Moment Multicalibration for Uncertainty Estimation

Christopher Jung
University of Pennsylvania Department of Computer and Information Science

Changhwa Lee
University of Pennsylvania Department of Economics

Mallesh M. Pai
Rice University Department of Economics

Aaron Roth
University of Pennsylvania Department of Computer and Information Science

Rakesh Vohra
University of Pennsylvania Department of Economics and Electrical and Systems Engineering

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

We show how to achieve the notion of "multicalibration" from Hébert-Johnson et al. (2018) not just for means, but also for variances and other higher moments. Informally, this means that we can find regression functions which, given a data point, can make point predictions not just for the expectation of its label, but for higher moments of its label distribution as well—and those predictions match the true distribution quantities when averaged not just over the population as a whole, but also when averaged over an enormous number of finely defined subgroups. It yields a principled way to estimate the uncertainty of predictions on many different subgroups—and to diagnose potential sources of unfairness in the predictive power of features across subgroups. As an application, we show that our moment estimates can be used to derive marginal prediction intervals that are simultaneously valid as averaged over all of the (sufficiently large) subgroups for which moment multicalibration has been obtained.

1. Introduction

Uncertainty estimation is fundamental to prediction and regression. Given a training set of labelled points $D \subseteq \mathcal{X} \times [0, 1]$ consisting of feature vectors $x \in \mathcal{X}$ and labels $y \in [0, 1]$, the standard regression problem is to find a function $\overline{\mu} : \mathcal{X} \to [0, 1]$ that delivers a good point estimate of $\mu(x) = \mathbb{E}[y|x]$. We also desire the *variance* of the label distribution $\mathbb{E}[(y-\mu(x))^2|x]$ as a measure of the inherent uncertainty of a prediction. Higher central moments would yield even more information about this uncertainty which can be represented by *prediction intervals*: An interval $[\ell(x), u(x)]$ that with high probability contains y, i.e., $\Pr_y[y \in [\ell(x), u(x)]|x] \geq 1 - \delta$ for some $\delta \in (0, 1)$.

If the data are generated according to a parametric model as in the classic ordinary least squares setting, one can form confidence regions around the underlying model parameters, and translate these into both mean and uncertainty estimates about individual predictions. In non-parametric settings it is unclear how one should reason about uncertainty. We typically observe each feature vector x infrequently, and so we have essentially no information about the true distribution on y conditional on x. One solution to this problem is to compute *marginal* prediction intervals which average over *data points* x to give guarantees of the form: $\Pr_{x,y}[y \in [\ell(x), u(x)]] \geq 1 - \delta$. This

is the approach that is taken in the *conformal prediction* literature — see e.g. Shafer and Vovk (2008). The impossibility of obtaining conditional prediction intervals and the necessity to satisfy some relaxation is well known in this literature Lei and Wasserman (2014); Barber et al. (2020); Foygel Barber et al. (2021).

Marginal prediction intervals, unlike prediction intervals, do *not* condition on x. They offer a promise not over the randomness of the label conditional on the features, but over an average over data points. To make the distinction vivid, imagine one is a patient with high blood pressure, and a statistical model asserts that a certain drug will lower one's diastolic blood pressure to between 70 and 80 mm Hg. If $[70, 80]$ were a 95% prediction interval *conditional on all of one's observable features*, then one could reason that over the unrealized randomness in the world, there is a 95% chance that one's new blood pressure will lie in $[70, 80]$. If $[70, 80]$ is a 95% marginal prediction interval, however, it means that *95% of all patients who take the drug* will see their blood pressure decline to a level contained within the interval. Because the average is taken over a large, heterogeneous collection of people, the guarantee of the marginal prediction interval offers no meaningful promise to individuals. For example, it is possible that patients that share one's demographic characteristics (e.g. women of Sephardic Jewish descent with a family history of diabetes) will tend to see their blood pressure elevated by the drug. Estimating separate marginal predictions for different demographic groups does not solve this problem, because a single person can be a member of many medically relevant groups simultaniously, and there is no clear way to combine many (potentially very different) estimates. Foygel Barber et al. (2021) also study a similar problem – for more discussion, see our related work in Section 1.2.

This fundamental problem with uncertainty estimation in non-parametric settings is also a problem for mean estimation: what does it mean that a point prediction $\overline{\mu}(x)$ is an estimate of $\mathbb{E}[y|x]$ if we have no knowledge of the distribution on y conditional on x (because we have observed no samples from this distribution)? A standard performance measure is *calibration* (Dawid, 1982), which similarly averages over data points: a predictor $\overline{\mu}$ is calibrated (roughly) if $\mathbb{E}_{(x,y)}[\overline{\mu}(x) - y|\overline{\mu}(x) = i] = 0$ for all predictions i: i.e. for every i, conditioned on x being such that the prediction $\overline{\mu}(x)$ was (close to) i, the expected outcome y is also (close to) i. Just as with marginal prediction intervals, guarantees of calibration mean little to individuals, who differ substantially from the majority of people over whom the average is taken.

Hébert-Johnson et al. (2018) proposed *multicalibration* as a way to interpolate between the (unattainable) ideal of being able to correctly predict $\mathbb{E}[y|x]$ for each x and offering a guaranteed averaged over the entire data distribution. The idea is to fix a large, structured set of (possibly overlapping) sub-populations ($\mathcal{G} \in 2^X$). A predictor $\overline{\mu}$ is multicalibrated if, informally, for all predictions i and groups $G \in \mathcal{G}$, $\mathbb{E}_{(x,y)}[\overline{\mu}(x) - y|\overline{\mu}(x) = i, x \in G] = 0$. Thus, $\overline{\mu}$ is calibrated not just on the overall population, but also simultaneously on many different finely defined sub-populations that one might care about (e.g. different demographic groups). Hébert-Johnson et al. (2018) show how to compute an approximately multicalibrated predictor $\overline{\mu}$ on all subgroups in $\mathcal{G}$ that have substantial probability mass—we provide a high level description of their algorithm, which we use, below.

The main contribution of this paper is to show how to achieve what can loosely be termed multicalibration for higher moment estimates. We provide not just estimates $\overline{\mu}(x)$ of means ($\mu(x) = \mathbb{E}[y|x]$), but also estimates, $\overline{m}_k(x)$, for higher central moments, ($m_k(x) = \mathbb{E}[(y - \mu(x))^k|x]$) such that all of these forecasts are appropriately multicalibrated in a sense made precise below. This is useful for a number of basic tasks. One we briefly highlight is that it can help diagnose data

iniquities: for example, if the set of collected features is much less predictive of the target label on certain demographic groups $G \in \mathcal{G}$ this will necessarily manifest itself in multicalibrated moment predictions by having higher variance predictions on individual members of those populations.

As an important application, we show that standard concentration inequalities which could be applied using the true moments of a distribution to obtain prediction intervals can also be applied using our multicalibrated moment estimates. Doing so produces intervals $[\ell(x), u(x)]$ for each data point that are *simultaneously* valid marginal prediction intervals not just overall, but also conditioned on x lying in any of the (sufficiently large) subgroups over which we are multicalibrated. This allows one to interpret these prediction intervals as predicting something meaningful not just an average over all people, but — simultaneously — as averages over all of the people who were given the same prediction, across many finely defined subgroups (like women of Sephardic Jewish descent with a family history of diabetes). Note that because the groups $G \in \mathcal{G}$ may overlap, a single individual can belong to many such groups, and can at her option interpret the prediction interval as averaging over any of them.

