{G}, \Gamma^0, \Gamma^1)$-*jointly multicalibrated* if for all $G \in \mathcal{G}$, $\gamma^0 \in \text{Range}_{f^0}$, $\gamma^1 \in \text{Range}_{f^1}$:

$$\sum_{\gamma^0 \in \text{Range}_{f^0}} \mu_f(\gamma^0 | G, \gamma^1) \cdot (\gamma^0 - \Gamma^0(Y_{(G,\gamma^0,\gamma^1)}))^2 \le \tfrac{\alpha^0}{\mu_f(G,\gamma^1)},$$

$$\sum_{\gamma^1 \in \text{Range}_{f^1}} \mu_f(\gamma^1 | G, \gamma^0) \cdot (\gamma^1 - \Gamma^1(Y_{(G,\gamma^0,\gamma^1)}))^2 \le \tfrac{\alpha^1}{\mu_f(G,\gamma^0)}.$$

Similarly, given id functions V^0, $\{V^1_{\gamma^0}\}_{\gamma^0 \in \text{Range}_{\Gamma^0}}$, predictor $f = (f^0, f^1)$ is (α^0, α^1)-*approximately* $(\mathcal{G}, V^0, V^1)$-*jointly multicalibrated* if for $G \in \mathcal{G}$, $\gamma^0 \in \text{Range}_{f^0}$, $\gamma^1 \in \text{Range}_{f^1}$:

$$\sum_{\gamma^0 \in \text{Range}_{f^0}} \mu_f(\gamma^0 | G, \gamma^1) \cdot \left(V^0(\gamma^0, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \le \tfrac{\alpha^0}{\mu_f(G,\gamma^1)}, \tag{1}$$

$$\sum_{\gamma^1 \in \text{Range}_{f^1}} \mu_f(\gamma^1 | G, \gamma^0) (V^1_{\Gamma^0(Y_{(G,\gamma^0,\gamma^1)})}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)}))^2 \le \tfrac{\alpha^1}{\mu_f(G,\gamma^0)}. \tag{2}$$

The Canonical Joint Multicalibration Algorithm: We now introduce `JointMulticalibration` (Algorithm 2), a canonical algorithm for learning a jointly multicalibrated predictor $f = (f^0, f^1)$ for (Γ^0, Γ^1). Algorithm 2 significantly generalizes the (mean, moment)-multicalibration algorithm of Jung et al. (2021), leading to some key differences in the analysis. We defer the discussion of these differences, as well as the proof of its convergence guarantees in Theorem 5.6, to Appendix E.1, and give only a brief overview of the algorithmic ideas below.

[2] In fact, we only need this to (much less restrictively) hold for $\gamma^0 \in [\frac{1}{m}]$, as Algorithm 2 gives predictors ranging in $[\frac{1}{m}] \times [\frac{1}{m}]$.

Algorithm 2 JointMulticalibration$((\Gamma^0,\Gamma^1),\mathcal{G},m,(f^0,f^1))$

Let V^0 and $\{V^1_{\gamma^0}\}_{\gamma^0\in[1/m]}$ be id functions for Γ^0 and Γ^1 satisfying Assumptions 4.1, 5.2, 5.3, 5.4.
Initialize $t=1$ **and** $f_1=(f^0,f^1)$.
while $\exists(\gamma^0,\gamma^1,G)\in[\frac{1}{m}]\times[\frac{1}{m}]\times\mathcal{G}$ s.t. $\Pr_{x\in\mathcal{X}}[f_t(x)=(\gamma^0,\gamma^1),x\in G]\left(V^0(\gamma^0,Y_{(\gamma^0,\gamma^1,G)})\right)^2\geq\frac{\alpha^0}{m}$ **do**
  Let $\mathcal{G}^0_t\leftarrow\{G\cap\{x\in\mathcal{X}\colon f^1_t(x)=\gamma^1\}\colon G\in\mathcal{G},\gamma^1\in[\frac{1}{m}]\}$
  Update $f^0_{t+1}\leftarrow\text{BatchMulticalibration}^V(V^0,\mathcal{G}^0_t,m,f^0_t,\alpha^0)$
  for $\gamma^0\in[1/m]$ **do**
    Let $\mathcal{G}^{1,\gamma^0}_t\leftarrow\{G\cap\{x\in\mathcal{X}:f^0_{t+1}(x)=\gamma^0\}:G\in\mathcal{G}\}$
    Let $f^{1,\gamma^0}_{t+1}\leftarrow\text{BatchMulticalibration}^V(V^1_{\gamma^0},\mathcal{G}^{1,\gamma^0}_t,m,f^1_t,\alpha^1)$
  end for
  Update $f^1_{t+1}(x)\leftarrow\sum_{\gamma^0\in[1/m]}\mathbb{1}[f^0_{t+1}(x)=\gamma^0]\cdot f^{1,\gamma^0}_{t+1}(x),\forall x\in\mathcal{X}$
  Update $t\leftarrow t+1$.
end while
Output $f_t=(f^0_t,f^1_t)$.

Theorem 5.6 (Guarantees of Algorithm 2)**.** *Consider any property $\Gamma=(\Gamma^0,\Gamma^1)$, with Γ^0 elicitable and Γ^1 elicitable conditionally on Γ^0, whose id functions satisfy Assumptions 4.1, 5.2, 5.3, 5.4. Fix any group family $\mathcal{G}\subseteq 2^{\mathcal{X}}$ and discretization $m\geq 1$. Set $\alpha^0=\frac{4(L^0)^2}{m}$ and $\alpha^1=\frac{4(L^1)^2}{m}$. Let $\alpha^1_*=\frac{8((L^0L^0_aL_c)^2+(L^1)^2)}{m}$. Then, `JointMulticalibration` (Algorithm 2) will output an (α^0,α^1_*)-approximately $(\mathcal{G},V^0,V^1)$-jointly multicalibrated Γ-predictor $f=(f^0,f^1)$, via at most $\frac{B^0B^1m^4}{L^0L^1}$ updates to f. Here, $B^0:=\sup_{\gamma,y\in[0,1]}S^0(\gamma,y)-\inf_{\gamma,y\in[0,1]}S^0(\gamma,y)$ for S^0 an antiderivative of $V^0(\gamma^0,y)$ wrt. γ^0, and $B^1:=\max_{\gamma^0\in[1/m]}\left(\sup_{\gamma,y\in[0,1]}S^1_{\gamma^0}(\gamma,y)-\inf_{\gamma,y\in[0,1]}S^1_{\gamma^0}(\gamma,y)\right)$ for each $S^1_{\gamma^0}$ an antiderivative of $V^1_{\gamma^0}(\gamma^1,y)$ wrt. γ^1.*

To train a two-dimensional predictor, we employ a two-stage structure whereby we alternately multicalibrate f^0 on the *current* level sets of f^1, and f^1 on the current level sets of f^0, until the desired level of joint multicalibration error is reached (in the sense of Equations 1, 2 of Definition 5.5).

As in Section 4, our predictor is discretized: $\text{Range}_{f^0}=\text{Range}_{f^1}=[\frac{1}{m}]$. Both f^0 and f^1 are updated via calls to a subroutine `BatchMulticalibration`^{V}, which is very similar to `BatchMulticalibration` (Algorithm 1) but has a stricter stopping rule to meet the extra demands of *joint* multicalibration, and accepts id functions V rather than properties Γ to simplify notation. We defer the pseudocode and the analysis for `BatchMulticalibration`^{V} to Algorithm 3 and Lemma E.1 in Appendix E.

Throughout the execution of Algorithm 2, the subroutine is invoked on auxiliary group families consisting of pairwise intersections of groups in $\mathcal{G}$ with the level sets of f^0, f^1. Due to updates to f^0 and f^1, these auxiliary groups are always in drift across these invocations, and careful bookkeeping is needed to verify that this does not prevent overall convergence. A key fact we prove towards this is that across *all* invocations on V^0 throughout Algorithm 2, the subroutine will perform boundedly many updates on f^0, implying that also f^1 will be re-calibrated at most that many times.

6. Applications

By combining our theory with known results from the elicitation literature in an essentially blackbox way, we can obtain several novel positive and negative results shedding light on an important question: when is it possible to produce (multi)calibrated predictors for various *risk measures*? We now summarize our results informally, and relegate the corresponding formal statements to Appendix G.

Joint Multicalibration of Bayes Pairs An elicitable property Γ by definition minimizes some scoring function S. Thus, we can view S as a *loss* which yields an accurate Γ-predictor when $\mathbb{E}[S]$ is minimized over a dataset. For instance, if Γ is the *mean*, we would minimize $\mathbb{E}_{(x,y)\sim D}[S(\gamma,y)]$ for $S(\gamma,y)=(\gamma-y)^2$ — which is just the familiar L_2 *regression*. As another example, τ-*quantiles* are elicited by optimizing the expected *pinball loss* $S_\tau(\gamma,y):=(1-\tau)\gamma+\max\{y-\gamma,0\}$; this procedure is known as *quantile regression*.

In loss minimization settings, one may care not about the minimizer per se, but rather about the loss value it induces, effectively asking: how large is the expected (strictly consistent) loss S at the true property value of Γ, i.e. $\min_\gamma\mathbb{E}_{(x,y)\sim D}[S(\gamma,y)]$? This object is known as the *Bayes risk* Γ^B of Γ with respect to S, and the two-dimensional property $\Gamma^{\text{BP}}:=(\Gamma,\Gamma^B)$ is a *Bayes pair* with respect to S.

Definition 6.1 (Bayes Risk, Bayes Pair)**.** Fix an elicitable $\Gamma:\mathcal{P}\to\mathbb{R}$ and a strictly consistent Γ-scoring function S. The property $\Gamma^B:\mathcal{P}\to\mathbb{R}$ given by $\Gamma^B(P):=S(\Gamma(P),P)$ for $P\in\mathcal{P}$ is the *Bayes risk* of Γ, and the property $\Gamma^{\text{BP}}:=(\Gamma,\Gamma^B)$ is the *Bayes pair* with respect to S.

For instance, (mean, variance) is a Bayes pair with respect to the squared loss. Another important Bayes pair, (quantile, CVaR), will be discussed shortly. As it turns out, Bayes risks are often not elicitable *per se* (Embrechts et al., 2021). However, any Bayes risk Γ^B is evidently elicitable *conditionally* on its underlying property Γ: knowing the value of Γ reveals the value of Γ^B. Thus, Bayes pairs are amenable to our joint multicalibration techniques:

Theorem 6.2 (Informal)**.** *Under mild assumptions, all Bayes pairs $\Gamma^{BP}:=(\Gamma,\Gamma^B)$ with respect to Lipschitz losses S are jointly multicalibratable using Algorithm 2.*

CVaR Multicalibration Conditional Value at Risk (CVaR), known also as Expected Shortfall (ES), is a tail risk measure of central significance in the literature. Proposed by Artzner et al. (1999) and Rockafellar & Uryasev (2000), τ-CVaR (for any $\tau \in [0,1]$) measures the mean of the top $(1-\tau)$-fraction of a random variable's largest values, defined (with $q_\tau(P)$ denoting the τ-quantile of P) as:

$$\text{CVaR}_\tau(P) := \mathbb{E}_{Y\sim P}[Y|Y > q_\tau(P)].$$

Given its ability to capture tail risk beyond the quantile, as well as many appealing features such as its *coherence* (see Artzner et al. (1999)), the CVaR has gained widespread prominence in areas ranging from finance to robust optimization. Its real-world significance is underscored by its recent introduction into international banking regulations, known as the Basel Accords, as a replacement for quantiles (called Value-at-Risk (VaR) in finance) for the purposes of market risk capital calculations (Embrechts et al., 2014). Thus, it is very important to ask if the CVaR is sensible for calibration — as this would let us train multicalibrated CVaR predictors using our canonical batch and online methods, complementing recent algorithmic quantile multicalibration results of Bastani et al. (2022) and Jung et al. (2023).

The answer to this question turns out to be nuanced. We show that while CVaR_τ is *not* sensible for calibration for any $\tau \in [0,1]$ (eliminating the possibility of directly training multicalibrated predictors for it), it can be multicalibrated *jointly* with the corresponding quantile q_τ.

Theorem 6.3 (Informal)**.** *For $\tau \in [0,1]$, τ-CVaR is* not sensible for calibration*. However, τ-CVaR is multicalibratable* jointly with the τ-quantile*, by instantiating Algorithm 2.*

For the negative part of this theorem, we simply invoke our Theorem 3.4 with a well-known result of (Gneiting, 2011) that *CVaR violates CxLS*. For the positive part, we invoke our joint multicalibration Theorem 6.2 by using the known fact that $(q_\tau, \text{CVaR}_\tau)$ is a Bayes pair relative to the (rescaled) τ-pinball loss (see e.g. Frongillo & Kash (2021)).

An Impossibility Result for Distortion Risk Measures Distortion risk measures (Wang et al., 1997) are a large theoretically and practically important class of risk measures. Its representatives include means, quantiles, CVaR, spectral risk measures, and numerous other important risk measures; see e.g. Kou & Peng (2016) or Gzyl & Mayoral (2008).

Definition 6.4 (Distortion Risk Measure)**.** Given a *distortion function* $h : [0,1] \to [0,1]$ (i.e., h is nondecreasing and satisfies $h(0) = 0$ and $h(1) = 1$), the corresponding *distortion risk measure* $\Gamma^h : \mathcal{P} \to \mathbb{R}$ is given by:[3]

$$\Gamma^h(P) := \int_{-\infty}^{0} \left(h\left(1-F_P(x)\right) - 1\right) dx + \int_{0}^{\infty} h(1-F_P(x))\, dx$$

for $P \in \mathcal{P}$, where F_P is the CDF of P.

For instance, the choice $h(x) = x$ for $x \in [0,1]$ leads to Γ^h being the distribution *mean*. Letting $h_\tau(x) = \mathbb{1}[x > 1-\tau]$ leads to Γ^{h_τ} being the *τ-quantile*.