1.1. Overview of Our Approach and Results

1.1.1. MEAN MULTICALIBRATION AND IMPEDIMENTS TO EXTENSIONS TO HIGHER MOMENTS

We first review the algorithm of Hébert-Johnson et al. (2018), recast in the framework in which we will conduct our analysis. We here elide some issues such as how we deal with discretization and how calibration error is parameterized — see Section 2 for the formal model and definitions. Fix a feature space $\mathcal{X}$, labels $\mathcal{Y} = [0, 1]$, and an unknown distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. Given are sets $\mathcal{G} \subseteq 2^{\mathcal{X}}$, corresponding to sub-populations of interest. The goal is to construct a predictor, $\overline{\mu} : \mathcal{X} \to \mathcal{Y}$, that is multicalibrated, i.e. calibrated on each group $G \in \mathcal{G}$. This means that we want a predictor, $\overline{\mu}$, that is *mean-consistent* on every set of the form $G(\overline{\mu}, i) = \{x \in G : \overline{\mu}(x) = i\}$ for some i: in other words, for every such set $G(\overline{\mu}, i)$ we want $\mathbb{E}_{(x,y) \sim \mathcal{P}}[\overline{\mu}(x) - y]|x \in G(\overline{\mu}, i)] = 0$. We describe the algorithm as if it has direct access to the true distribution $\mathcal{P}$, and defer for now a description of how to implement the algorithm using a finite sample.

It is helpful to conceive of the task as a zero-sum game between two players: a "(mean) consistency player", and an "audit player" who knows the true distribution $\mathcal{P}$. The consistency player chooses a predictor $\overline{\mu}$, and the audit player, given a predictor, attempts to identify a subset S of $\mathcal{X}$ on which the predictor is not mean consistent.[1] Given a pair of choices, the corresponding cost (which the consistency player wishes to minimize and the audit player wishes to maximize) is the absolute value difference between the average prediction of the consistency player and the average expected label on the subset S identified by the audit player. The value of this game is 0, since the consistency player can obtain perfect consistency using the true conditional label distribution $\overline{\mu}(x) = \mathbb{E}[y|x]$. The algorithm of Hébert-Johnson et al. (2018) can be interpreted as solving this zero sum game by simulating repeated play, using online gradient descent for the consistency player, and "best response" for an audit player, who stops play if there are no remaining sets $S = G(\overline{\mu}, i)$ witnessing violations of multicalibration. This works because by linearity of expectation, we can formulate the game so that the consistency player's utility function is *linear* in her individual predictions $\overline{\mu}(x)$. A formal description and proof of correctness can be found in Section 3.1.

---

1. Here, and in what follows, we adopt the convention that G refers to a group in $\mathcal{G}$, while S refers to any generic subset of $\mathcal{X}$.

There are two—related—impediments to extending this approach to higher moments, i.e., finding predictors $\overline{m}_k(x) \approx m_k(x) = \mathbb{E}[(y - \mathbb{E}[y|x])^k|x]$, that are "consistent" with $\mathcal{P}$ on many sets. The first of these is definitional—what do we mean by "consistent" for higher moments? The second is algorithmic—given a definition, how do we achieve it? Both are impediments because, unlike means, higher moments are not linear functionals of the distribution. A consequence is that moments for $k > 1$ do not combine linearly in the way expectations do. In particular for $S = S_1 \cup S_2$ where S_1 and S_2 are disjoint, $\mathbb{E}[(y - \mathbb{E}[y|x \in S])^k|S] \neq \Pr(x \in S_1|S)\mathbb{E}[(y - \mathbb{E}[y|x \in S_1])^k|S_1] + \Pr(x \in S_2|S)\mathbb{E}[(y - \mathbb{E}[y|x \in S_2])^k|S_2]$. It is therefore silly to require that moment predictions $\overline{m}_k(x)$ satisfy the same "average consistency" condition asked of means: i.e. we cannot demand that the population variance on the subset of the population on which we predict variance v be v, because this is not a property that the true moments $m_k(x)$ satisfy. Consider, for example, a setting in which there are two types of points, x_1 and x_2. The true distribution is uniform over $\{(x_1, 0), (x_2, 1)\}$ (and so in particular the label y is deterministically fixed by the features). We have that for all $k > 1$, $\mu(x_1) = 0, \mu(x_2) = 1$, and $m_k(x_0) = m_k(x_1) = 0$. Nevertheless, the variance over the set of points on which the true distribution satisfies $m_k(x) = 0$ is $1/4$, not 0. We cannot ask that our "moment calibrated" predictors satisfy properties violated by the true distribution, because we would have no guarantee of feasibility — and our ultimate goal in multicalibration is to find a set of mean and moment predictors that are indistinguishable from the true distribution with respect to some class of tests.

1.1.2. MEAN CONDITIONED MOMENT MULTICALIBRATION AND MARGINAL PREDICTION INTERVALS

A key observation (Observation 2.1) is that higher moments *do* linearize over sets that have the same mean: in other words, if we have $S = S_1 \cup S_2$ for disjoint S_1 and S_2 such that $\mathbb{E}[y|x \in S_1] = \mathbb{E}[y|x \in S_2]$, then, it follows that $\mathbb{E}[(y - \mathbb{E}[y|x \in S])^k|S] = \Pr(x \in S_1|S)\mathbb{E}[(y - \mathbb{E}[y|x \in S_1])^k|S_1] + \Pr(x \in S_2|S)\mathbb{E}[(y - \mathbb{E}[y|x \in S_2])^k|S_2]$. An implication of this is that the true distribution does satisfy what we term *mean-conditioned moment multicalibration*. Namely, if for a fixed $k > 1$ we define for each set $G \in \mathcal{G}$ and each pair of mean and k^{th} moment values i, j the sets: $G(\mu, m_k, i, j) = \{x \in G : \mu(x) = i, m_k(x) = j\}$, then we have *both* mean consistency: $\mathbb{E}[(y - i)|x \in G(\mu, m_k, i, j)] = 0$ *and* moment consistency: $\mathbb{E}[(y - i)^k - j|x \in G(\mu, m_k, i, j)] = 0$ over these sets. Therefore, we require the same condition to hold for our mean and moment predictors $\overline{\mu}$ and $\{\overline{m}_a\}_{a=1}^k$: namely that simultaneously for every a, that over each of the sets $G(\overline{\mu}, \overline{m}_a, i, j)$, the true label mean should be i and the true label a-th moment should be j. In other words, if we have a set of predictors that are mean conditioned moment multicalibrated, then an individual who receives a particular mean and (e.g.) variance prediction can be assured that amongst all the people who received the same mean and variance prediction *even averaged over any of the possibly large number of sub-groups G of which the individual is a member*, the true mean and variance are faithful to the prediction.

Appendix F demonstrates a key application of mean-conditioned moment-multicalibrated estimators: They can be used in place of real distributional moments to derive prediction intervals. Given moments of a random variable X, a standard way to derive concentration inequalities for X is by using the following inequality for any even moment (for $k = 2$ this is Cheybychev's inequality): $\Pr[|X - \mu(X)| \geq t] \leq \frac{\mathbb{E}[(X - \mu(X))^k]}{t^k}$.

If X is the label distribution conditional on features x, this yields the prediction interval:

$$\Pr_{y}\left[y \in \left[\mu(x) - \left(\frac{m_k(x)}{\delta}\right)^{1/k}, \mu(x) + \left(\frac{m_k(x)}{\delta}\right)^{1/k}\right]\middle| x\right] \geq 1 - \delta.$$

In Appendix F we show that if we have a mean-conditioned moment-multicalibrated pair $(\overline{\mu}, \overline{m}_k)$, we can replace the true mean and moments in the derivation of this prediction interval, and get *marginal* prediction intervals, which are valid not just averaged over all points, but simultaneously as averaged over all point that received the same prediction within any of the groups within which we are mean-conditioned moment multicalibrated. In other words, for all $G \in \mathcal{G}$ and for all i, j:

$$\Pr_{(x,y)\sim P}\left[y \in \left[\overline{\mu}(x) - \left(\frac{\overline{m}_k(x)}{\delta}\right)^{1/k}, \overline{\mu}(x) + \left(\frac{\overline{m}_k(x)}{\delta}\right)^{1/k}\right]\middle| x \in G(\overline{\mu}, \overline{m}_k, i, j)\right] \geq 1 - \delta.$$

1.1.3. ACHIEVING MEAN CONDITIONED MOMENT MULTICALIBRATION

What is the difficulty with finding sets of predictors $(\overline{\mu}, \{\overline{m}_a\}_{a=2}^k)$ such that simultaneously each pair $(\overline{\mu}, \overline{m}_a)$ are mean-conditioned moment multicalibrated? It is that moments do not have the linear structure that means do. Hence, the zero-sum game formulation we describe for mean-multicalibration cannot be applied directly. A naïve approach (which fails, but which will be a useful sub-routine for us) is to first train a mean-multicalibrated predictor $\overline{\mu}$, and then define "pseudo-moment" labels for each x as $\widetilde{m}_{k,\overline{\mu}}(x) = (y - \overline{\mu}(x))^k$. Since these are constant values, we can then use the algorithm for mean multicalibration to achieve "pseudo-moment calibration with respect to $\overline{\mu}$" — i.e. mean consistency on each set $G(\overline{\mu}, \overline{m}_k, i, j)$ with respect to our pseudo-moment labels $\widetilde{m}_{k,\overline{\mu}}(x)$. By itself this doesn't guarantee any sort of "moment consistency," but we show in Section 3.2 that if we can:

1. Find moment predictors $\overline{m}_k$ that satisfy pseudo-moment calibration with respect to $\overline{\mu}$, *and*

2. Our mean predictor $\overline{\mu}$ satisfies mean consistency on every set of the form $G(\overline{\mu}, \overline{m}_k, i, j)$,

then, the pair $(\overline{\mu}, \overline{m}_k)$ will satisfy mean-conditioned moment multicalibration.

The difficulty is that these two requirements are circularly defined. Once we have a *fixed* mean predictor $\overline{\mu}$, we can use a gradient descent procedure to find moment predictors $\{\overline{m}_a\}_{a=2}^k$ that are pseudo-calibrated with respect to $\overline{\mu}$. However, we also require our mean predictor to be mean consistent on the sets $G(\overline{\mu}, \overline{m}_a, i, j)$, which are undefined until we fix our moment predictors $\{\overline{m}_a\}_{a=2}^k$. Section 3.3 resolves the circularity by using an alternating descent procedure that toggles between updating $\overline{\mu}$ and $\{\overline{m}_a\}_{a=2}^k$, each aiming for a mean calibration target that is defined with respect to the other. We prove that this alternating gradient descent procedure is guaranteed to converge after only a small number of rounds.

Finally, we show in Appendix D how to implement our algorithm using a finite sample from the distribution and furnish sample complexity bounds, in a way analogous to Hébert-Johnson et al. (2018). The sample complexity bounds are logarithmic in the number of groups $|\mathcal{G}|$ that we wish to be multicalibrated with respect to, and polynomial in our desired calibration error parameters and the number of moments k with which we wish to achieve mean-conditioned moment multicalibrated predictors. In particular, because dependence on $|G|$ is only logarithmic, we can satisfy mean-conditioned moment-multicalibration on an exponentially large collection of intersecting sets

$\mathcal{G}$ from just a polynomial sample of data from the unknown population distribution. Note, however, that despite our polynomial dependence on k, the natural scale of the k'th moment decreases exponentially in k, and so to obtain non-trivial approximation guarantees for k'th moments with polynomial sample complexity, we should think of taking k at most logarithmic in the relevant parameters of the problem. See Theorem 20 and Corollary 21 for details. Our running time scales polynomially with our approximation error parameters, the number of moments k with which we wish to be multicalibrated, and the running time of solving learning problems over $\mathcal{G}$ (which is at most linear in $|\mathcal{G}|$, but can be much faster). See Theorems 20 and 27 for details. In other words, our algorithms are "oracle efficient" in the sense that if we have a subroutine for solving learning problems over $\mathcal{G}$, then we can use it to solve mean-conditioned moment-multicalibration problems with at most polynomial overhead. In theory, for almost every interesting class $\mathcal{G}$, learning over $\mathcal{G}$ is hard in the worst case — but oracle efficiency has proven to be a useful paradigm in the design of learning algorithms (especially in the fairness in machine learning literature — see e.g. (Agarwal et al., 2018; Kearns et al., 2018; Hébert-Johnson et al., 2018; Kim et al., 2019)) because in practice we have extremely powerful heuristics for solving complex learning problems. Moreover, this kind of oracle efficiency is the best running time guarantee that we can hope for, because as shown by Hébert-Johnson et al. (2018), even mean-multicalibration is as hard as solving arbitrary learning problems over $\mathcal{G}$ in the worst case.

1.2. Additional Related Work

Calibration as a means of evaluating forecasts of expectations dates back to Dawid (1982). This literature focuses on a simple online forecasting setting, motivated by weather prediction problems: in a sequence of rounds, nature chooses the probability of some binary event (e.g. rain), and a forecaster predicts a probability of that event. Dawid (1982) shows that a Bayesian forecaster will always be subjectively calibrated (i.e. he will believe himself to be calibrated). Foster and Vohra (1998) show that there exist *randomized* forecasters that can asymptotically satisfy calibration against arbitrary sequences of outcomes (this is impossible for deterministic forecasters (Oakes, 1985)). These papers focus on the online setting, because simple calibration is trivial in a batch/distributional setting: simply predicting the mean outcome on every point satisfies calibration. Within this literature, the most related works are Lehrer (2001) and Sandroni et al. (2003), which give very general asymptotic results that are able to achieve (mean) multicalibration as a special case. Lehrer (2001), operating in the sequential online setting, asks for calibration to hold not just on the entire sequence of realized outcomes, but on countably many infinite *sub-sequences* (e.g. the set of all computable subsequences). He proves that there exists an online forecasting algorithm which can asymptotically achieve this. Sandroni et al. (2003) extend this result to subsequences which can be defined in terms of the forecasters predictions as well. Both of these papers operate in a setting that is general enough to encode the constraint of mean multicalibration (by encoding the features of datapoints in the "state space") even in an online, adversarial setting — albeit not in a computationally or sample efficient way. In contrast, Hébert-Johnson et al. (2018), who define the notion of mean multicalibration, give an algorithm for achieving it in a batch distributional setting — in a much more computationally and sample efficient manner than could have been achieved by applying the machinery of Lehrer (2001); Sandroni et al. (2003). Recently, Zhao et al. (2020) gave a notion of "individual level" (mean) calibration, defined over the randomness of the forecaster, that is valid conditional on individual data points (i.e. without needing to average over a population).

They provide promising empirical results, but the theoretical guarantees of predictors satisfying this notion do not provide non-trivial information about a data distribution, because (as the authors note) their notion of individual calibration can be satisfied without observing any data.