As Wang & Ziegel (2015) and Kou & Peng (2016) showed, means and quantiles are essentially[4] *the only* distortion risks with convex level sets on finite-support distributions. Paired with our Theorem 3.4 (no CxLS $\implies$ not sensible for calibration), this yields a sweeping negative result:

Theorem 6.5 (Informal)**.** *No distortion risk measures, other than (essentially) means and quantiles, are sensible for calibration on any dataset family $\mathcal{D}$ which allows for finite-support label distributions.*

Thus, we learn that there will not be another multicalibration algorithm for distortion risks: the existing mean and quantile multicalibration methods are (essentially) the only ones.

Acknowledgments

This research was done in part while Georgy Noarov was visiting the Simons Institute for the Theory of Computing, and was supported in part by a Gift from AWS AI for Research in Trustworthy AI, the Simons Collaboration on the Theory of Algorithmic Fairness, and NSF grants FAI-2147212 and CCF-2217062. We warmly thank Christopher Jung and Arpit Agarwal for enlightening conversations at an early stage of this work.

References

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Bastani, O., Gupta, V., Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Practical adversarial multivalid conformal prediction. In *Neural Information Processing Systems (NeurIPS)*, 2022.

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

Deng, Z., Dwork, C., and Zhang, L. Happymap: A generalized multicalibration method. In *Innovations in Theoretical Computer Science (ITCS)*, 2023.

Diestel, J. and Uhl, J. J. Vector measures, vol. 15 of mathematical surveys. *American Mathematical Society, Providence, RI, USA*, 1977.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. Outcome indistinguishability. In *Proceedings*

[3] We assume the integrals exist for all $P \in \mathcal{P}$.

[4] Up to two further quantile-like properties; see Definition G.3 and Theorem G.4 in Section G for the precise statement.

of the 53rd Annual ACM SIGACT Symposium on Theory of Computing, pp. 1095–1108, 2021.

Dwork, C., Kim, M. P., Reingold, O., Rothblum, G. N., and Yona, G. Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature. In *International Conference on Algorithmic Learning Theory*, pp. 342–380. PMLR, 2022.

Embrechts, P., Puccetti, G., Rüschendorf, L., Wang, R., and Beleraj, A. An academic response to basel 3.5. *Risks*, 2 (1):25–48, 2014.

Embrechts, P., Mao, T., Wang, Q., and Wang, R. Bayes risk, elicitability, and the expected shortfall. *Mathematical Finance*, 31(4):1190–1217, 2021.

Finocchiaro, J. and Frongillo, R. Convex elicitation of continuous properties. *Advances in Neural Information Processing Systems*, 31, 2018.

Frongillo, R. M. and Kash, I. A. Elicitation complexity of statistical properties. *Biometrika*, 108(4):857–879, 2021.

Garg, S., Jung, C., Reingold, O., and Roth, A. Oracle efficient online multicalibration and omniprediction. *Manuscript*, 2023.

Globus-Harris, I., Harrison, D., Kearns, M., Roth, A., and Sorrell, J. Multicalibration as boosting for regression. *International Conference on Machine Learning (ICML)*, 2023.

Gneiting, T. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Good, I. J. Rational decisions. In *Breakthroughs in statistics*, pp. 365–377. Springer, 1992.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. Omnipredictors. In *ITCS*, 2022a.

Gopalan, P., Kim, M. P., Singhal, M. A., and Zhao, S. Low-degree multicalibration. In *Conference on Learning Theory*, pp. 3193–3234. PMLR, 2022b.

Gupta, V., Jung, C., Noarov, G., Pai, M. M., and Roth, A. Online multivalid learning: Means, moments, and prediction intervals. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

Gzyl, H. and Mayoral, S. On a relationship between distorted and spectral risk measures. *Rev Econ Financ*, 2008.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Jung, C., Lee, C., Pai, M., Roth, A., and Vohra, R. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pp. 2634–2678. PMLR, 2021.

Jung, C., Noarov, G., Ramalingam, R., and Roth, A. Batch multivalid conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2023.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F., and Reingold, O. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4): e2108097119, 2022.

Kou, S. and Peng, X. On the measurement of economic tail risk. *Operations Research*, 64(5):1056–1072, 2016.

Lambert, N. S., Pennock, D. M., and Shoham, Y. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 129–138, 2008.

Lee, D., Noarov, G., Pai, M. M., and Roth, A. Online minimax multiobjective optimization: Multicalibeating and other applications. In *Neural Information Processing Systems (NeurIPS)*, 2022.

Osband, K. H. *Providing Incentives for Better Cost Forecasting (Prediction, Uncertainty Elicitation)*. PhD thesis, University of California, Berkeley, 1985.

Rockafellar, R. T. and Uryasev, S. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Roth, A. Uncertain: Modern topics in uncertainty estimation. https://www.cis.upenn.edu/ aaroth/uncertainty-notes.pdf, 2022.

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

Steinwart, I., Pasin, C., Williamson, R., and Zhang, S. Elicitation and identification of properties. In *Conference on Learning Theory*, pp. 482–526. PMLR, 2014.

Wang, R. and Ziegel, J. F. Elicitable distortion risk measures: A concise proof. *Statistics & Probability Letters*, 100: 172–175, 2015.

Wang, S. S., Young, V. R., and Panjer, H. H. Axiomatic characterization of insurance prices. *Insurance: Mathematics and economics*, 21(2):173–183, 1997.

A. Additional Related Work

The main conceptual contribution of our paper is to connect the literature on *multicalibration* with the literature on *property elicitation*, both of which have a number of related threads.

Multicalibration (Mean) multicalibration in the batch setting was introduced by Hébert-Johnson et al. (2018), and has subsequently been generalized in a number of ways. As already discussed, Jung et al. (2021) study mean conditioned moment multicalibration in the batch setting, Gupta et al. (2022) study mean, quantile, and mean conditioned moment multicalibration in the sequential setting, and Jung et al. (2023) study quantile multicalibration in the batch setting. These generalizations can be seen as asking for calibration with respect to different distributional properties — i.e. they are generalizations of the type that we characterize in our paper. Dwork et al. (2021) study outcome indistinguishability, generalizing batch multi-calibration in a binary-label setting to allow for *distinguishers* that can evaluate the expectations of arbitrary predicates of triples $(x, f(x), y)$, and consider strengthenings in which the distinguisher gets further access to f — e.g. by being able to query it in arbitrary points, or by having access to its code. In particular, they show that without additional access to f, outcome indistinguishability can be reduced to (mean) multicalibration. Note that many distributional properties (e.g. variances, quantiles, etc) only become interesting when the label space is not binary, but rather consists of more than two labels or is real valued.

The most closely related works study abstract generalizations of multi-calibration. Dwork et al. (2022) study a generalization of outcome indistinguishability to real valued labels, and along the way consider batch multicalibration with respect to *linearizing statistics*. In our language, these are distributional properties Γ that behave linearly over mixture distributions — informally that for any two distributions D_1, D_2 and any $\alpha \in [0, 1]$, $\Gamma(\alpha D_1 + (1-\alpha)D_2) = \alpha\Gamma(D_1) + (1-\alpha)\Gamma(D_2)$. We adopt one of their definitions of multicalibration with respect to general distributional properties. All linearizing statistics have convex level sets (and so are elicitable), but not all elicitable properties are linear in this sense — for example, quantiles do not linearize. So our characterization of multicalibration implies that it is possible to multicalibrate with respect to a broader class of properties than are studied by Dwork et al. (2022). Lee et al. (2022) study a general online learning problem that they call "Online Minimax Multiobjective Optimization", and derive algorithms for multicalibration along with a number of other applications in this framework. We make use of this framework to derive our sequential multicalibration bounds, using an *identification function* that arises from the connection we make to property elicitation. Recent work of Deng et al. (2023) studies a one-dimensional generalization of multicalibration that asks for the condition that $\mathbb{E}[c(f(x), x)s(f(x), y)] = 0$ for abstract functions c and s, and derives sufficient (but not necessary) conditions under which this can be achieved in the batch setting. The algorithms we derive for batch multicalibration are similar to theirs, where an identification function takes the place of their s function, and a scoring function takes the place of their potential function; they do not consider sequential or multi-dimensional problems. Relative to this line of work, our result is the first to provide a characterization of when property multicalibration can be obtained, and to provide unifying results for both batch and sequential multicalibration.

There are also generalizations in orthogonal directions. Gopalan et al. (2022b) define "low degree multicalibration", which is a hierarchy of properties of predictors that are still trying to predict *means*, but relax the conditioning event that $f(x) = v$. At the bottom of the hierarchy is *multi-accuracy* (Hébert-Johnson et al., 2018; Kim et al., 2019) which does not condition on $f(x)$ at all. Multicalibration lies at the top of the hierarchy; in between are conditions that depend on $f(x)$ only smoothly, through a degree k polynomial. Gopalan et al. (2022b) show that intermediate levels of this hierarchy have some of the desirable properties of multicalibration and can be easier to obtain. Several works (Kim et al., 2019; Gopalan et al., 2022a; Kim et al., 2022; Globus-Harris et al., 2023) study generalizations of mean multicalibration in which "groups" representing subsets of the data domain are relaxed to arbitrary real valued functions and give a number of applications. In this setting, Globus-Harris et al. (2023) provide a characterization of when mean multicalibration implies Bayes optimality. See Roth (2022) for an introductory exposition of much of this work.

Property elicitation Brier (1950), Good (1992), and Savage (1971) study *proper scoring rules* — which are contracts for paying an expert as a function of their prediction and of the realized outcome, with the property that they maximally reward the expert (in expectation) for *truthfully* reporting their estimate of the probability of the outcome event. Since scoring rules directly elicit probabilities, they could in principle be used to elicit an entire probability distribution by eliciting the probability of every event in its support, but this is generally infeasible. Instead, Lambert et al. (2008) introduce the problem of *property elicitation*, whose goal is to design contracts that incentivize experts to truthfully report some *property* of a large or infinite support distribution — like its mean, variance, median, etc. Informally a property is *elicitable* if there

exists some function of a report and an outcome that in expectation over the outcome is minimized at the property value. There is now a large literature on property elicitation; we make use of several key results. Osband (1985) and Gneiting (2011) define the notion of an identification function for a property, which like a scoring rule is a function of a report and an outcome; an identification function takes value 0 in expectation over the outcome if the report is equal to the property value. Steinwart et al. (2014) prove a central characterization theorem (subject to mild technical conditions) — a continuous property is elicitable if and only if it has an identification function if and only if it has convex level sets. The characterization of Steinwart et al. (2014) holds generally for continuous outcome spaces. When the outcome space is finite, Finocchiaro & Frongillo (2018) show that (subject to technical conditions), elicitable properties can be elicited with convex scoring rules.

B. Proof of Theorem 3.6

Theorem 3.6 (CxLS $\implies$ Sensible for Calibration)**.** *Consider a continuous property $\Gamma : \mathcal{P} \to \mathbb{R}$, and any family $\mathcal{D}$ of $\mathcal{P}$-regular dataset distributions. Then, if Γ has convex level sets on $\mathcal{P}$, it is sensible for calibration over $\mathcal{D}$.*

Proof. Suppose that Γ has convex level sets on $\mathcal{P}$. We now establish that for every dataset distribution $D \in \mathcal{D}$, the true distributional predictor f_Γ^D is calibrated: namely, that for all $\gamma \in \text{Range}_{f_\Gamma^D} \subseteq \text{Range}_\Gamma$, we have $\Gamma(Y_{f_\Gamma^D,\gamma}) = \gamma$. For the remainder of the proof, fix any dataset distribution $D \in \mathcal{D}$ (which is $\mathcal{P}$-regular by assumption) and any value $\gamma \in \text{Range}_{f_\Gamma^D}$.

Let $\text{LS}_\Gamma(\gamma) = \{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ be the γ-level set of property Γ — which is convex by our assumption on Γ. Further, let $Q_\gamma := \{x \in \mathcal{X} : Y_x \in \text{LS}_\Gamma(\gamma)\}$ be the feature space region consisting of all points $x \in \mathcal{X}$ whose label distributions have property value γ. Let X_γ be the conditional probability measure on Q_γ induced by the dataset distribution D.

Proving that $\Gamma(Y_{f_\Gamma^D,\gamma}) = \gamma$ is equivalent to establishing that $Y_{f_\Gamma^D,\gamma} \in \text{LS}_\Gamma(\gamma)$. The conditional distribution $Y_{f_\Gamma^D,\gamma}$ is a mixture distribution over the individual label distributions Y_x for $x \in Q_\gamma$, so we can write the following formal expression:

$$Y_{f_\Gamma^D,\gamma} = \mathop{\mathbb{E}}_{x\sim X_\gamma}[Y_x] := \int_{Q_\gamma} Y_x \, dX_\gamma \, .$$

If $\mathcal{X}$ is finite: In this case, $Y_{f_\Gamma^D,\gamma}$ is just a convex combination of the constituent distributions Y_x for $x \in Q_\gamma$:

$$Y_{f_\Gamma^D,\gamma} = \sum_{x\in Q_\gamma} X_\gamma(x) \cdot Y_x.$$

The convexity of the level set $\text{LS}_\Gamma(\gamma)$ then implies that indeed $Y_{f_\Gamma^D,\gamma} \in \text{LS}_\Gamma(\gamma)$, since $Y_{f_\Gamma^D,\gamma}$ is a convex combination, over $x \in Q_\gamma$, of distributions $Y_x \in \text{LS}_\Gamma(\gamma)$.

If $\mathcal{Y}$ is finite: In this case, $Y_{f_\Gamma^D,\gamma} = \mathbb{E}_{x\sim X_\gamma}[Y_x]$ is naturally given by a *Lebesgue integral* of the random variable Y_x over the simplex $\Delta(d) \subset \mathbb{R}^d$. By our assumption that the dataset is $\mathcal{P}$-regular, we have that Y_x is a bounded (e.g. in the ℓ_∞ norm) and Lebesgue measurable random variable. Consequently, Y_x is in fact Lebesgue integrable, so the expectation $\mathbb{E}_{x\sim X_\gamma}[Y_x]$ is well-defined and evaluates to some point $u \in \mathbb{R}^d$. It remains to show that $u \in \text{LS}_\Gamma(\gamma)$.

For this, introduce the indicator function $\mathbb{1}_{\text{LS}_\Gamma(\gamma)} : \Delta(d) \to \{0\} \cup \{+\infty\}$, defined to be 0 for $Y_x \in \text{LS}_\Gamma(\gamma)$, and ∞ otherwise. As the set $\text{LS}_\Gamma(\gamma)$ is convex, its indicator function $\mathbb{1}_{\text{LS}_\Gamma(\gamma)}$ is convex. Therefore, we can apply Jensen's inequality to $\mathbb{1}_{\text{LS}_\Gamma(\gamma)}$ to conclude that:

$$\mathbb{1}_{\text{LS}_\Gamma(\gamma)}(u) = \mathbb{1}_{\text{LS}_\Gamma(\gamma)}\left(\mathop{\mathbb{E}}_{x\sim X_\gamma}[Y_x]\right) \leq \mathop{\mathbb{E}}_{x\sim X_\gamma}\left[\mathbb{1}_{\text{LS}_\Gamma(\gamma)}(Y_x)\right] = 0,$$

implying that $\mathbb{1}_{\text{LS}_\Gamma(\gamma)}(u) = 0$. By definition of $\mathbb{1}_{\text{LS}_\Gamma(\gamma)}$, this demonstrates that $u \in \text{LS}_\Gamma(\gamma)$, as desired.

$\mathcal{X}, \mathcal{Y}$ infinite: In this case, we define $Y_{f_\Gamma^D,\gamma} = \mathbb{E}_{x\sim X_\gamma}[Y_x]$ as the *Bochner integral* of the Bochner measurable map ξ_D. (Recall that ξ_D is defined in Definition 3.5, and see Diestel & Uhl (1977) for formal definitions and properties of Bochner measurability and integrability.) By a standard Bochner integrability criterion (see Theorem 2 on p. 45 of Diestel & Uhl (1977)), this integral indeed exists and evaluates to a point in the ambient space W_{TV}, as it is easy to check that $\mathbb{E}_{x\sim X_\gamma}[||Y_x||_{\text{TV}}] < \infty$ (indeed, the TV norm of any probability distribution is 1 so $\mathbb{E}_{x\sim X_\gamma}[||Y_x||_{\text{TV}}] = 1 < \infty$). Again, we want to show that $Y_{f_\Gamma^D,\gamma} = \mathbb{E}_{x\sim X_\gamma}[Y_x] \in \text{LS}_\Gamma(\gamma)$. For this, we use the following result, which can be interpreted as a mean value theorem for Bochner integrals:

Fact 2 (Corollary 8 on p. 48 of Diestel & Uhl (1977)). Let $(\Omega, \mathcal{A}, \mu)$ be a finite measure space, E a Banach space, and $f : \Omega \to E$ a Bochner μ-integrable map. For $G \subseteq E$, let $\overline{\text{co}}(G)$ be the closure of the convex hull of G. Then, for any $A \in \mathcal{A}$ with $\mu(A) > 0$, the Bochner integral of f over A belongs to the closure of the convex hull of the image of A under f:

$$\frac{1}{\mu(A)} \int_A f \, d\mu \in \overline{\text{co}}(f(A)).$$

To instantiate this fact, we let: (1) $\Omega := \mathcal{X}$, together with the probability measure induced by the dataset over $\mathcal{X}$; (2) the Banach space $E := W_{\text{TV}}$; (3) the Bochner integrable mapping $f := \xi_D$; and (4) the measurable event $A := Q_\gamma \subseteq \mathcal{X}$.

Note that $f(A) = f(Q_\gamma) \subseteq \text{LS}_\Gamma(\gamma)$ (with this inclusion being strict whenever there is some label distribution $P \in \text{LS}_\Gamma(\gamma)$ that is *not* induced by the dataset distribution D conditional on any $x \in \mathcal{X}$), so we have that $\overline{\text{co}}(f(A)) \subseteq \overline{\text{co}}(\text{LS}_\Gamma(\gamma))$. Observe that $\text{LS}_\Gamma(\gamma)$ is convex by assumption, and it is also closed in the standard metric topology on W_{TV} since it is the preimage under the continuous mapping Γ of the closed singleton $\{\gamma\} \in \text{Range}_\Gamma$. Thus, $\text{LS}_\Gamma(\gamma)$ is a closed convex set so $\overline{\text{co}}(\text{LS}_\Gamma(\gamma)) = \text{LS}_\Gamma(\gamma)$. As a result, we in fact see that $\overline{\text{co}}(f(A)) \subseteq \text{LS}_\Gamma(\gamma)$.

Since X_γ is the conditional distribution induced by D over $x \in Q_\gamma$, we can see that $\frac{1}{\mu(A)} \int_A f d\mu = \mathbb{E}_{x \sim X_\gamma}[Y_x] = Y_{f_\Gamma^D, \gamma}$. Together with our observation that $\overline{\text{co}}(f(A)) \subseteq \text{LS}_\Gamma(\gamma)$, this lets us conclude by Fact 2 that $Y_{f_\Gamma^D, \gamma} \in \text{LS}_\Gamma(\gamma)$, as desired. □

C. Convergence Analysis for Batch Multicalibration (Proof of Theorem 4.3)

Our convergence analysis of Algorithm 1 will utilize the following natural *potential function*:

Definition C.1. The *potential* for Algorithm 1 at round t is:

$$\Phi_t := \mathop{\mathbb{E}}_{(x,y) \sim D}[S(f_t(x), y)] = \mathop{\mathbb{E}}_{x \sim X}[S(f_t(x), Y_x)],$$

where $f_t : \mathcal{X} \to \mathbb{R}$ is the property predictor at the beginning of iteration t of the algorithm and S is a strictly consistent scoring function for property Γ satisfying Assumption 4.1.

First, we prove the following helper Lemma that bounds the change in S — the potential function of Algorithm 1 — in the scenario where an incorrect prediction γ for the property value $\Gamma(Y)$ is corrected on a label distribution Y. We will later use this fact to bound the progress of the algorithm after every update to the predictor f for Γ.

Lemma C.2. *Consider any property $\Gamma : \mathcal{P} \to \mathbb{R}$. Suppose $S : \text{Range}_\Gamma \times \mathcal{Y} \to \mathbb{R}$ and $V : \text{Range}_\Gamma \times \mathcal{Y} \to \mathbb{R}$ with $V(\gamma, y) = \frac{\partial S(\gamma, y)}{\partial \gamma}$ are a strictly consistent scoring function and the corresponding identification function for Γ that satisfy Assumption 4.1. Then for any $\gamma \in \text{Range}_\Gamma$ and any label distribution $Y \in \mathcal{P}$, letting L_Y be the Lipschitz constant of $V(\cdot, Y)$, it holds that:*

$$\frac{(V(\gamma, Y))^2}{2L_Y} \leq S(\gamma, Y) - S(\Gamma(Y), Y) \leq V(\gamma, Y)(\gamma - \Gamma(Y)) - \frac{(V(\gamma, Y))^2}{2L_Y}.$$

Proof. We will prove this result with the help of the following claim.

Claim 1. For any L-Lipschitz increasing function h defined on any interval $[a, b]$, it holds that:

$$h(a)(b-a) + \frac{(h(b) - h(a))^2}{2L} \leq \int_a^b h(t)dt \leq h(b)(b-a) - \frac{(h(b) - h(a))^2}{2L}.$$

Proof. Under these constraints on h, the largest value of the integral $\int_a^b h(t)dt$ would be obtained if $h(t)$ first increased from $h(a)$ to $h(b)$ for $t \in [a, t']$, where $t' \in [a, b]$ is defined by $(t' - a)L = h(b) - h(a)$, at the fastest rate possible (that is, at the rate L), and stayed constant at the value $h(b)$ for $t \in [t', b]$. The integral of this piecewise linear function on $[a, b]$ gives the upper bound.

Conversely, the smallest value of the integral $\int_a^b h(t)dt$ would be obtained if $h(t)$ first stayed constant at the value $h(a)$ for $t \in [a, t']$, where t' is defined so that $(b - t')L = h(b) - h(a)$, and then increased from $h(a)$ to $h(b)$ at the fastest rate possible (that is, at the rate L) for $t \in [t', b]$. Integrating this function on $[a, b]$ gives the claimed lower bound. □

As V is a derivative of S, by the fundamental theorem of calculus we get: $S(\gamma, Y) - S(\Gamma(Y), Y) = \int_{\Gamma(Y)}^{\gamma} V(t, Y)dt$.

First assume $\gamma \geq \Gamma(Y)$. By Assumption 4.1, V continuously increases from $\Gamma(Y)$ to γ, and has Lipschitz constant L_Y. Then, by Claim 1, and using that $V(\Gamma(Y), Y) = 0$, we obtain

$$\frac{(V(\gamma, Y))^2}{2L_Y} \leq \int_{\Gamma(Y)}^{\gamma} V(t, Y)dt \leq V(\gamma, Y)(\gamma - \Gamma(Y)) - \frac{(V(\gamma, Y))^2}{2L_Y}.$$

Now assume $\gamma < \Gamma(Y)$. Then, we have:

$$S(\gamma, Y) - S(\Gamma(Y), Y) = \int_{\Gamma(Y)}^{\gamma} V(t, Y)dt = -\int_{\gamma}^{\Gamma(Y)} V(t, Y)dt.$$

By Assumption 4.1, V continuously increases from γ to $\Gamma(Y)$, and has Lipschitz constant L_Y. By Claim 1, we have: $-V(\Gamma(Y), Y)(\Gamma(Y) - \gamma) + \frac{(V(\Gamma(Y),Y) - V(\gamma,Y))^2}{2L_Y} \leq -\int_{\gamma}^{\Gamma(Y)} V(t, Y)dt \leq -V(\gamma, Y)(\Gamma(Y) - \gamma) - \frac{(V(\Gamma(Y),Y) - V(\gamma,Y))}{2L_Y}$, which from $V(\Gamma(Y), Y) = 0$ simplifies to: $\frac{(V(\gamma,Y))^2}{2L_Y} \leq -\int_{\gamma}^{\Gamma(Y)} V(t, Y)dt \leq V(\gamma, Y)(\gamma - \Gamma(Y)) - \frac{(V(\gamma,Y))^2}{2L_Y}$. Thus, we have shown our bound for both cases $\gamma \geq \Gamma(Y)$ and $\gamma < \Gamma(Y)$. □

Now, we are ready to prove Theorem 4.3, which gives the convergence rate for Algorithm 1. We restate the theorem here for convenience.

Theorem 4.3 (Guarantees of Algorithm 1)**.** *Fix data distribution $D \in \Delta\mathcal{Z}$ and groups $\mathcal{G} \subseteq 2^{\mathcal{X}}$. Fix an elicitable property Γ with its scoring function S and id function $V = \frac{\partial S}{\partial \gamma}$ satisfying Assumption 4.1, so that $V(\cdot, Y_Q)$ is L-Lipschitz on all label distributions Y_Q (for $Q \subseteq \mathcal{X}$) induced by D. Set discretization $m \geq 1$. If Algorithm 1 is initialized with predictor $f_1 : \mathcal{X} \to \mathbb{R}$ with score $\mathbb{E}_{(x,y)\sim D}[S(f_1(x), y)] = C_{init}$, and $C_{opt} = \mathbb{E}_{(x,y)\sim D}[S(f_\Gamma^D(x), y)]$ is the score of the true distributional predictor f_Γ^D, then Algorithm 1 produces a $\frac{4L^2}{m}$-approximately $(\mathcal{G}, V)$-multicalibrated Γ-predictor f after at most $(C_{init} - C_{opt})\frac{m^2}{L}$ updates.*

Proof. Suppose the algorithm has not halted at round t. Thus, f_t does not yet satisfy α-approximate $(\mathcal{G}, V)$-multicalibration, so by the pigeonhole principle there is a pair $(G, \gamma) \in \mathcal{G} \times [1/m]$ such that on the set $Q_t := \{x \in \mathcal{X} : x \in g, f_t(x) = \gamma\}$:

$$|V(\gamma, Y_{Q_t})| \geq \sqrt{\frac{\alpha/m}{\Pr_{x\sim X}[x \in Q_t]}}. \tag{3}$$

Now, letting $\gamma' = \operatorname{argmin}_{\gamma'' \in [1/m]} |V(\gamma'', Y_{Q_t})|$, the algorithm will update $f_t \to f_{t+1}$ via the rule:

$$f_{t+1}(x) := \mathbb{1}[x \notin Q_t] \cdot f_t(x) + \mathbb{1}[x \in Q_t] \cdot \gamma'.$$

From the definition of the potential function values Φ_t and Φ_{t+1}, we have

$$\begin{aligned}
\Phi_{t+1} &= \Pr_{x\sim X}[x \in Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x)|x \in Q_t] + \Pr_{x\sim X}[x \notin Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x)|x \notin Q_t] \\
&= \Pr_{x\sim X}[x \in Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x)|x \in Q_t] + \Pr_{x\sim X}[x \notin Q_t] \mathop{\mathbb{E}}_{x\in X}[S(f_t(x), Y_x)|x \notin Q_t] \\
&= \Phi_t + \Pr_{x\sim X}[x \in Q_t] \left(\mathop{\mathbb{E}}_{x\in X}[S(f_{t+1}(x), Y_x) - S(f_t(x), Y_x)|x \in Q_t] \right) \\
&= \Phi_t + \Pr_{x\sim X}[x \in Q_t] \left(\mathop{\mathbb{E}}_{x\in X}[S(\gamma', Y_x) - S(\gamma, Y_x)|x \in Q_t] \right) \\
&= \Phi_t + \Pr_{x\sim X}[x \in Q_t] \left(S(\gamma', Y_{Q_t}) - S(\gamma, Y_{Q_t}) \right).
\end{aligned}$$

Here Step 2 follows because $f_t(x) = f_{t+1}(x)$ for all x outside Q_t, and Step 5 uses the fact that f_t and f_{t+1} are both constant on Q_t to rewrite expected scores of f_t, f_{t+1} over $x \sim X$ simply as the scores of the predictor values γ, γ' with respect to the mixture distribution Y_{Q_t} of labels over the region Q_t.