Hébert-Johnson et al. (2018) also proposed the notion of "multi-accuracy," a weaker notion than multicalibration which asks for a predictor $\overline{\mu}$ that satisfies mean consistency on each set $G \in \mathcal{G}$, but not on sets $G(\overline{\mu}, i)$. Kim et al. (2019) gave a practical algorithm for achieving multi-accuracy, and a promising set of experiments suggesting that it could be used to correct for error disparities between different demographic groups on realistic data sets, without sacrificing overall accuracy. Dwork et al. (2019) propose notions of fairness and evidence consistency for ranking individuals by their "probability of success" when historical data only records binary outcomes: they show that their proposed notions are closely related to multicalibration of the probability predictions implicitly underlying the rankings. Shabat et al. (2020) prove uniform convergence bounds for multicalibration error over hypothesis classes of bounded complexity. In our paper, as in Hébert-Johnson et al. (2018), we learn over hypothesis classes that are only implicitly defined by the set of groups $\mathcal{G}$, and so we bound generalization error in the same manner that Hébert-Johnson et al. (2018) do, rather than using uniform convergence arguments.

Conformal prediction is similarly motivated to calibration, but is focused on finding marginal prediction intervals rather than mean estimates: see e.g. Shafer and Vovk (2008) for an overview of this literature. Finding marginal prediction intervals on its own (i.e. when prediction intervals only have to be valid on average over the entire population) is easy in the batch/distributional setting, and so just as with the calibration literature, the conformal prediction literature is primarily focused on the online setting in which predictions must be made as points arrive. The most closely related paper in this literature is Foygel Barber et al. (2021) who study the batch distributional (with a more general label space than considered in this paper), and also aim to find marginal prediction intervals with coverage guarantees which hold not just over the entire population, but on a collection $\mathcal{G}$ of more finely defined sub-populations. Foygel Barber et al. (2021) obtain prediction intervals of this sort by using a holdout set method from conformal prediction: roughly speaking, they compute empirical $1 - \delta$ coverage intervals on each set $G \in \mathcal{G}$ in the holdout set, and then for an individual x, select the *widest* such interval amongst all groups G that contain x. In general, this is very conservative, but they show that under certain structural assumption on $\mathcal{G}$, the width of the interval can be bounded. In contrast, our methods always give tight estimates of variance and other higher moments without any assumption on the group structure, and our prediction intervals are tight insofar as moment based bounds are tight (which is to say, tight in the worst case, but not instance optimal). The algorithm given by Foygel Barber et al. (2021) relies on explicit enumeration of groups $G \in \mathcal{G}$ over the holdout set.

There are also several papers in the "fairness in machine learning" literature (in addition to Hébert-Johnson et al. (2018); Kim et al. (2019)), which are similarly motivated by replacing coarse statistical constraints with constraints that come closer to offering individual guarantees: see Chouldechova and Roth (2020) for a survey. Kearns et al. (2018, 2019) propose to learn classifiers which equalize statistical measures of harm like false positive or negative rates across a very large number of demographic subgroups $G \in \mathcal{G}$, and give practical algorithms for this problem by solving a zero-sum game formulation using techniques from no-regret learning. Kim et al. (2018) give algorithms for satisfying a notion of metric fairness which similarly enforces constraints averaged over a large number of subgroups $G \in \mathcal{G}$. Rothblum and Yona (2018) define a PAC-like version of the individual fairness notion of Dwork et al. (2012) and prove generalization bounds showing how to achieve

their notion out of sample on all sufficiently large groups of individuals. Sharifi-Malvajerdi et al. (2019) show how to equalize statistical measures of harm like false positive rates across *individuals* — when the rates in question are defined over the randomness of the problem distribution and the classifier. Joseph et al. (2016, 2018) propose an individual-level notion of "weakly meritocratic fairness" that can be satisfied in bandit learning settings whenever it is possible to compute confidence or prediction intervals around individual labels. They analyze the parametric setting, when actual (conditional) prediction and confidence intervals are possible — but the techniques from our paper could be used for learning in the assumption-free setting (with a slightly weaker notion of fairness) using marginal prediction intervals.

2. Preliminaries

Let $\mathcal{X}$ be the domain of features, $\mathcal{Y} = [0, 1]$ the label domain, and $\mathcal{P}$ the true (unknown) probability distribution over $\mathcal{X} \times \mathcal{Y}$.[2] Let $\mathcal{P}_{\mathcal{X}}$ refer to the induced marginal distribution on $\mathcal{X}$ and define $\mathcal{P}_{\mathcal{Y}}$ analogously. Going forward, we refer to the associated random variables with capital letters (e.g. X, Y), and realizations with lowercase letters (x, y).

Let $\mathcal{G} \subseteq 2^{\mathcal{X}}$ be a collection of subsets of $\mathcal{X}$,[3] and for each $G \in \mathcal{G}$, let χ_G denote that associated indicator function, i.e. $\chi_G(x) = 1 \iff x \in G$. For implementation purposes, we assume that each indicator function $\chi_G(x)$ can be computed by a polynomially sized circuit.[4]

Definition 1 *Given the true distribution $\mathcal{P}$, we write $\mu = \mathbb{E}_{\mathcal{P}}[y]$, and its kth central moment is: $m_k = \mathbb{E}_{\mathcal{P}}\left[(y - \mu)^k\right]$. Given a set $S \subseteq \mathcal{X}$, we abuse notation and write $\mu(S) = \mathbb{E}_{\mathcal{P}}[y|x \in S]$ and $m_k(S) = \mathbb{E}_{\mathcal{P}}\left[(y - \mu(S))^k | x \in S\right]$ for the conditional mean and k^{th} central moment of labels on the distribution conditional on $x \in S$.*

We are given n independent draws from $\mathcal{X} \times \mathcal{Y}$ according to distribution $\mathcal{P}$, denoted $D = \{(x_b, y_b)\}_{b=1}^n$. The goal is to predict means and higher moments of $\mathcal{Y}|\mathcal{X}$, i.e. to construct functions $\overline{\mu} : \mathcal{X} \to [0, 1]$, (we shall refer to this as a mean predictor) and $\overline{m}_k : \mathcal{X} \to [0, 1]$ (analogously, k^{th}-moment predictor)—as $\mathcal{Y}$ is the unit interval, means and moments also lie in the unit interval.

To define calibration, we need to reason about all points that receive a particular prediction. For real valued predictors, this can be a measure zero set. One solution is to to restrict attention to predictors that are discretized to lie on the grid $G_m = \{\frac{1}{2m}, \frac{3}{2m}, \ldots, \frac{2m-1}{2m}\}$, for some (large) number m. If one were to do this, the discretization parameter m would be coupled to the error one could ultimately obtain: since it may be inevitable to suffer error at least $1/2m$ if one is restricted to making predictions on a discrete grid. Alternately, one can define calibration by "bucketing" real valued predictions into m buckets of width $\frac{1}{m}$ each. This allows us to treat m (a parameter controlling the fineness of our calibration constraint) as an orthogonal parameter to our calibration

---

2. Our approach applies for both finite and infinite feature domains. If $\mathcal{X}$ is uncountably infinite, define an associated measure space, and $\mathcal{P}$ is a countably additive probability measure on this space. We omit the associated notation since it will have no use in what follows.

3. If $\mathcal{X}$ is uncountably infinite, then $\mathcal{G}$ is a collection of measurable, computable sets. We abuse notation and write $2^{\mathcal{X}}$ to denote this.