From here, we have:

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \Pr_{x\sim X}[x \in Q_t]\Big((S(\gamma', Y_{Q_t}) - S(\Gamma(Y_{Q_t}), Y_{Q_t})) - (S(\gamma, Y_{Q_t}) - S(\Gamma(Y_{Q_t}), Y_{Q_t}))\Big) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t})) - \frac{(V(\gamma', Y_{Q_t}))^2}{2L} - \frac{(V(\gamma, Y_{Q_t}))^2}{2L}\Big) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t})) - \frac{(V(\gamma', Y_{Q_t}))^2}{2L}\Big) - \frac{\alpha}{2Lm} \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t}))\Big) - \frac{\alpha}{2Lm} \\
&\leq V(\gamma', Y_{Q_t})(\gamma' - \Gamma(Y_{Q_t})) - \frac{\alpha}{2Lm} \\
&\leq L|\gamma' - \Gamma(Y_{Q_t})| \cdot |\gamma' - \Gamma(Y_{Q_t})| - \frac{\alpha}{2Lm} \\
&\leq \frac{L}{m^2} - \frac{\alpha}{2Lm}.
\end{aligned}$$

The equality follows by introducing an added and subtracted term $S(\Gamma(Y_{Q_t}), Y_{Q_t})$. The 1st inequality applies the upper and lower bound of Lemma C.2 to the two score differences. The 2nd inequality follows by the α-miscalibration condition of Equation 3. The 3rd inequality drops the nonpositive term $-\Pr_{x\sim X}[x \in Q_t]\frac{(V(\gamma', Y_{Q_t}))^2}{2L}$. The 4th inequality drops the factor $\Pr_{x\sim X}[x \in Q_t] \leq 1$. The 5th inequality holds since by the Lipschitzness of V, we have $|V(\gamma', Y_{Q_t})| = |V(\gamma', Y_{Q_t}) - V(\Gamma(Y_{Q_t}), Y_{Q_t})| \leq L|\gamma' - \Gamma(Y_{Q_t})|$. The 6th inequality holds because γ' must be at most $\frac{1}{m}$ away from the true property value $\Gamma(Y_{Q_t})$ on Q_t: the algorithm cannot select a farther grid point γ'', as that would result in a greater value of $|V(\gamma'', Y_{Q_t})|$, by the structure of the m-discretization and the monotonic increase of $|V(\cdot, Y_{Q_t})|$ in both directions away from $\Gamma(Y_{Q_t})$.

Now setting $\alpha = \frac{4L^2}{m}$, we get $\Phi_{t+1} - \Phi_t \leq \frac{L}{m^2} - \frac{\alpha}{2Lm} = -\frac{L}{m^2}$. Telescoping this over the rounds $t = 1, \ldots, T$, where T is the total number of iterations before convergence, we obtain:

$$\Phi_T - \Phi_1 \leq -T\frac{L}{m^2}.$$

By assumption $\Phi_1 = C_{\text{init}}$ and $\Phi_T \geq C_{\text{opt}}$, so we have $T\frac{L}{m^2} \leq \Phi_1 - \Phi_T \leq C_{\text{init}} - C_{\text{opt}}$, and thus

$$T \leq (C_{\text{init}} - C_{\text{opt}})\frac{m^2}{L},$$

concluding the proof. □

D. Finite Sample Guarantees for Batch Multicalibration

We have described Algorithm 1 as if it has direct access to the underlying distribution D (since it computes expectations of the identification function V on the underlying distribution). In general we do *not* have access to D directly, and instead have access only to a sample $\hat{D} \sim D^n$ of n points sampled i.i.d. from D. In practice, we would run the algorithm on the empirical distribution over the n points in $\hat{D}$, and its guarantees would carry over to the underlying distribution D from which these points were sampled. Jung et al. (2023) proved this for the special case of quantiles, but in fact their proof uses nothing other than the conditions in Assumption 4.1. We state the more general version of the theorem here (implicit in Jung et al. (2023)) and briefly sketch the argument. We note that this argument is to establish that Algorithm 1 generalizes when used as an empirical risk minimization algorithm. An alternative way to obtain similar generalization bounds would be to follow the strategy of Hébert-Johnson et al. (2018) and use techniques from adaptive data analysis to implement a statistical query oracle, and to then modify the algorithm so as to compute the quantities $V(\gamma, Q_t)$ only through this oracle.

Theorem D.1 (Implicit in Jung et al. (2023))**.** *Fix a distribution $D \in \Delta\mathcal{Z}$ and a property Γ together with a bounded identification function V. Suppose Algorithm 1 is run using the empirical distribution $\hat{D} \sim D^n$ over n i.i.d. samples drawn from D. Then if Algorithm 1 halts after T rounds and returns a model f_T, with probability $1 - \delta$ over the randomness of the data distribution, f^T satisfies α-approximate $(\mathcal{G}, V)$-multicalibration with respect to D for:*

$$\alpha = \frac{4L^2}{m} + O\left(\sqrt{\frac{\ln(1/\delta) + T\ln(m^2|\mathcal{G}|)m}{n}} + \frac{m\ln(1/\delta) + T\ln(m^2|\mathcal{G}|)}{n}\right).$$

The proof has a simple structure. Since V is bounded, expectations of V over D (i.e. the quantities of the form $V(\gamma, Y_{G,\gamma})$ that appear in the definition of approximate $(\mathcal{G}, V)$-multicalibration) concentrate around their expectations with high probability when evaluated on the empirical distribution $\hat{D}$. Thus, if the model f_T was *fixed*, we would get that its $(\mathcal{G}, V)$-multicalibration error is similar in-sample and out-of-sample by union-bounding over each $G \in \mathcal{G}$ and $\gamma \in \text{Range}_{f_t} = [1/m]$. Of course the model f_T output by the algorithm is not fixed before $\hat{D}$ is sampled, so to establish the claim, it is also necessary that we union-bound over *all* models f_T that might be output. But we can do this; since Algorithm 1 produces models with range restricted to $[1/m]$, then for any $t \in [T]$, we can easily see for any fixed model f_t that at most $|G|m^2$ models f_{t+1} could possibly result at the next step — at most one for every choice of $G \in \mathcal{G}$, $\gamma \in [1/m]$, and $\gamma' \in [1/m]$ at iteration t of Algorithm 1. Thus, fixing any initial model $f_1 = f$, the number of models f_T that might be output after the final step T of the algorithm is bounded by $(|G|m^2)^T$. Theorem D.1 then follows by union-bounding over all such models.

Theorem D.1 upper bounds the generalization error of Algorithm 1 in terms of the number of rounds T before it halts. Thus, when paired with an upper bound on T, it gives a worst-case bound on generalization error. Theorem 4.3 upper bounds the round complexity by $T \leq O\left(\frac{m^2}{L}\right)$, but there is a catch: L here is the Lipschitz constant for expectations of V taken over the true underlying distribution D, and this will generally not be preserved over the empirical distribution $\hat{D}$. Nevertheless, Jung et al. (2023) show that the same convergence bound holds when run on $\hat{D}$ (up to constants) — by arguing that each round of the algorithm run on $\hat{D}$ decreases the potential function as measured on D (where the Lipschitz assumption has been made). This is because the algorithm decides on its update each round by measuring quantities of the form $V(\gamma, Q)$ which are expectations of a bounded function V, and so concentrate around their true values.

Theorem D.2 (Implicit in Jung et al. (2023))**.** *Fix a distribution $D \in \Delta\mathcal{Z}$ (which induces a set of conditional label distributions Y_Q for each $Q \subset \mathcal{X}$) and a property Γ together with an identification function V. Assume that Γ and V together with the set of label distributions $\mathcal{P} = \{Y_Q : Q \subset \mathcal{X}\}$ together satisfy Assumption 4.1 with Lipschitz constant L. Suppose Algorithm 1 is run using the empirical distribution on a dataset $\hat{D} \sim D^n$ consisting of n i.i.d. samples from D. Then for any $\delta > 0$, if:*

$$n \geq \Omega\left(\ln\left(\frac{m^2}{L\delta}\right) + \frac{m^2}{L}\ln\left(\frac{|G|m}{L}\right)\frac{m^4}{L^2}\right).$$

with probability $1 - \delta$ over the randomness of D, Algorithm 1 halts after at most $T = O\left(\frac{m^2}{L}\right)$ many steps.

Together with Theorem D.1, this establishes a worst-case generalization bound for batch property multicalibration that is polynomial in all of the parameters of the problem and in the assumed Lipschitz constant L of the property's identification function V.

E. Joint Multicalibration Guarantees

We begin by formally defining the subroutine $\texttt{BatchMulticalibration}^V$ as the following Algorithm 3:

Algorithm 3 $\text{BatchMulticalibration}^V(V, \mathcal{G}, m, f, \alpha)$

Initialize $t = 1$ **and** $f_1 = f$**.**
while $\exists(\gamma, G) \in [1/m] \times \mathcal{G}$ such that $\Pr_{x\in\mathcal{X}}[f_t(x) = \gamma, x \in G]\left(V(\gamma, Y_{(\gamma,G)})\right)^2 \geq \alpha/m$ **do**
  Let $Q_t = \{x : f_t(x) = \gamma, x \in G\}$
  Let:

$$\gamma' = \underset{\gamma''\in[1/m]}{\text{argmin}} |V(\gamma'', Y_{Q_t})|$$

  Update: $f_{t+1}(x) := \mathbb{1}[x \notin Q_t] \cdot f_t(x) + \mathbb{1}[x \in Q_t] \cdot \gamma'$ **for all** $x \in \mathcal{X}$**, and** $t \leftarrow t + 1$**.**
end while
Output f_t.

We here state the guarantees enjoyed by Algorithm 3. This statement is stronger than that of the guarantees for the similar Algorithm 1, in two ways: (1) the notion of achieved multicalibration error at convergence (Equation 4) is stronger than that of Algorithm 1; and (2) we show that even with the input group family $\mathcal{G}$ not fixed beforehand (and thus potentially changing over time), Algorithm 3 will never perform more than a certain number of updates to the predictor f, and if it does

perform that many updates then it will be approximately calibrated conditional on *all* measurable subsets of $\mathcal{X}$ (rather than just the ones it explicitly performed updates on).

Lemma E.1. *Set $\alpha = \frac{4L^2}{m}$. $\texttt{BatchMulticalibration}^V$ (Algorithm 3), when run on a function V that is monotonically increasing and L-Lipschitz in its first argument, outputs a $[1/m]$-discretized predictor f that satisfies:*

$$\Pr_{x\in\mathcal{X}}[f(x) = \gamma, x \in G]\left(V(\gamma, Y_{(\gamma,G)})\right)^2 \leq \frac{\alpha}{m} \quad \textit{for all } \gamma \in [1/m], G \in \mathcal{G}. \tag{4}$$

Moreover, Algorithm 3 terminates in at most $\frac{Bm^2}{L}$ iterations, where $B = \sup_{\gamma,y\in[0,1]} S(\gamma, y) - \inf_{\gamma,y\in[0,1]} S(\gamma, y)$ for S an antiderivative of V, and if it runs for that long, the resulting predictor will satisfy (4) for all *(measurable) regions $G \subseteq \mathcal{X}$.*

Proof. Denote by S an antiderivative of V, and define, similar to the proof of Theorem 4.3, the potential value at iteration t of Algorithm 3 as:

$$\Phi_t := \mathop{\mathbb{E}}_{(x,y)\sim D}[S(f_t(x), y)] = \mathop{\mathbb{E}}_{x\sim X}[S(f_t(x), Y_x)].$$

Suppose that at round t, the algorithm finds a violation of its `while` loop condition for some $G \in \mathcal{G}$ and $\gamma \in [1/m]$. Let $Q_t = \{x \in \mathcal{X} : x \in \mathcal{G}, f(x) = \gamma\}$. Via the same calculations as in the proof of Theorem 4.3, we have that

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &\leq \Pr_{x\sim X}[x \in Q_t]\Big(V(\gamma', Y_{Q_t})(\gamma' - \gamma_t^*) - \frac{(V(\gamma, Y_{Q_t}))^2}{2L}\Big) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(|V(\gamma', Y_{Q_t})||\gamma' - \gamma_t^*| - \frac{(V(\gamma, Y_{Q_t}))^2}{2L}\Big) \\
&\leq \Pr_{x\sim X}[x \in Q_t]\Big(L|\gamma' - \gamma_t^*|^2 - \frac{(V(\gamma, Y_{Q_t}))^2}{2L}\Big),
\end{aligned}$$

where we denote by γ_t^* the unique point such that $V(\gamma_t^*, Y_{Q_t}) = 0$ (it is the analog of $\Gamma(Y_{Q_t})$ in the proof of Theorem 4.3). This argument is still valid as it rests on Lemma C.2, which requires properties of V and S that are still satisfied here.

Now, just as in the aforementioned proof, we have by the monotonicity of $V(\cdot, Y_{Q_t})$ that since the algorithm chooses $\gamma' = \operatorname{argmin}_{\gamma''\in[1/m]} |V(\gamma'', Y_{Q_t})|$, it must be that $|\gamma' - \gamma_t^*| \leq \frac{1}{m}$, and so we obtain

$$\Phi_{t+1} - \Phi_t \leq \frac{L}{m^2} - \frac{1}{2L}\Pr_{x\sim X}[x \in Q_t]|V(\gamma, Y_{Q_t})|^2.$$

Since the condition of the `while` loop demands that $\Pr_{x\sim X}[x \in Q_t]\,|V(\gamma, Y_{Q_t})|^2 \geq \alpha/m$, we get

$$\Phi_{t+1} - \Phi_t \leq \frac{L}{m^2} - \frac{\alpha}{2Lm}.$$

Setting $\alpha = \frac{4L^2}{m}$, we then have $\Phi_{t+1} - \Phi_t \leq -\frac{L}{m^2}$, and thus, by telescoping, $\Phi_T - \Phi_0 \leq -\frac{TL}{m^2}$. Since by definition $B = \sup_{\gamma,y\in[0,1]} S(\gamma, y) - \inf_{\gamma,y\in[0,1]} S(\gamma, y)$, we also have $\Phi_T - \Phi_0 \geq -B$, and therefore $T \leq \frac{Bm^2}{L}$, providing an upper bound on the number of iterations of the algorithm.