4. Our algorithm in the end will need to manipulate these indicator functions. We might imagine e.g. that $\mathcal{G}$ is the hypothesis class of some learning algorithm for a binary prediction problem, and that the functions $\chi_G(x)$ are particular hypotheses from this class — e.g. linear threshold functions.

error. To that end, given a set $S \subseteq \mathcal{X}$, mean predictor $\overline{\mu}$, and some $i \in [m]$, define

$$S(\overline{\mu}, i) \equiv \left\{ x \in S : \left| \overline{\mu}(x) - \frac{2i-1}{2m} \right| \leq \frac{1}{2m} \right\}$$

to be the set of points in S whose mean predictions fall into the ith bucket, i.e. $[\frac{2i-1}{2m} - \frac{1}{2m}, \frac{2i-1}{2m} + \frac{1}{2m}]$. Analogously, define

$$S(\overline{\mu}, \overline{m}_k, i, j) \equiv \left\{ x \in S : \left| \overline{\mu}(x) - \frac{2i-1}{2m} \right| \leq \frac{1}{2m}, \left| \overline{m}_k(x) - \frac{2j-1}{2m} \right| \leq \frac{1}{2m} \right\}$$

to be the set of points in S that receive mean predictions in the ith bucket *and* k^{th} moment predictions in the jth bucket. Given mean and k^{th} moment predictors $\overline{\mu}$ and $\overline{m}_k$, and any set $S \subseteq \mathcal{X}$ we write

$$\overline{\mu}(S) = \mathbb{E}_{\mathcal{P}}[\overline{\mu}(x) | x \in S] \quad \text{and} \quad \overline{m}_k(S) = \mathbb{E}_{\mathcal{P}}[\overline{m}_k(x) | x \in S],$$

i.e. $\overline{\mu}(S)$ is the average mean prediction of $\overline{\mu}$ when x's are drawn according to the true distribution, $\mathcal{P}$, conditional on $x \in S$, and $\overline{m}_k(S)$ is the analogous quantity for k'th moment predictions.

To be clear, we will maintain the convention for means and higher moments that quantities with bars refer to predictions ($\overline{\mu}, \overline{m}_k$) and unmodified notation (μ, m_k) refer to true (unknown) population values.

Definition 2 (Consistency) *Call a mean predictor $\overline{\mu}$ (α, ϵ)-mean consistent on a set S if $|\mu(S) - \overline{\mu}(S)| \leq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} + \epsilon$. Similarly, a moment predictor $\overline{m}_k$ is called (α, ϵ)-moment consistent on a set S if $|m_k(S) - \overline{m}_k(S)| \leq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} + \epsilon$.*

When $\epsilon = 0$, we say $\overline{\mu}$ is α-mean consistent and $\overline{m}_k$ is α-moment consistent. Note that (α, ϵ)-mean consistency implies $(\alpha + \epsilon)$-mean consistency.

Remark 3 *Our notion of consistency on a set S corresponds to error that smoothly degrades with the size (measure) of the set S. This is essential to giving out of sample guarantees. Hébert-Johnson et al. (2018) handles this slightly differently, by giving uniform guarantees, but only for sets that have measure at least γ. Our approach of giving smoothly parameterized error guarantees for all sets is only stronger (up to a reparameterization of $\alpha \leftarrow \alpha\gamma$), and makes the analysis of our algorithms more transparent because it corresponds more directly to the guarantees they achieve.*

The following simple observation will be useful in understanding our approach.

Observation 2.1 *Let $\mathcal{P}$ be a mixture distribution over m component distributions $\mathcal{P}_\ell$ with mixture weights $w_\ell \geq 0$, $\sum_{\ell=1}^m w_\ell = 1$. Let $\mu_\ell, m_{k\ell}$ be the mean and k^{th} moment associated with $\mathcal{P}_\ell$. Then:*

$$m_k = \sum_{\ell=1}^m w_\ell \left(\sum_{a=0}^k \binom{k}{a} (\mu_\ell - \mu)^{k-a} m_{a\ell} \right).$$

If the mixture variables have the same mean, i.e. $\mu_\ell = \mu$ for all ℓ, then, the above expression reduces to: $m_k = \sum_{\ell=1}^m w_\ell m_{k\ell}$

Observation 2.1 highlights the key challenge: unlike means, higher moments combine non-linearly over mixtures. That is to say, that although $\overline{m}_k(S)$ is defined to be an average over the values $\overline{m}_k(x)$ for $x \in S$, $m_k(S)$ is *not* an average over the values $m_k(x)$ for $x \in S$ for $k > 1$. Observation 2.1 also makes clear what we are trying to exploit in defining mean-*conditioned* moment calibration: $m_k(S)$ *is* an average over the values $m_k(x)$ for $x \in S$ whenever $\mu(x)$ is constant over S.

We are now ready to define calibration, which asks for mean and moment consistency on particular sets defined by the mean and moment predictors themselves:

Definition 4 (Calibration) *Fix a set $S \subseteq \mathcal{X}$ and a true distribution $\mathcal{P}$.*

1. *A mean predictor $\overline{\mu}$ is (α, ϵ)-mean calibrated on a set S if it is (α, ϵ)-mean consistent on every set $S(\overline{\mu}, i)$, i.e. if for each $i \in [m]$:*

$$|\mu\left(S(\overline{\mu}, i)\right) - \overline{\mu}(S(\overline{\mu}, i))| \leq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S(\overline{\mu}, i))} + \epsilon.$$

   Again, if $\epsilon = 0$, we say $\overline{\mu}$ is α-mean calibrated.

2. *Predictors $(\overline{\mu}, \overline{m}_k)$ are $(\alpha, \beta, \epsilon)$-mean-conditioned-moment calibrated on a set S if they are (α, ϵ)-mean and (β, ϵ)-moment consistent on every set $S(\overline{\mu}, \overline{m}_k, i, j)$, i.e. if for every $i, j \in [m]$:*

$$|\mu\left(S(\overline{\mu}, \overline{m}_k, i, j)\right) - \overline{\mu}\left(S(\overline{\mu}, \overline{m}_k, i, j)\right)| \leq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S(\overline{\mu}, \overline{m}_k, i, j))} + \epsilon,$$

$$\text{and } |m_k\left(S(\overline{\mu}, \overline{m}_k, i, j)\right) - \overline{m}_k\left(S(\overline{\mu}, \overline{m}_k, i, j)\right)| \leq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(S(\overline{\mu}, \overline{m}_k, i, j))} + \epsilon.$$

   If $\epsilon = 0$, we say $(\overline{\mu}, \overline{m}_k)$ are (α, β)-mean-conditioned-moment calibrated.

We say that $\overline{\mu}, \overline{m}_k$ are (α, ϵ)-multicalibrated and $(\alpha, \beta, \epsilon)$-mean-conditioned-moment multicalibrated with respect to (a collection of sets) $\mathcal{G}$ if they are (α, ϵ)-mean calibrated and $(\alpha, \beta, \epsilon)$-mean conditioned moment calibrated respectively on every $G \in \mathcal{G}$.

Remark 5 *Observe that by construction, the true feature conditional mean and moment functions $\mu(x), m_k(x)$ are mean-conditioned-moment multicalibrated on every collection of sets G. We can view the goal of multicalibration as coming up with mean and moment predictors $\overline{\mu}, \overline{m}_k$ that are almost indistinguishable from the true distributional means and moments, with respect to a class of consistency checks defined by $\mathcal{G}$. Note that it is only because we have defined our goal as mean conditioned moment calibration that the true moments $m_k(x)$ of the distribution satisfy these consistency conditions, which are defined as expectations.*

We highlight the difference between calibration and consistency on a given set S in terms of mean prediction $\overline{\mu}$; an analogous discussion applies to higher moments. Consistency requires that the prediction $\overline{\mu}(x)$, averaged over x's in S according to the conditional distribution, approximately equals the true label average $\mu(S)$. It doesn't impose a similar requirement on subsets of S. Therefore, a predictor consistent on S will be correct on average for the set S but could be systematically biased for each prediction in S that it makes.