Importantly, in this argument we never referenced the actual definition of Q_t (i.e. that $Q_t = \{x \in \mathcal{X} : x \in \mathcal{G}, f(x) = \gamma\}$) — we only used that it satisfies the `while` loop condition, i.e. $\Pr_{x\sim X}[x \in Q_t]\,|V(\gamma, Y_{Q_t})|^2 \geq \alpha/m$. Therefore, the upper bound $\frac{Bm^2}{L}$ on the total number of iterations in fact holds for any arbitrary sequence of regions $Q_1, Q_2, \ldots$ where each $Q_i \subseteq \mathcal{X}$ is measurable with respect to the marginal data distribution over $\mathcal{X}$. As a result, we know that *if* $\texttt{BatchMulticalibration}^V$ does run for at least $\frac{Bm^2}{L}$ iterations, then as soon as it finishes iteration $t = \frac{Bm^2}{L}$, there will not exist *any* measurable $Q \subseteq \mathcal{X}$ violating condition (4). Thus, no matter which group family $\mathcal{G}$ $\texttt{BatchMulticalibration}^V$ is run on (and even if the group family were to change arbitrarily during the execution), it will never update the predictor f more than $\frac{Bm^2}{L}$ times, concluding the proof. □

E.1. Convergence Analysis for the `Joint Multicalibration` Algorithm 2

Now we are ready to prove our convergence guarantee for the canonical `Joint Multicalibration` Algorithm 2. As mentioned in Section 5, Algorithm 2 significantly generalizes the (mean, moment)-multicalibration algorithm of Jung et al. (2021), leading to some key differences in the analysis.

Notably, in our terminology, in their specific case the re-calibration of f^1 given f^0 can be cast as a single mean multicalibration subroutine using what they call a "pseudo-label" technique. At our level of generality, this does not work anymore as we are forced to work with different id functions $V^1_{\gamma^0}$ for Γ^1 on each level set $\{f^0 = \gamma^0\}$. This is why our inner `for` loop iterates over the level sets of f^0, re-calibrating f^1 using m separate invocations of the subroutine (fortunately, these can actually be run in parallel, since f^0's level sets are disjoint). Even with this construction in hand, our potential function argument from Section 4 does not easily port over: each level set $\{f^0 = \gamma^0\}$ can overlap with multiple level sets of Γ^0, so the true property Γ^1 will generally not admit a single scoring function on $\{f^0 = \gamma^0\}$ that could be used as a potential. This is where our assumptions on the behavior of $V^1_{\gamma^0}$ with respect to γ^0 crucially enable us to show that, subject to f^0 being sufficiently multicalibrated, using the proxy id $V^1_{\gamma^0}$ on the level set $\{f^0 = \gamma^0\}$ will not cause the multicalibration subroutines for Γ^1 to fail to converge.

Theorem 5.6 (Guarantees of Algorithm 2)**.** *Consider any property $\Gamma = (\Gamma^0, \Gamma^1)$, with Γ^0 elicitable and Γ^1 elicitable conditionally on Γ^0, whose id functions satisfy Assumptions 4.1, 5.2, 5.3, 5.4. Fix any group family $\mathcal{G} \subseteq 2^{\mathcal{X}}$ and discretization $m \geq 1$. Set $\alpha^0 = \frac{4(L^0)^2}{m}$ and $\alpha^1 = \frac{4(L^1)^2}{m}$. Let $\alpha^1_* = \frac{8((L^0 L^0_a L_c)^2 + (L^1)^2)}{m}$. Then, `JointMulticalibration` (Algorithm 2) will output an (α^0, α^1_*)-approximately $(\mathcal{G}, V^0, V^1)$-jointly multicalibrated Γ-predictor $f = (f^0, f^1)$, via at most $\frac{B^0 B^1 m^4}{L^0 L^1}$ updates to f. Here, $B^0 := \sup_{\gamma,y\in[0,1]} S^0(\gamma, y) - \inf_{\gamma,y\in[0,1]} S^0(\gamma, y)$ for S^0 an antiderivative of $V^0(\gamma^0, y)$ wrt. γ^0, and $B^1 := \max_{\gamma^0\in[1/m]} \left(\sup_{\gamma,y\in[0,1]} S^1_{\gamma^0}(\gamma, y) - \inf_{\gamma,y\in[0,1]} S^1_{\gamma^0}(\gamma, y)\right)$ for each $S^1_{\gamma^0}$ an antiderivative of $V^1_{\gamma^0}(\gamma^1, y)$ wrt. γ^1.*

Proof. **Runtime:** First, observe that the `while` loop in Algorithm 2 will stop after at most $\frac{B^0 m^2}{L^0}$ iterations if we set $\alpha^0 = \frac{4(L^0)^2}{m}$. Indeed, all invocations of $\texttt{BatchMulticalibration}^V$ on f^0 with the identification function V^0 can be pieced together into a single process that first multicalibrates f^0 with respect to $\mathcal{G}^0_1$, then takes the resulting predictor and multicalibrates it with respect to $\mathcal{G}^0_2$, and so on until the stopping condition of the `while` loop in `JointMulticalibration` is met. This is equivalent to a single run of $\texttt{BatchMulticalibration}^V$ where the group family is externally updated from time to time: $\mathcal{G}^0_1 \to \mathcal{G}^0_2 \to \ldots \to \mathcal{G}^0_t \to \ldots$. But by Lemma E.1, this process cannot perform a total of more than $\frac{B^0 m^2}{L^0}$ updates on the predictor f^0. Since the predictor f^0 is updated at least once in each iteration of the `while` loop of `JointMulticalibration`, this also bounds the number of iterations of the `while` loop.

Now, for each iteration of the `while` loop, we have m calls to $\texttt{BatchMulticalibration}^V$ as applied to all identification functions $V^1_{\gamma^0}$ for $\gamma^0 \in [1/m]$. Again by Lemma E.1, each of them takes at most $\frac{B^1 m^2}{L^1}$ updates to converge. This follows directly from Assumption 5.4, which states that $V_{\gamma^0}(\cdot, P)$ is L^1-Lipschitz and monotonically increasing for all $P \in \mathcal{P}$ (not just for P such that $\Gamma^0(P) = \gamma^0$). Naively, running the subroutine m times, once for each level set of f^0_{t+1}, would amount to a total of $m \cdot \frac{B^1 m^2}{L^1} = \frac{B^1 m^3}{L^1}$ iterations. But in fact, all these m invocations can be viewed as a single invocation of $\texttt{BatchMulticalibration}^V$ that updates the predictor f^1 for Γ^1 using an identification function V^1_* defined as $V^1_{\gamma^0}$ on each level set $\{f^0_{t+1} = \gamma^0\}$ (which is well defined since these level sets partition the domain $\mathcal{X}$). Therefore, there will be only at most $\frac{B^1 m^2}{L^1}$ across *all* these m invocations of $\texttt{BatchMulticalibration}^V$.

Taking the above observations together, Algorithm 2 will therefore terminate after at most $\frac{B^0 m^2}{L^0} \cdot \frac{B^1 m^2}{L^1} = \frac{B^0 B^1 m^4}{L^0 L^1}$ updates to f^0 and to f^1, as claimed.

Multicalibration Guarantees: Now, we show that the predictors f^0_T, f^1_T output at termination satisfy the conditions of Definition 5.5 of approximate joint multicalibration.

By the stopping condition of the while loop, at termination we have for all γ^0, γ^1, G that $\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0, \gamma^1), x \in G] \left(V^0(\gamma^0, Y_{(\gamma^0,\gamma^1,G)})\right)^2 \leq \frac{\alpha^0}{m}$, implying after dividing by $\Pr_{x\in\mathcal{X}}[x \in G, f^1_T(x) = \gamma^1]$ that

$$\Pr_{x\in\mathcal{X}}[f^0_T(x) = \gamma^0 | x \in G, f^1_T(x) = \gamma^1] \left(V^0(\gamma^0, Y_{(\gamma^0,\gamma^1,G)})\right)^2 \leq \frac{\alpha^0/m}{\Pr[x \in G, f^1_T(x) = \gamma^1]} \text{ for all } G \in \mathcal{G}, \gamma^1 \in \text{Range}_{f^1_T}. \tag{5}$$

For every $G \in \mathcal{G}$ and $\gamma^1 \in \text{Range}_{f^1_T}$, summing this inequality over all at most m values $\gamma^0 \in \text{Range}_{f^0_T}$, we obtain that:

$$\sum_{\gamma^0\in\text{Range}_{f^0_T}} \Pr_{x\sim X}[f^0_T(x) = \gamma^0 | x \in G, f^1_T(x) = \gamma^1] \cdot \left(V^0(\gamma^0, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \leq \frac{\alpha^0}{\Pr_{x\in X}[x \in G, f^1_T(x) = \gamma^1]},$$

so the predictor f_T^0 satisfies its joint multicalibration condition (1) of Definition 5.5.

Now we show that the predictor f_T^1 for Γ^1 satisfies its joint multicalibration condition (2). By construction, for each $\gamma^0 \in [1/m]$ the function f_T^1 is equal to f_T^{1,γ^0} in the region $\{x \in \mathcal{X} : f_T^0(x) = \gamma^0\}$. Since f_T^{1,γ^0} is output by the corresponding call to $\texttt{BatchMulticalibration}^V$, by Lemma E.1 this guarantees for each $G' \in \mathcal{G}_T^{1,\gamma^0} = \{G \cap \{x \in \mathcal{X} : f_T^0(x) = \gamma^0\} : G \in \mathcal{G}\}$ and for each $\gamma^1 \in [1/m]$ that $\Pr_{x\in\mathcal{X}}[f_T^1(x) = \gamma^1, x \in G'] \left(V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^1,G')})\right)^2 \leq \frac{\alpha^1}{m}$, implying that $\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G] \left(V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})\right)^2 \leq \frac{\alpha^1}{m}$ for all $G \in \mathcal{G}$.

Therefore, we have for all γ^0, γ^1, G the bound

$$\left|V_{\gamma^0}^1(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})\right| \leq \sqrt{\frac{\alpha^1/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}}. \tag{6}$$

But observe that we instead want to bound $\left|V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}\left(\gamma^1, Y_{(G,\gamma^0,\gamma^1)}\right)\right|$, the absolute value of the *true* identification function on this set. This is where we can make use of Assumption 5.3, which gives us L_c-Lipschitzness of $V_{\gamma^0}^1(\cdot,\cdot)$ as a function of γ^0, as well as Assumption 5.2, which gives us L_a^0-anti-Lipschitzness of $V^0(\gamma^0,\cdot)$ as a function of γ^0: we obtain that

$$\begin{aligned}
|V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})| &\leq |V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)}) - V^1_{\gamma^0}(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| + |V^1_{\gamma^0}(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| \\
&\leq L_c|\gamma^0 - \Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)| + |V^1_{\gamma^0}(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| \\
&\leq L_c L_a^0 |V^0(\gamma^0, Y_{(G,\gamma^0,\gamma^1)})| + |V^1_{\gamma^0}(\gamma^1, Y_{(\gamma^0,\gamma^1,G)})| \\
&\leq L_c L_a^0 \sqrt{\frac{\alpha^0/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}} + \sqrt{\frac{\alpha^1/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}},
\end{aligned}$$

where the fourth step is by substituting in Inequalities 5 and 6.

From here, for all γ^0, γ^1, G we have the bound

$$\left|V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})\right| \leq \frac{(L_c L_a^0\sqrt{\alpha^0} + \sqrt{\alpha^1})/\sqrt{m}}{\sqrt{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]}},$$

and after squaring both sides of the inequality, we obtain:

$$\left(V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \leq \frac{(L_c L_a^0\sqrt{\alpha^0} + \sqrt{\alpha^1})^2/m}{\Pr_{x\in\mathcal{X}}[f_T(x) = (\gamma^0,\gamma^1), x \in G]} \quad \text{for all } \gamma^0, \gamma^1, G.$$

Now multiplying both sides by $\Pr_{x\in\mathcal{X}}[f_T^1(x) = \gamma^1 | x \in G, f_T^0(x) = \gamma^0]$ and noting that

$$(L_c L_a^0\sqrt{\alpha^0} + \sqrt{\alpha^1})^2 \leq 2((L_c L_a^0)^2\alpha^0 + \alpha^1) = 2((L_c L_a^0)^2 \cdot 4(L^0)^2 + 4(L^1)^2)/m = 8((L^0 L_a^0 L_c)^2 + (L^1)^2)/m = \alpha_*^1,$$

we get:

$$\Pr_{x\in\mathcal{X}}[f_T^1(x) = \gamma^1 | x \in G, f_T^0(x) = \gamma^0] \left(V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \leq \frac{\alpha_*^1/m}{\Pr_{x\in\mathcal{X}}[f_T^0(x) = \gamma^0, x \in G]} \quad \text{for all } \gamma^0, \gamma^1, G.$$

For every $G \in \mathcal{G}$ and $\gamma^0 \in \text{Range}_{f_T^0}$, summing this inequality over all at most m values $\gamma^1 \in \text{Range}_{f_T^1}$, we obtain that:

$$\sum_{\gamma^1 \in \text{Range}_{f_T^1}} \Pr_{x\in\mathcal{X}}[f_T^1(x) = \gamma^1 | x \in G, f_T^0(x) = \gamma^0] \left(V^1_{\Gamma^0\left(Y_{(G,\gamma^0,\gamma^1)}\right)}(\gamma^1, Y_{(G,\gamma^0,\gamma^1)})\right)^2 \leq \frac{\alpha_*^1}{\Pr_{x\in\mathcal{X}}[f_T^0(x) = \gamma^0, x \in G]},$$

so the predictor f_T^1 satisfies its joint multicalibration condition (2) of Definition 5.5, thus concluding the proof. □

F. Sequential Multicalibration

We now turn to the sequential adversarial setting, in which there is no underlying distribution, and our goal will be to obtain approximate $(\mathcal{G}, V)$-multicalibration (Definition 4.2) on the *empirical distribution* defined by the transcript π of an interaction between the Learner and an Adversary. This generalizes sequential multicalibration for means and quantiles studied by Gupta et al. (2022) to arbitrary elicitable properties. In fact, even for quantiles, we give a strengthening of the result of Gupta et al. (2022) — they give an ℓ_∞ variant of calibration that makes use of "bucketing" in its conditioning event — we give a bound on the same ℓ_2-notion of calibration we use for batch calibration, without any bucketing. Garg et al. (2023) similarly obtain this guarantee for sequential mean multicalibration.

F.1. Setup and Preliminaries

F.1.1. The Sequential Learning Setting

In the sequential setting, a *Learner* interacts with an *Adversary* in rounds $t = 1$ to T as follows:

1. The Adversary chooses a feature vector $x_t \in \mathcal{X}$ and a distribution $Y_t \in \Delta\mathcal{Y}$ (possibly subject to some restrictions), and reveals x_t to the Learner.
2. The Learner makes a prediction $p_t \in \mathbb{R}$.
3. The Adversary samples $y_t \sim Y_t$ and reveals y_t to the Learner.

The record of the interaction accumulates in a transcript $\pi = \{(x_t, p_t, y_t)\}_{t=1}^{T}$. For any $s \leq T$ and transcript π, the prefix of the transcript $\pi^{<s}$ is defined as $\pi^{<s} = \{(x_t, p_t, y_t)\}_{t=1}^{s-1}$. We write $\Pi^{<s}$ for the domain of all transcripts of length $< s$. A Learner is a collection of mappings (for each round $t \leq T$) $\mathcal{L}_t : \Pi^{<t} \times \mathcal{X} \to \Delta\mathbb{R}$, and an Adversary is a collection of mappings $\mathcal{A}_t : \Pi^{<t} \to \mathcal{X} \times \Delta\mathcal{Y}$, specifying their behavior given their observations thus far.

Now we can introduce our strong, ℓ_2, definition of online multicalibration that we will then show how to achieve.

Definition F.1 (Online Multicalibration)**.** Fix a transcript $\pi = \{(x_t, p_t, y_t)\}_{t=1}^{T}$. Let $n(\pi, G) = |\{t : x_t \in G\}|$ denote the number of rounds containing a member of group G in π, and $n(\pi, \gamma, G) = |\{t : x_t \in G, p_t = \gamma\}|$ denote the number of rounds containing a group G in which the prediction p_t was γ.

Fix π, a collection of groups $\mathcal{G}$, a property Γ, and an identification function V for Γ. We say that the transcript π is *α-approximately $(\mathcal{G}, V)$-multicalibrated* if for all $G \in \mathcal{G}$:

$$\sum_{\gamma} \frac{n(\pi, \gamma, G)}{n(\pi, G)} \left(\sum_{t : p_t = \gamma, x_t \in G} \frac{V(\gamma, y_t)}{n(\pi, \gamma, G)} \right)^2 \leq \alpha \frac{T}{n(\pi, G)}.$$

Remark F.2. Observe that this is exactly the definition of approximate multicalibration we gave in Definition 4.2, in which the empirical distribution over π replaces the distribution $\mathcal{D}$.

We can simplify the notion of multicalibration somewhat by canceling terms:

Observation 1. Fix a transcript π, a collection of groups $\mathcal{G}$, a property Γ, and an identification function V for Γ. For each group $G \in \mathcal{G}$ define the quantity:

$$K_2(G, \pi) = \sum_{\gamma} \frac{1}{n(\pi, \gamma, G)} \left(\sum_{t : p_t = \gamma, x_t \in G} V(\gamma, y_t) \right)^2.$$

Then π is α-approximately $(\mathcal{G}, V)$-multicalibrated if $K_2(G, \pi) \leq \alpha T$ for all groups $G \in \mathcal{G}$.

In the online setting, our goal will be to control the growth of $K_2(G, \pi)$ as the transcript is generated, for each $G \in \mathcal{G}$. The following Lemma will be key:

Lemma F.3. *Fix a partial transcript $\pi^{<s} = \{(x_t, p_t, y_t)\}_{t=1}^{s-1}$ and a one-round continuation (x_s, p_s, y_s). Write $\pi^{\leq s} = \pi^{<s} \circ (x_s, p_s, y_s)$ for the transcript extended by one round. Define:*

$$R(\pi^{<s}, G, \gamma) = \sum_{t < s : p_t = \gamma, x_t \in G} V(\gamma, y_t).$$

Then for every $G \in \mathcal{G}$, if $x_s \notin G$, we have:

$$K_2(G, \pi^{\leq s}) - K_2(G, \pi^{<s}) = 0.$$

If $x_s \in G$ and $p_s = \gamma$, we have:

$$K_2(G, \pi^{\leq s}) - K_2(G, \pi^{<s}) \leq \frac{1}{n(\pi^{<s}, \gamma, G)} \left(2V(\gamma, y_s)R(\pi^{<s}, G, \gamma) + V(\gamma, y_s)^2\right).$$

Proof. If $x_s \notin G$, then $K_2(G, \pi^{\leq s}) = K_2(G, \pi^{<s})$ by definition and we are done. Otherwise, if $x_s \in G$ we can calculate:

$$\begin{aligned}
& K_2(G, \pi^{\leq s}) - K_2(G, \pi^{<s}) \\
= \; & \frac{1}{n(\pi^{<s}, \gamma, G) + 1} \left(\left(\sum_{t<s : p_t = \gamma, x_t \in G} V(\gamma, y_t) \right) + V(\gamma, y_s) \right)^2 - \frac{1}{n(\pi^{<s}, \gamma, G)} \left(\sum_{t<s : p_t = \gamma, x_t \in G} V(\gamma, y_t) \right)^2 \\
\leq \; & \frac{1}{n(\pi^{<s}, \gamma, G)} \left(\left(\sum_{t<s : p_t = \gamma, x_t \in G} V(\gamma, y_t) \right) + V(\gamma, y_s) \right)^2 - \frac{1}{n(\pi^{<s}, \gamma, G)} \left(\sum_{t<s : p_t = \gamma, x_t \in G} V(\gamma, y_t) \right)^2 \\
\leq \; & \frac{1}{n(\pi^{<s}, \gamma, G)} \left(2V(\gamma, y_s)R(\pi^{<s}, G, \gamma) + V(\gamma, y_s)^2\right).
\end{aligned}$$

This concludes the proof. □

F.1.2. Key Tool: Online Minimax Multiobjective Optimization

For our online algorithm below, we will use the Multiobjective Optimization framework introduced by Lee et al. (2022).

Definition F.4 (Online Minimax Multiobjective Optimization Setting)**.** A Learner plays against an Adversary over rounds $t \in [T] := \{1, \ldots, T\}$. Over these rounds, the Learner accumulates a d-dimensional loss vector ($d \geq 1$), where each round's loss vector lies in $[-C, C]^d$ for some $C > 0$. At each round t, the Learner and the Adversary interact as follows:

1. Before round t, the Adversary selects and reveals to the Learner an *environment* comprising:
   (a) The Learner's and Adversary's respective convex compact action sets $\mathcal{X}^t$, $\mathcal{Y}^t$ embedded into a finite-dimensional Euclidean space;
   (b) A continuous vector valued loss function $\ell^t(\cdot, \cdot) : \mathcal{X}^t \times \mathcal{Y}^t \to [-C, C]^d$, with each $\ell^t_j(\cdot, \cdot) : \mathcal{X}^t \times \mathcal{Y}^t \to [-C, C]$ (for $j \in [d]$) convex in the 1st and concave in the 2nd argument.
2. The Learner selects some $x^t \in \mathcal{X}^t$.
3. The Adversary observes the Learner's selection x^t, and responds with some $y^t \in \mathcal{Y}^t$.
4. The Learner suffers (and observes) the loss vector $\ell^t(x^t, y^t)$.

The Learner's objective is to minimize the value of the maximum dimension of the accumulated loss vector after T rounds—in other words, to minimize: $\max_{j \in [d]} \sum_{t \in [T]} \ell^t_j(x^t, y^t)$.

A key quantity in the analysis of the Learner's performance in the online minimax multi-objective optimization setting is the Adversary-Moves-First value of the stage games at each round t of the interaction — i.e. how well the Learner could do if (counter-factually) she knew the Adversary's action ahead of time.

Definition F.5 (Adversary-Moves-First (AMF) Value at Round t)**.** The *Adversary-Moves-First value* of the game defined by the environment $(\mathcal{X}^t, \mathcal{Y}^t, \ell^t)$ at round t is:

$$w^t_A := \sup_{y^t \in \mathcal{Y}^t} \min_{x^t \in \mathcal{X}^t} \left(\max_{j \in [d]} \ell^t_j(x^t, y^t) \right).$$

We can measure the performance of the Learner by comparing it to a benchmark defined by the Adversary moves first values of the games defined at each round.

Definition F.6 (Adversary-Moves-First (AMF) Regret)**.** On transcript $\pi^t = \{(\mathcal{X}^s, \mathcal{Y}^s, \ell^s), x^s, y^s\}_{s=1}^t$, we define the Learner's Adversary Moves First (AMF) Regret for the j^{th} dimension at time t to be:

$$R_j^t(\pi^t) := \sum_{s=1}^{t} \ell_j^s(x^s, y^s) - \sum_{s=1}^{t} w_A^s.$$

The overall *AMF Regret* is then defined as follows: $R^t(\pi^t) = \max_{j\in[d]} R_j^t$.

Lee et al. (2022) show that in any online minimax multiobjective optimization setting, the following Algorithm 4 obtains diminishing AMF regret.

Algorithm 4 General Algorithm for the Learner that Achieves Sublinear AMF Regret

for rounds $t = 1, \ldots, T$ **do**

Learn adversarially chosen $\mathcal{X}^t, \mathcal{Y}^t$, and loss function $\ell^t(\cdot, \cdot)$.

Let $$\chi_j^t := \frac{\exp\left(\eta \sum_{s=1}^{t-1} \ell_j^s(x^s, y^s)\right)}{\sum_{i\in[d]} \exp\left(\eta \sum_{s=1}^{t-1} \ell_i^s(x^s, y^s)\right)} \text{ for } j \in [d].$$

Play $$x^t \in \operatorname*{argmin}_{x\in\mathcal{X}^t} \max_{y\in\mathcal{Y}^t} \sum_{j\in[d]} \chi_j^t \cdot \ell_j^t(x, y).$$

Observe the Adversary's selection of $y^t \in \mathcal{Y}^t$.

end for

Theorem F.7 (AMF Regret Guarantee of Algorithm 4 (Lee et al., 2022))**.** *For any $T \geq \ln d$, Algorithm 4 with learning rate $\eta = \sqrt{\frac{\ln d}{4TC^2}}$ obtains, against any Adversary, AMF regret bounded by: $R^T \leq 4C\sqrt{T \ln d}$.*

F.2. Canonical Algorithm for Sequential Multicalibration

In the rest of this section, we show how for any bounded and continuous elicitable property Γ with a Lipschitz identification function V, and for any finite group structure $\mathcal{G}$, the problem of obtaining diminishing $(\mathcal{G}, V)$-multicalibration error in the sequential adversarial setting can be cast as an instance of online minimax multiobjective optimization, and so can be solved with an appropriate instantiation of Algorithm 4 with multicalibration error bounds following from an appropriate instantiation of Theorem F.7.

Assumptions Throughout this section, we fix a continuous elicitable property Γ with a bounded range, which we w.l.o.g. rescale such that $\text{Range}_\Gamma = [0, 1]$; and we also fix an identification function V for Γ. Moreover, we will assume that the label space $\mathcal{Y} = [0, 1]$.

In our current online setting, we need to make two continuity assumptions on the identification function V in relation to the Adversary's play. Our first assumption is a *weaker* version of the batch Assumption 4.1. Namely, we will assume that the Adversary's chosen distributions Y_t in every round t are such that $V(\cdot, Y_t)$ is Lipschitz in the Learner's prediction, but we *do not* assume anything about the magnitude of the individual Lipschitz constants in each round: only that they exist, and that their *average value* over all rounds is bounded by some L.

Assumption F.8 (Average Lipschitzness of V in the Learner's Action)**.** Assume that at each round t, the Adversary's label distribution Y_t is such that the identification function V for property Γ is L_t-Lipschitz for some $L_t < \infty$:

$$|V(\gamma, Y_t) - V(\gamma', Y_t)| \leq L_t|\gamma - \gamma'| \quad \text{for all } \gamma, \gamma'.$$

We make no assumption about the individual Lipschitz constants L_t, but *assume that their average value is bounded by L*:

$$\frac{1}{T}\sum_{t=1}^{T} L_t \leq L.$$

Our second continuity assumption on V requires that, holding the Learner's play fixed, V should be appropriately piecewise uniformly continuous in the Adversary's choice of label. Intuitively, this assumption's main function will be to ensure that the Adversary's action space is compact and finite-dimensional (thus satisfying the technical requirements of the online minimax multiobjective framework on the Adversary's action space), up to an arbitrarily small error.

Assumption F.9 (Either the Adversary Plays Finite-Support Distributions, or V Is Piecewise Uniformly Continuous in the Adversary's Action)**.** We assume that *either* the Adversary's distributions Y_t are finite-support in all rounds t, *or* otherwise the following condition on V must hold:

Fix any integer discretization parameter $m \geq 1$ for the Learner's play. Then, we assume that fixing any $\gamma \in [1/m]$, the function $V(\gamma, \cdot) : \mathcal{Y} \to \mathbb{R}$ is piecewise uniformly continuous in the Adversary's action, in the sense that there exist finitely many points $0 = \delta_1 < \delta_2 < \ldots < \delta_{M+1} = 1$ such that the function $V(\gamma, \cdot)$ is uniformly continuous on each subinterval (δ_i, δ_{i+1}) for all $i = 1 \ldots M$.