Calibration on S requires, for every prediction $i \in [m]$, that $\overline{\mu}$ is consistent on the set $S(\overline{\mu}, i)$. That is to say it ensures consistency on every subset of x's in S on which the predictor $\overline{\mu}$ makes predictions in some some fixed bucket i. Exact calibration implies exact consistency, but the reverse is not true.

3. Achieving Mean Conditioned Moment Multicalibration

In this section, we show how to construct a mean predictor that is mean multicalibrated and mean and moment predictors that are mean-conditioned-moment multicalibrated with respect to some given collection of groups $\mathcal{G}$. For the sake of clarity, we assume in this section that distributional quantities such as conditional mean $\mu(S)$ and conditional k^{th} moment $m_k(S)$ are known and have a perfect learning oracle for $\mathcal{G}$ – this allows us to present the core algorithmic ideas without the technical details of approximation, which are intricate but largely standard (i.e. Chernoff bounds). However, as these distributional quantities must be in fact estimated from data, we show in Appendix D how to generalize all the main results in this section to a setting when only finite sample from $\mathcal{P}$ is available.

3.1. Mean Multicalibration

We summarize an algorithm to achieve mean multicalibration. It is a modest extension to the one in Hébert-Johnson et al. (2018) that accommodates arbitrary distributions over a possibly infinite domain and arbitrary initializations. We present it in somewhat greater generality than needed for mean-calibration, because our final algorithm in Section 3.3 needs to achieve mean consistency on more sets than are required for mean calibration alone.

For intuition, consider the following mini-max problem, which captures a more difficult problem than mean multicalibration (as there is no restriction at all on the sets S):

$$\min_{\overline{\mu}:\mathcal{X}\to[0,1]} \quad \max_{\substack{S\subseteq\mathcal{X},\\ \lambda\in\{-1,1\}}} \quad \lambda \cdot \mathcal{P}_{\mathcal{X}}(S) \cdot (\overline{\mu}(S) - \mu(S)) \, .$$

We can associate a zero-sum game with this mini-max problem by viewing the minimization player as a *consistency* player who must commit to a mean predictor $\overline{\mu}$, and viewing the maximization player as an *auditor* who attempts to identify sets S on which the consistency player fails to be mean consistent. Observe that the inclusion of the measure term $\mathcal{P}_{\mathcal{X}}(S)$ in the objective makes the learner's utility function linear in her individual predictions $\overline{\mu}(x)$. There is a strategy for the consistency player that would guarantee her a payoff of 0—or in other words, would guarantee consistency on all possible sets S: she could simply set $\overline{\mu}(x) = \mathbb{E}[y|x]$. This establishes the value of the game, but of course it requires knowledge of $\mathcal{P}$. Given only a finite sample of the data, we will be unable to determine $\mathbb{E}[y|x]$ for all x, and so this strategy is not implementable.

One way to solve our problem absent knowledge of the distribution is to allow the consistency player to play online gradient descent (Zinkevich, 2003) on the set of mean predictors over rounds t, and to allow the auditor to "best respond" at every round, by exhibiting a set S corresponding to a large consistency violation[5]. This is guaranteed to converge quickly to an approximate equilibrium of the game: i.e. a mean predictor satisfying approximate consistency on all sets. If the auditor limits herself to choosing sets $S(\overline{\mu}^t, i)$ corresponding to mean calibration, then we converge quickly to approximate mean calibration. Here we give a direct analysis of a general gradient descent procedure of the sort we need, in terms of the sets that the auditor happens to choose during this interaction. For finite support distributions $\mathcal{P}$, this bound could be derived directly from the regret

---

5. Because the objective function of our game weights the consistency violations $\overline{\mu}(S) - \mu(S)$ by the measure of the set $\mathcal{P}_{\mathcal{X}}(S)$, these violations are linear functions of the individual predictions $\overline{\mu}(x)$. Thus it suffices to run gradient descent over the space of individual predictions $\mathcal{X}$, rather than the space of all possible functions $\overline{\mu} : \mathcal{X} \to [0, 1]$.

bound of online projected gradient descent (Zinkevich, 2003) or from the analysis of the similar algorithm in Hébert-Johnson et al. (2018). We reproduce a direct analysis in the Appendix to match the theorem statement we want for distributions which may have infinite support. (Note that for such distributions the mean predictor will have to be maintained implicitly). In Algorithm 1, after each gradient update, we project $\overline{\mu}^t$ back into the set of functions with range $[0, 1]$ using an ℓ_2 projection. Because squared ℓ_2 distance is linearly separable, it can be accomplished by a simple coordinate-wise operation which we write as $\text{project}_{[0,1]}(x) = \min(\max(x, 0), 1)$.

Start with an arbitrary initial mean predictor $\overline{\mu}^1 : \mathcal{X} \to [0, 1]$

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

  Audit player plays some $S^t \subseteq \mathcal{X}, \lambda^t \in \{-1, 1\}$

  $\overline{\mu}^{t+1}(x) = \begin{cases} \text{project}_{[0,1]} \left(\overline{\mu}^t(x) - \eta\lambda^t \right) & \text{if } x \in S^t, \\ \overline{\mu}^t(x) & \text{otherwise.} \end{cases}$

Algorithm 1: Projected Gradient Descent(η) for $\overline{\mu}$

Lemma 6 *For any initial mean predictor $\overline{\mu}^1 \in \mathcal{X} \to [0, 1]$ and any sequence of $(S^t, \lambda^t)_{t=1}^T$, Algorithm 1 satisfies: $\sum_{t=1}^T \lambda^t \mathcal{P}_{\mathcal{X}}(S^t) \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) \leq \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathcal{P}_{\mathcal{X}}(S^t)$.*

The proof is in the Appendix. A direct consequence of the bound in Lemma 6 is that, when interacting with a consistency player who uses gradient descent with learning rate $\eta = \alpha$, an auditor will be able to find sets that fail to be α-mean consistent for at most $1/\alpha^2$ many rounds.[6] The following theorem is a direct consequence of Lemma 6 — its short proof is in the Appendix.

Theorem 7 *Set $T = \frac{1}{\alpha^2} - 1$ and $\eta = \alpha = \frac{1}{\sqrt{T+1}}$ in Algorithm 1. Assume that for every $t \in [T]$, $\lambda^t \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S^t)}$, Then, for every $S \subseteq \mathcal{X}$, we have*

$$\left| \overline{\mu}^{T+1}(S) - \mu(S) \right| \leq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)}.$$

In particular, if the auditor selects sets $G(\overline{\mu}^t, i)$ that fail to satisfy approximate mean consistency whenever they exist, then we quickly converge to a mean-multicalibrated predictor. Either we reach a state in which $\overline{\mu}^t$ is approximately mean consistent on every set $G(\overline{\mu}^t, i)$ before T rounds, in which case we are done, or after T rounds, the conclusion of Theorem 7 implies not only that we are approximately mean-multicalibrated with respect to $\mathcal{G}$, but that we are approximately mean-consistent on every set.