We now introduce our canonical algorithm, and prove its guarantees subject to Assumptions F.8 and F.9.

Algorithm 5 OnlineMulticalibration$(\mathcal{G}, V, m)$

Initialize an empty transcript $\pi^{\leq 0}$.
for rounds $t = 1, \ldots, T$ **do**
Observe the Adversary's chosen feature vector x_t.
Define the loss function $\ell^t : [1/m] \times \mathcal{G} \to \mathbb{R}^{|\mathcal{G}|}$ such that for each $G \in \mathcal{G}$:

$$\ell^t_G(\gamma_t, y_t) = \sum_{\gamma \in [1/m]} \mathbb{1}[x_t \in G, \gamma_t = \gamma] \cdot \frac{1}{n(\pi^{<t}, \gamma, G)} \left(2V(\gamma, y_t)R(\pi^{<t}, G, \gamma) + V(\gamma, y_t)^2\right),$$

where:

$$R(\pi^{<t}, G, \gamma) = \sum_{s<t : p_s = \gamma, x_s \in G} V(\gamma, y_s).$$

Let
$$\chi^t_G := \frac{\exp\left(\eta \sum_{s=1}^{t-1} \ell^s_G(p^s, y^s)\right)}{\sum_{G' \in \mathcal{G}} \exp\left(\eta \sum_{s=1}^{t-1} \ell^s_{G'}(p^s, y^s)\right)} \text{ for } G \in \mathcal{G}.$$

Let
$$P^t \in \operatorname*{argmin}_{P \in \Delta[1/m]} \max_y \sum_{G \in \mathcal{G}} \mathop{\mathbb{E}}_{p \sim P} \left[\chi^t_G \cdot \ell^t_G(p, y)\right].$$

Sample $p_t \sim P^t$ **and make prediction** p_t.
Observe the Adversary's selection of y_t.
Update the transcript $\pi^{\leq t} = \pi^{\leq t-1} \circ (x_t, p_t, y_t)$.
end for

Theorem F.10 (Algorithmic Guarantees for Sequential Multicalibration)**.** *Fix any finite collection of groups $\mathcal{G}$ and any bounded elicitable property Γ with $\text{Range}_\Gamma = [0, 1]$ and with a bounded identification function V satisfying $|V(\gamma, y)| \leq C$ for all γ, y. Suppose that the Adversary chooses a sequence of distributions that together with V satisfy Assumption F.8 with Lipschitz constant L, and Assumption F.9. Then for any $m > 0$, and any $T \geq \max\{\ln |\mathcal{G}|, 3m\}$, there is a randomized algorithm for the Learner (Algorithm 5) that chooses amongst m discrete predictions at every round and that (together with the Adversary) induces a transcript distribution after T rounds that produces a transcript satisfying α-approximate $(\mathcal{G}, V)$-multicalibration for:*

$$\mathop{\mathbb{E}}_{\pi}[\alpha] \leq \frac{2CL}{m} + \frac{2mC^2 \ln\left\lceil \frac{T}{m} \right\rceil}{T} + 12C^2\sqrt{\frac{\ln |\mathcal{G}|}{T}}.$$

Taking $m = \Theta\left(\sqrt{T}\right)$, this gives:

$$\mathop{\mathbb{E}}_{\pi}[\alpha] \leq O\left(\sqrt{\frac{\max\left\{\ln^2 T, \ln |\mathcal{G}|\right\}}{T}}\right).$$

Proof. We embed our learning problem into the online minimax optimization setting so that we can apply Theorem F.7. First, what are the Learner's and the Adversary's strategy spaces? We now define them, and show that both of these are convex, compact and finite dimensional sets as required.

At each round we let the Learner's strategy space be $\Delta[1/m]$, the simplex of probability distributions over predictions γ_t discretized at the granularity of $1/m$. This clearly is a convex, compact, and finite-dimensional space.

For the Adversary's strategy space, we would have wanted it to be the set of all probability distributions over $\mathcal{Y}$, but that would not be compact or finite-dimensional; so instead we assume that in each round t, the Adversary's strategy space is the (convex, compact and finite-dimensional) set of all distributions over $\mathcal{Y} = [0, 1]$ supported on at most some (arbitrary and unspecified) finite number N_t of points in $[0, 1]$. (To be clear, our algorithm below does not need to know N_t, nor will it be included in the performance bounds for it.) Recalling Assumption F.9, we easily see that this is without loss of generality, up to an arbitrarily small error term $\epsilon > 0$. Indeed, for any distribution $Y_t \in \Delta\mathcal{Y}$, this assumption lets us find, for any $\epsilon > 0$, a finite-support distribution $\tilde{Y}_t$ (supported over sufficiently many points) such that for any $\gamma \in [1/m]$, $|\mathbb{E}_{y\sim Y_t}[V(\gamma, y)] - \mathbb{E}_{y\sim \tilde{Y}_t}[V(\gamma, y)]| < \epsilon$. (Which would have its support consist of the points δ_i from the assumption, plus sufficiently many discretization points inside each interval (δ_i, δ_{i+1}).)

Next, we need to define the loss function ℓ^t used at each round. We take the dimension of the loss function to be $d = |\mathcal{G}|$ — with a coordinate devoted to each group $G \in \mathcal{G}$. Suppose at round t, the Adversary has chosen feature vector x_t (which, recall, is shown to the Learner before she must make a prediction). Then we define the loss vector ℓ^t as follows. For each $G \in \mathcal{G}$:

$$\ell^t_G(\mathrm{Alg}_t, Y_t) = \mathop{\mathbb{E}}_{\gamma_t\sim\mathrm{Alg}_t, y_t\sim Y_t} \sum_{i\in[m]} \left[\mathbb{1}\left[x_t \in G, \gamma_t = \frac{i}{m}\right] \cdot \frac{1}{n\left(\pi^{<t}, \frac{i}{m}, G\right)} \left(2V\left(\frac{i}{m}, y_t\right) R\left(\pi^{<t}, G, \frac{i}{m}\right) + V\left(\frac{i}{m}, y_t\right)^2\right)\right],$$

where $\mathrm{Alg}_t \in \Delta[1/m]$ is the distribution over predictions chosen by the Learner, and Y_t is the label distribution chosen by the Adversary. By linearity of expectation, this loss function is linear in the actions of both players, and so in particular is convex-concave as required. By the boundedness of V, the definition of R, and the fact that $\mathbb{1}\left[x_t \in G, \gamma_t = \frac{i}{m}\right] = 1$ for exactly one value of i, this loss function takes values in $[-C', C']$ as required, for $C' \le 3C^2$.

Next, we need to upper-bound the Adversary-Moves-First value of the game at round t:

$$w^A_t = \sup_{Y_t} \min_{\mathrm{Alg}_t\in\Delta[\frac{1}{m}]} \max_{G\in\mathcal{G}} \mathop{\mathbb{E}}_{\gamma_t\sim\mathrm{Alg}_t, y_t\sim Y_t} \left[\sum_{i\in[m]} \mathbb{1}\left[x_t \in G, \gamma_t = \frac{i}{m}\right] \cdot \frac{2V(\frac{i}{m}, y_t)R(\pi^{<t}, G, \frac{i}{m}) + V(\frac{i}{m}, y_t)^2}{n(\pi^{<t}, \frac{i}{m}, G)}\right].$$

To bound w^A_t, consider what the Learner should do if the Adversary goes first, revealing the true label distribution Y_t. The Learner then knows the true property value $\gamma^*_t = \Gamma(Y_t)$, so if she could play $\gamma_t = \gamma^*_t$, this would ensure that $V(\gamma_t, Y_t) = 0$, implying that

$$w^A_t = \frac{(V(\gamma^*_t, y_t))^2}{n(\pi^{<t}, \gamma^*_t, G)} \le \frac{C^2}{n(\pi^{<t}, \gamma^*_t, G)}.$$

The Learner cannot generally play γ^*_t (since it may not be a multiple of $1/m$ and hence not in her strategy space), but she can select the discrete point $\gamma_t \in [1/m]$ that is closest to γ^*_t (in particular, this γ_t will satisfy $|\gamma^*_t - \gamma_t| \le \frac{1}{m}$). With this action, the Learner will achieve 0 loss in all coordinates corresponding to groups G such that $x_t \notin G$ (since for each of these coordinates, the indicator $\mathbb{1}[x_t \in G, \gamma_t = \frac{i}{m}] = 0$ for all i). Thus, it remains to consider the coordinates corresponding to groups G such that $x_t \in G$. Let i be such that $\gamma_t = i/m$. Then, the indicator $\mathbb{1}[x_t \in G, \gamma_t = \frac{i}{m}] = 1$, so the loss value in this coordinate can be bounded as:

$$\mathop{\mathbb{E}}_{y_t\sim Y_t}\left[\frac{1}{n\left(\pi^{<t}, \frac{i}{m}, G\right)} \left(2V\left(\frac{i}{m}, y_t\right) R\left(\pi^{<t}, G, \frac{i}{m}\right) + V\left(\frac{i}{m}, y_t\right)^2\right)\right] \le \frac{2CL_t}{m} + \frac{C^2}{n\left(\pi^{<t}, \frac{i}{m}, G\right)},$$

where we used that $\mathbb{E}_{y_t\sim Y_t}[V(\gamma^*_t, y_t)] = 0$, that $|\gamma_t - \gamma^*_t| \le \frac{1}{m}$, and that $V(\cdot, Y_t)$ is L_t-Lipschitz by Assumption F.8.

This latter expression thus serves as an upper bound on the AMF value w^A_t. With this bound in hand, we can now apply

Theorem F.7 to conclude that Algorithm 4 obtains the following AMF regret bound:

$$\begin{aligned}
& \max_{G\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\gamma_t\sim \mathrm{Alg}_t, y_t\sim Y_t}\left[\frac{1}{n\left(\pi^{<t},\gamma_t,G\right)}\left(2V\left(\gamma_t,y_t\right)R\left(\pi^{<t},G,\gamma_t\right)+V\left(\gamma_t,y_t\right)^2\right)\right] \\
\leq\ & \max_{G\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T}\left(\frac{2CL_t}{m}+\frac{C^2}{n\left(\pi^{<t},\gamma_t,G\right)}\right)+12C^2\sqrt{\frac{\ln|\mathcal{G}|}{T}} \\
\leq\ & \frac{2CL}{m}+\max_{G\in\mathcal{G}} \frac{1}{T}\sum_{t=1}^{T}\frac{C^2}{n\left(\pi^{<t},\gamma_t,G\right)}+12C^2\sqrt{\frac{\ln|\mathcal{G}|}{T}} \\
=\ & \frac{2CL}{m}+\max_{G\in\mathcal{G}} \frac{C^2}{T}\left(\sum_{\gamma\in[1/m]}\sum_{t=1}^{n\left(\pi^{<T},\gamma,G\right)}\frac{1}{t}\right)+12C^2\sqrt{\frac{\ln|\mathcal{G}|}{T}} \\
\leq\ & \frac{2CL}{m}+\frac{C^2}{T}\left(m\sum_{t=1}^{\lceil T/m\rceil}\frac{1}{t}\right)+12C^2\sqrt{\frac{\ln|\mathcal{G}|}{T}} \\
\leq\ & \frac{2CL}{m}+\frac{2mC^2\ln\left\lceil\frac{T}{m}\right\rceil}{T}+12C^2\sqrt{\frac{\ln|\mathcal{G}|}{T}},
\end{aligned}$$

where the second inequality uses our Lipschitz Assumption F.8, the equality uses that $n(\pi^{\leq t}, \frac{i}{m}, G) = n(\pi^{<t}, \frac{i}{m}, G) + 1$ on any round t in which $x_t \in G$ and $\gamma_t = \frac{i}{m}$, and the second-to-last inequality uses that $\sum_{\gamma\in[1/m]} n(\pi^{<T}, \gamma, G) = T$.

By Lemma F.3, this implies that for all groups $G \in \mathcal{G}$, their miscalibration error is bounded as:

$$\mathbb{E}[K_2(G,\pi)] \quad = \quad \sum_{t=1}^{T}\mathbb{E}\left[K_2\left(G,\pi^{\leq t}\right)-K_2\left(G,\pi^{<t}\right)\right] \leq \frac{2CL}{m}+\frac{2mC^2\ln\left\lceil\frac{T}{m}\right\rceil}{T}+12C^2\sqrt{\frac{\ln|\mathcal{G}|}{T}},$$

which completes the proof. □

G. Formal Statements of Results in Section 6

G.1. Joint Multicalibration of Bayes Risks

Consider any Bayes pair (Γ, Γ^B) with respect to a strictly consistent scoring function $S(\gamma, y)$. As in Section 5, we assume that $\text{Range}_\Gamma \subseteq [0,1]$ and $\text{Range}_{\Gamma^B} \subseteq [0,1]$. To show that Bayes pairs are jointly multicalibratable, we will need to set up several assumptions on the scoring and identification functions associated with (Γ, Γ^B), in order to ensure the satisfaction of Assumptions 4.1, 5.2, 5.3, and 5.4 that the generic joint multicalibration result of Section 5 relies on.

To satisfy Assumption 4.1, we assume that the property Γ has an identification function V that is strictly increasing and L-Lipschitz in its first argument. To satisfy Assumption 5.2, we additionally assume that $V(\cdot, P)$ is L_a-anti-Lipschitz for $P \in \mathcal{P}$.

Now note that for all $\gamma \in \text{Range}_\Gamma$, the Bayes risk Γ^B by definition satisfies $\Gamma^B(P) = S(\gamma, P)$ on the level set $\{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ of Γ. As a result, the identification function for the Bayes risk Γ^B on the level set $\{P \in \mathcal{P} : \Gamma(P) = \gamma\}$ can be simply taken to be:

$$V^B_\gamma(\gamma^B, y) := \gamma^B - S(\gamma, y)$$

for all $\gamma^B, y \in [0,1]$. Taking the expectation over any $P \in \mathcal{P}$, we can thus write the expected conditional identification function of Γ^B conditioned on $\Gamma = \gamma$ as $V^B_\gamma(\gamma^B, P) := \gamma^B - S(\gamma, P)$.