3.2. Pseudo-Moment Consistency

In this section we make a simple observation: Algorithm 1 from Section 3.1 for achieving mean consistency and calibration did not depend on *any* properties of the labels y. It would have worked equally well had we invented an arbitrary label for each datapoint x, and asked for mean consistency with respect to that label. Using this observation, we consider a (naïve, and incorrect) attempt at

---

6. A somewhat better bound is achievable by using a non-uniform learning rate that depends on the measure of the sets S^t chosen by the auditor; we use a uniform learning rate throughout this paper for clarity.

achieving calibration for higher moments — but one that will be a useful subroutine in our final algorithm. Recall that $m_k(S) = \mathbb{E}[(y - \mu(x))^k | x \in S]$. If we have a mean predictor $\overline{\mu}(x)$, it is therefore tempting to imagine that each point x is associated with an alternative label $\tilde{y}(x) = \widetilde{m}_{k,\overline{\mu}}(x)$, where: $\widetilde{m}_{k,\overline{\mu}}(x) = \mathbb{E}\left[(y - \overline{\mu}(x))^k \Big| x\right]$. We could then use the algorithm from Section 3.1 to construct an predictor $\overline{m}_k$ that was *mean* multicalibrated with respect to these labels. We refer to the property of being mean consistent with respect to the moment-like labels $\widetilde{m}_{k,\overline{\mu}}(x)$ as "pseudo-moment-consistency":

Definition 8 (Pseudo-Moment-Consistency) *Fixing a mean predictor $\overline{\mu}$, define the kth pseudo-moment labels to be $\widetilde{m}_{k,\overline{\mu}}(x) = \mathbb{E}\left[(y - \overline{\mu}(x))^k \Big| x\right]$. A moment predictor $\overline{m}_k$ is (β, ϵ)-pseudo-moment-consistent on a set S, with respect to a mean predictor $\overline{\mu}$ if $|\overline{m}_k(S) - \widetilde{m}_{k,\overline{\mu}}(S)| \leq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(S)} + \epsilon$. We simply say β-pseudo-moment consistent if the predictor is $(\beta, 0)$-pesudo-moment-consistent.*

We can achieve pseudo-moment consistency using the gradient descent procedure of Algorithm 2 in the appendix which is essentially analogous to Algorithm 1. In particular, we obtain the following theorem, whose proof is deferred to the Appendix.

Theorem 9 *Let $T = \frac{1}{\beta^2} - 1$ and $\eta = \frac{1}{\sqrt{T+1}} = \beta$ in Algorithm 2, and fix any mean predictor $\overline{\mu}$, which defines the function $\widetilde{m}_{k,\overline{\mu}}(x)$. Assume that for every $t \in [T]$, $\left|\overline{m}_k^t(R^t) - \widetilde{m}_{k,\overline{\mu}}(R^t)\right| \geq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(R^t)}$. Then, for every $R \subseteq \mathcal{X}$, we have $|\overline{m}_k(R) - \widetilde{m}_{k,\overline{\mu}}(R)| \leq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(R)}$ – that is $\overline{m}_k$ is β-pseudo-moment-consistent on every set R.*

Now, a guarantee of "pseudo-moment-consistency" is really a guarantee of *mean* consistency with respect to "moment-like" labels $\widetilde{m}_{k,\overline{\mu}}(x)$, and does *not* correspond to moment consistency. This is because moments m_k for $k > 1$ don't combine linearly the way means do: recall Observation 2.1. But also recall from Observation 2.1 that higher moments *do* happen to combine linearly if we average only over points that share the same mean.

We take advantage of this to prove the following key lemma: if we achieve pseudo-moment consistency on all sets $G(\overline{\mu}, \overline{m}_k, i, j)$ (for $G \in \mathcal{G}$) with respect to a mean predictor $\overline{\mu}$ that happens also to be mean-consistent on all sets $G(\overline{\mu}, \overline{m}_k, i, j)$, then, the pair of predictors is in fact approximately mean-conditioned moment multicalibrated with respect to $\mathcal{G}$.

Lemma 10 *Assume $\overline{\mu}$ is such that for all $G \in \mathcal{G}$ and $i, j \in [m]$, $\overline{\mu}$ is α-mean consistent on every set $G(\overline{\mu}, \overline{m}_k, i, j))$:*

$$|\overline{\mu}(G(\overline{\mu}, \overline{m}_k, i, j)) - \mu(G(\overline{\mu}, \overline{m}_k, i, j))| \leq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(G(\overline{\mu}, \overline{m}_k, i, j))}.$$

Assume also that $\overline{m}_k$ is β-pseudo-moment-consistent with respect to $\overline{\mu}$ on every set $G(\overline{\mu}, \overline{m}_k, i, j))$ for $G \in \mathcal{G}$ and $i, j \in [m]$:

$$|\overline{m}_k(G(\overline{\mu}, \overline{m}_k, i, j)) - \widetilde{m}_{k,\overline{\mu}}(G(\overline{\mu}, \overline{m}_k, i, j))| \leq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(G(\overline{\mu}, \overline{m}_k, i, j))}.$$

Then, for every $G \in \mathcal{G}$, $i, j \in [m]$, we have

$$|\overline{m}_k(G(\overline{\mu}, \overline{m}_k, i, j)) - m_k(G(\overline{\mu}, \overline{m}_k, i, j))| \leq \frac{\beta + k\alpha}{\mathcal{P}_{\mathcal{X}}(G(\overline{\mu}, \overline{m}_k, i, j)} + \frac{k}{m}.$$

This implies in particular that $(\overline{\mu}, \overline{m}_k)$ are $(\alpha, \beta', \epsilon)$-mean-conditioned moment multicalibrated with respect to $\mathcal{G}$, for $\beta' = \beta + k\alpha$ and $\epsilon = \frac{k}{m}$.

Lemma 10 reduces the problem of finding mean-conditioned-moment multicalibrated predictors $(\overline{\mu}, \overline{m}_k)$ to the problem of finding a pair of predictors $(\overline{\mu}, \overline{m}_k)$ satisfying mean-consistency and pseudo-moment-consistency on the sets $G(\overline{\mu}, \overline{m}_k, i, j)$. It is unclear how to do this, because these conditions have a circular dependency: pseudo-moment consistency of $\overline{m}_k$ with respect to $\overline{\mu}$ is not defined until we have fixed a mean predictor $\overline{\mu}$, because the "labels" $\widetilde{m}_{k,\overline{\mu}}(x)$ with respect to which pseudo-moment consistency is defined depend on $\overline{\mu}$. On the other hand, the sets $G(\overline{\mu}, \overline{m}_k, i, j)$ on which $\overline{\mu}$ must satisfy mean consistency are not defined until we fix the moment predictor $\overline{m}_k$. The next section is devoted to resolving this circularity and finding predictors satisfying the conditions of Lemma 10.

3.3. Mean-Conditioned Moment Multicalibration

We arrive at the last block upon which our main result rests: an alternating gradient descent procedure that on any distribution finds a mean multicalibrated predictor $\overline{\mu}$ together with moment predictors $\{\overline{m}_a\}_{a=2}^k$ such that each pair $(\overline{\mu}, \overline{m}_a)$ is approximately mean-conditioned moment multicalibrated on $\mathcal{G}$. We continue, for clarity's sake, to assume access to the underlying distribution $\mathcal{P}$, and postpone to Appendix D the details of implementing this approach with a polynomially sized sample of points. Our strategy is to obtain a set of predictors that together satisfy the hypotheses of Lemma 10: mean consistency and pseudo-moment-consistency on every set of the form $G(\overline{\mu}, \overline{m}_a, i, j)) \subseteq G \in \mathcal{G}, 1 < a \le k$, and $i, j \in [m]$. We have already seen in Section 3.1 that for a *fixed* collection of sets, a simple gradient-descent procedure can obtain mean consistency on each of the sets. Section 3.2 demonstrates that for a *fixed* mean predictor $\overline{\mu}$, a similar gradient descent procedure can obtain pseudo-moment-consistency with respect to $\overline{\mu}$ on each set $G(\overline{\mu}, \overline{m}_a, i, j))$. Our algorithm simply alternates between these two procedures. In rounds t, we maintain hypothesis predictors $\overline{\mu}^t, \{\overline{m}_a^t\}_{a=2}^k$. In alternating rounds, we perform updates of gradient descent using Algorithm 3 to arrive at a mean predictor $\overline{\mu}^t$ that has taken a step towards consistency on sets $G(\overline{\mu}^t, \overline{m}_a^{t-1}, i, j)$, and then using the newly updated mean predictor $\overline{\mu}^t$, run Algorithm 4 to obtain moment predictors $\overline{m}_a^t$ that obtain pseudo-moment-consistency with respect to $\overline{\mu}^t$ on all sets $G(\overline{\mu}^t, \overline{m}_a^t, i, j)$. This is coordinated via a wrapper algorithm, Algorithm 5. We prove this alternating procedure terminates after $1/\alpha^2 - 1$ many rounds and outputs predictors $\overline{\mu}, \{\overline{m}_a\}_{a=2}^k$ that are jointly mean-conditioned moment-multicalibrated. The actual details of Algorithm 3, 4, and 5 can be found in the appendix.