To satisfy Assumption 5.3, we need to enforce the Lipschitzness of V^B_γ be Lipschitz with respect to its subscript γ. To do so, we assume that the scoring function S for the Bayes pair (Γ, Γ^B) is L_S-Lipschitz in its first argument. For any γ^B and any P, this lets us write $|V_\gamma(\gamma^B, P) - V_{\gamma'}(\gamma^B, P)| = |S(\gamma', P) - S(\gamma, P)| \leq L_S|\gamma - \gamma'|$, implying that V^B_γ is L_S-Lipschitz in γ.

Finally, we verify Assumption 5.4 of Section 5. Note that the identification function $V^B_\gamma(\cdot, P)$ for the Bayes risk Γ^B is well-defined for every $\gamma \in [0,1]$ and $P \in \mathcal{P}$, even when $\Gamma(P) \neq \gamma$. Furthermore, $V^B_\gamma(\gamma^B, P)$ is *linear* in γ^B *with slope* 1. Thus, $V^B_\gamma(\cdot, P)$ is strictly increasing and, in fact, 1-Lipschitz for $\gamma \in [0,1]$ and $P \in \mathcal{P}$, as desired.

With all requisite assumptions on the scoring and identification functions for Γ and Γ^B satisfied, we can now invoke Theorem 5.6 to obtain the following joint multicalibration guarantees for Bayes pairs:

Theorem G.1 (Bayes Pairs Are Jointly Multicalibratable)**.** *Consider any Bayes pair (Γ, Γ^B) with respect to a strictly consistent scoring function S. Let V be an identification function for Γ. Assume that: (1) The scoring function S is L_S-Lipschitz in its first argument; (2) V is strictly increasing, L-Lipschitz and L_a-anti-Lipschitz in its first argument.*

Pick a discretization factor $m \geq 1$. Set $\alpha^0 = \frac{4L^2}{m}$ and $\alpha^1 = \frac{4}{m}$. Let $\alpha^1_* = \frac{8}{m}((LL_aL_S)^2 + 1)$. Given any $\mathcal{G} \subseteq 2^{\mathcal{X}}$, instantiate `JointMulticalibration` *(Algorithm 2) using the id function V for Γ, and the id function collection V^B for Γ^B, such that $V^B_\gamma(\gamma^B, P) := \gamma^B - S(\gamma, P)$ for all $\gamma, \gamma^B \in [0,1], P \in \mathcal{P}$.*

Then, Algorithm 2 will output an $\left(\frac{4L^2}{m}, \frac{8}{m}((LL_aL_S)^2 + 1)\right)$-approximately $(\mathcal{G}, V, V^B)$-jointly multicalibrated predictor $f = (f^0, f^1)$ for (Γ, Γ^B), in at most $O\left(\frac{m^4}{L}\right)$ updates[5] to the joint predictor f.

G.2. Joint (Quantile, CVaR) Multicalibration

By itself, the CVaR is not sensible for calibration. Using our Theorem 3.4, this follows automatically from a well-known negative result of Gneiting (2011), who showed that CVaR_τ is not elicitable as it has nonconvex level sets for various simple distribution families $\mathcal{P}$.

Fact 3 (CVaR_τ has nonconvex level sets (Gneiting, 2011)). For any $\tau \in [0,1]$, CVaR_τ has nonconvex level sets relative to any class $\mathcal{P}$ of distributions over some interval $I \subseteq \mathbb{R}$ that includes the finite-support distributions, or the finite mixtures of compact-support distributions with well-defined PDF.

On the positive side, as an easy corollary of Theorem G.1, we obtain our next result that the pair (quantile, CVaR) can be jointly multicalibrated. To be able to apply Theorem G.1, it suffices to identify a strictly consistent scoring function S_τ for which the pair (τ-quantile, CVaR_τ) for any $\tau \in [0,1]$ is a Bayes pair, and then obtain the Lipschitz constant for S_τ, as well as the Lipschitz and anti-Lipschitz constants for a strictly increasing identification function V_τ for the τ-quantile.

And indeed, it is well-known (see e.g. Example 1 in Embrechts et al. (2021)) that (τ-quantile, CVaR_τ) is a Bayes pair for a scoring function S_τ that is the rescaled (by a factor of $\frac{1}{1-\tau}$) *pinball loss*:

Fact 4 ((τ-quantile, CVaR_τ) is a Bayes pair). Fix any $\tau \in [0,1]$ and let $\Gamma := q_\tau$ be a τ-quantile, and $\Gamma^B := \text{CVaR}_\tau$ be the τ-CVaR. Then (Γ, Γ^B) is a Bayes pair with respect to the strictly Γ-consistent scoring function S_τ defined, for all $\gamma, y \in [0,1]$, as:

$$S_\tau(\gamma, y) := \gamma + \frac{1}{1-\tau}(y - \gamma)_+,$$

where we have denoted $(u)_+ = \max\{u, 0\}$.

To bound the Lipschitz constant of S_τ, note that its derivative in the first argument is $\frac{\partial S_\tau(\gamma, y)}{\partial \gamma} = \mathbb{1}[y \leq \gamma] - \frac{\tau}{1-\tau}\mathbb{1}[y > \gamma]$. Thus S_τ has Lipschitz constant $L_{S_\tau} \leq \sup_{\gamma^*, y^*} \left|\frac{\partial S_\tau(\gamma^*, y^*)}{\partial \gamma}\right| = \max\{1, \frac{\tau}{1-\tau}\}$.

Now we need to settle on a strictly increasing (in the first argument) identification function V_τ for the τ-quantile q_τ and investigate its Lipschitz properties. Specifically, let us use the standard quantile id function defined as $V_\tau(\gamma, P) := \Pr_{y \sim P}[y \leq \gamma] - \tau$ for all γ and all $P \in \mathcal{P}$. Evidently, $V_\tau(\cdot, P)$ is just the CDF of P shifted by τ. Thus, by assuming that all distributions in $\mathcal{P}$ have a strictly increasing CDF, we ensure that V_τ is strictly increasing in γ.

To conveniently quantify the Lipschitzness of V_τ, assume that it is differentiable in γ: this is equivalent to all $P \in \mathcal{P}$ having a well-defined PDF pdf_P, which will then be the derivative of $V_\tau(\cdot, P)$: namely, $\frac{\partial V_\tau(\gamma, P)}{\partial \gamma} = pdf_P(\gamma)$. Therefore, enforcing a Lipschitz and an anti-Lipschitz constant on V_τ simply translates to assuming an upper and a lower bound on

[5] Specifically, Algorithm 2 will perform at most $R^- R^+ \frac{m^4}{L}$ updates on the predictor f, where we have denoted $R^- = \sup_{\gamma, y \in [0,1]} S(\gamma, y) - \inf_{\gamma, y \in [0,1]} S(\gamma, y)$ and $R^+ = \frac{1}{2}\max_{\gamma \in [1/m]}\left(\sup_{\gamma^B, y \in [0,1]} (\gamma^B - S(\gamma, y))^2 - \inf_{\gamma^B, y \in [0,1]} (\gamma^B - S(\gamma, y))^2\right)$.

the PDF of the distributions in the underlying family $\mathcal{P}$. Indeed, if we now assume that for all $P \in \mathcal{P}$, the PDF satisfies $0 < M_1 \leq pdf_P(y) \leq M_2 < \infty$ for all $y \in [0,1]$, this gives us that V_τ is M_2-Lipschitz and M_1-anti-Lipschitz.

Plugging the above Lipschitz and anti-Lipschitz bounds on S_τ and V_τ into Theorem G.1, we thus obtain the following joint (quantile, CVaR) multicalibration result:

Theorem G.2 (Joint Multicalibration of (τ-quantile, CVaR$_\tau$))**.** *Fix any constants $0 < M_1 < M_2$, and take any family $\mathcal{P}$ of probability distributions over $[0,1]$ such that each $P \in \mathcal{P}$ has a strictly increasing CDF and a well-defined density function pdf_P satisfying $M_1 \leq pdf_P(y) \leq M_2$ for all $y \in [0,1]$.*

Fix any target coverage level $\tau \in [0,1]$, and any group structure $\mathcal{G} \subseteq 2^{\mathcal{X}}$ on the dataset. Pick a discretization $m \geq 1$. Set $\alpha^0 = \frac{4M_2^2}{m}$ and $\alpha^1 = \frac{4}{m}$. Let $\alpha^1_* = \frac{8}{m}((M_1 M_2 \max\{1, \tau/(1-\tau)\})^2 + 1)$.

Then, by appropriately instantiating `JointMulticalibration` *(Algorithm 2), we can compute a*

$$\left(\frac{4M_2^2}{m}, \frac{8}{m}\left(\left(M_1 M_2 \max\left\{1, \frac{\tau}{1-\tau}\right\}\right)^2 + 1\right)\right) - \textit{approximately jointly } \mathcal{G}\textit{-multicalibrated predictor}$$

$f = (f^0, f^1)$ for the pair (τ-quantile, CVaR$_\tau$), after at most $O\left(\frac{m^4}{M_2}\right)$ updates to the joint predictor f.

G.3. Most Distortion Risk Measures Are Not Sensible for Calibration

We begin by formally stating the result of Kou & Peng (2016) and Wang & Ziegel (2015) that we will use. It shows that out of all distortion risk measures, the only ones that have convex level sets across the family of all finite-support distributions are: (1) means, (2) quantiles, and (3) two other risk measures which are quantile variants; here are the corresponding definitions.

Definition G.3. Consider any family $\mathcal{P}$ of probability distributions. For any distribution $P \in \mathcal{P}$, let its CDF (which need not be strictly increasing or continuous) be denoted F_P. We define the following distributional properties over $\mathcal{P}$:

1. For any $\tau \in [0,1]$, the τ-quantile is defined by:
$$q_\tau(P) = \inf\{y : F_P(y) \geq \tau\} \quad \text{for } P \in \mathcal{P}.$$
2. For any $\tau \in [0,1]$ and $c \in [0,1]$, define the property:
$$q^1_{\tau,c}(P) := c \cdot \inf\{y : F_P(y) \geq \tau\} + (1-c) \cdot \inf\{y : F_P(y) > \tau\} \quad \text{for } P \in \mathcal{P}.$$
3. For any $\tau \in [0,1]$ and $c \in [0,1]$, define the property:
$$q^2_{\tau,c}(P) := c \cdot \inf\{y : F_P(y) > 0\} + (1-c) \cdot \inf\{y : F_P(y) = 1\} \quad \text{for } P \in \mathcal{P}.$$

Observe that (1) $q^2_{\tau,c}$ is just a convex combination of the 0% quantile and the 100% quantile of the distribution; and (2) $q^1_{\tau,c}$ in fact *is* (for all $c \in [0,1]$) the τ-quantile subject to the CDF F_P being *strictly* increasing.

Kou & Peng (2016) showed that distribution means, together with the three (parametric) properties listed in Definition G.3, are the only distortion risk measures with convex level sets. The proof of this result was then simplified and refined by Wang & Ziegel (2015), who showed that it holds even over the family of distributions supported on at most 3 points.

Theorem G.4 (On Distortion Risk Measures with Convex Level Sets (Kou & Peng, 2016; Wang & Ziegel, 2015))**.** *Let $\mathcal{P}_3$ be the set of all probability distributions supported on at most 3 real-valued points. Let $\mathcal{P}_{bd}$ be the set of all bounded distributions over the reals with a well-defined PDF. Let $\mathcal{P}$ be any family of distributions over the reals such that either $\mathcal{P} \supseteq \mathcal{P}_3$, or $\mathcal{P} \supseteq \mathcal{P}_{bd}$.*

Consider any distortion risk measure $\Gamma : \mathcal{P} \to \mathbb{R}$. Then, Γ violates the convex level sets assumption on $\mathcal{P}$, unless it is one of the following distributional properties:

1. *The distributional mean;*

2. *A τ-quantile q_τ, for some $\tau \in [0,1]$;*
3. *The property $q^1_{\tau,c}$, for some $\tau, c \in [0,1]$;*
4. *The property $q^2_{\tau,c}$, for some $\tau, c \in [0,1]$.*

Now, our Theorem 3.4 lets us immediately conclude that for any $\mathcal{P}$ as in Theorem G.4, no distortion risk measure — other than means, quantiles, or the two parametric properties $q^1_{\tau,c}$ or $q^2_{\tau,c}$ — is sensible for calibration over any $\mathcal{P}$-compatible family of dataset distributions $\mathcal{D}$ that includes all the $\mathcal{P}$-compatible 2-point dataset distributions. To formally restate this:

Theorem G.5 (Sensibility for Calibration for Distortion Risk Measures)**.** *Let $\mathcal{P}_3$ be the set of all probability distributions supported on at most 3 real-valued points. Let $\mathcal{P}_{bd}$ be the set of all bounded distributions over the reals with a well-defined PDF. Let $\mathcal{P}$ be any convex space of distributions over the reals such that either $\mathcal{P} \supseteq \mathcal{P}_3$, or $\mathcal{P} \supseteq \mathcal{P}_{bd}$.*

Consider any distortion risk measure $\Gamma : \mathcal{P} \to \mathbb{R}$*, and any family $\mathcal{D}$ of $\mathcal{P}$-compatible dataset distributions that includes all the $\mathcal{P}$-compatible 2-point dataset distributions.*

Then Γ is not sensible for calibration over $\mathcal{D}$*, unless Γ is one of the following distributional properties:*

1. *The distributional mean;*
2. *A τ-quantile q_τ, for some $\tau \in [0,1]$;*
3. *The property $q^1_{\tau,c}$, for some $\tau, c \in [0,1]$;*
4. *The property $q^2_{\tau,c}$, for some $\tau, c \in [0,1]$.*