Theorem 11 *Let T be the final iterate t of Algorithm 5 (i.e. its output is $(\overline{\mu}^T, \{\overline{m}_a^T\}_{a=2}^k)$. Algorithm 5 has the following guarantees:*

1. **Total Iterations**: *The algorithm halts. The final iterate T is s.t. $T \le \frac{1}{\alpha^2} - 1$. The total number of gradient descent update operations is at most $\left(\frac{1}{\alpha^2} - 1\right)\left(1 + (k-1)\left(\frac{1}{\beta^2} - 1\right)\right)$.*

2. **Mean multicalibration**: *Output $\overline{\mu}^T$ is α-mean multicalibrated with respect to $\mathcal{G}$.*

3. **Mean Conditioned Moment multicalibration**: *For every $a \in \{2, \dots, k\}$, the pair $(\overline{\mu}^T, \overline{m}_a^T)$ is $(\alpha, \beta + a\alpha, \frac{a}{m})$-mean-conditioned moment-multicalibrated with respect to $\mathcal{G}$.*

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M Wallach. A reductions approach to fair classification. In *ICML*, 2018.

Rina Foygel Barber et al. Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*, 14(2):3487–3524, 2020.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

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

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 106–125. IEEE, 2019.

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

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 61–70. IEEE, 2010.

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

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Meritocratic fairness for infinite and contextual bandits. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 158–163, 2018.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.

Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852, 2018.

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

Ehud Lehrer. Any inspection is manipulable. *Econometrica*, 69(5):1333–1347, 2001.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 71–96, 2014.

David Oakes. Self-calibrating priors do not exist. *Journal of the American Statistical Association*, 80(390):339–339, 1985.

Thomas K Philips and Randolph Nelson. The moment bound is tighter than chernoff's bound for positive tail probabilities. *The American Statistician*, 49(2):175–178, 1995.

Guy Rothblum and Gal Yona. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688, 2018.

Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. Calibration with many checking rules. *Mathematics of operations Research*, 28(1):141–153, 2003.

Jeanette P Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff–hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2):223–250, 1995.

Eliran Shabat, Lee Cohen, and Yishay Mansour. Sample complexity of uniform convergence for multicalibration. *arXiv preprint arXiv:2005.01757*, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average individual fairness: Algorithms, generalization and experiments. In *Advances in Neural Information Processing Systems*, pages 8242–8251, 2019.

Peng-Jun Wan, Ding-Zhu Du, Panos Pardalos, and Weili Wu. Greedy approximations for minimum submodular cover with submodular cost. *Computational Optimization and Applications*, 45(2): 463–474, 2010.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. *arXiv preprint arXiv:2006.10288*, 2020.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Appendix A. Details and Proofs from Section 3.1

Lemma 12 *For any initial mean predictor $\overline{\mu}^1 \in \mathcal{X} \to [0,1]$ and any sequence of $(S^t, \lambda^t)_{t=1}^T$, Algorithm 1 satisfies: $\sum_{t=1}^T \lambda^t \mathcal{P}_{\mathcal{X}}(S^t) \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) \leq \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathcal{P}_{\mathcal{X}}(S^t)$.*

Proof Because $\overline{\mu}^{t+1}(x) = \overline{\mu}^t(x)$ for $x \notin S^t$, we can lower bound the "progress" made towards μ at each round $t \in [T]$ as:

$$
\begin{aligned}
& \mathbb{E}_{\mathcal{P}} \left[(\overline{\mu}^t(x) - \mu(x))^2 - (\overline{\mu}^{t+1}(x) - \mu(x))^2 \right] \\
& = \mathcal{P}_{\mathcal{X}}(S^t) \mathbb{E}_{\mathcal{P}} \left[(\overline{\mu}^t(x) - \mu(x))^2 - (\overline{\mu}^{t+1}(x) - \mu(x))^2 | x \in S^t \right] \\
& \geq \mathcal{P}_{\mathcal{X}}(S^t) \mathbb{E}_{\mathcal{P}} \left[(\overline{\mu}^t(x) - \mu(x))^2 - (\overline{\mu}^t(x) - \eta\lambda^t - \mu(x))^2 | x \in S^t \right] \\
& = \mathcal{P}_{\mathcal{X}}(S^t) \mathbb{E}_{\mathcal{P}} \left[(\overline{\mu}^t(x) - \mu(x))^2 - \left((\overline{\mu}^t(x) - \mu(x))^2 - 2\eta\lambda^t(\overline{\mu}^t(x) - \mu(x)) + (\eta\lambda^t)^2 \right) | x \in S^t \right] \\
& = \mathcal{P}_{\mathcal{X}}(S^t) \mathbb{E}_{\mathcal{P}} \left[2\eta\lambda^t(\overline{\mu}^t(x) - \mu(x)) | x \in S^t \right] - \mathcal{P}_{\mathcal{X}}(S^t)(\eta\lambda^t)^2 \\
& = 2\eta\lambda^t \mathcal{P}_{\mathcal{X}}(S^t) \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) - \mathcal{P}_{\mathcal{X}}(S^t)(\eta\lambda^t)^2.
\end{aligned}
$$

The inequality would be an equality if we did not project $\overline{\mu}^t$ into the range $[0,1]$. Performing the projection only decreases its ℓ_2 distance to μ, which yields the inequality. Rearranging terms and observing that $(\lambda^t)^2 = 1$ yields

$$
\lambda^t \mathcal{P}_{\mathcal{X}}(S^t) \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) \leq \frac{1}{2\eta} \mathbb{E}_{\mathcal{P}} \left[(\overline{\mu}^t(x) - \mu(x))^2 - (\overline{\mu}^{t+1}(x) - \mu(x))^2 \right] + \frac{\eta \mathcal{P}_{\mathcal{X}}(S^t)}{2}
$$

Therefore we have that

$$
\begin{aligned}
& \sum_{t=1}^T \lambda^t \mathcal{P}_{\mathcal{X}}(S^t) \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) \\
& \leq \sum_{t=1}^T \left(\frac{1}{2\eta} \mathbb{E}_{\mathcal{P}} \left[(\overline{\mu}^t(x) - \mu(x))^2 - (\overline{\mu}^{t+1}(x) - \mu(x))^2 \right] + \frac{\eta \mathcal{P}_{\mathcal{X}}(S^t)}{2} \right) \\
& = \frac{1}{2\eta} \mathbb{E}_{\mathcal{P}} \left[(\overline{\mu}^1(x) - \mu(x))^2 - (\overline{\mu}^{T+1}(x) - \mu(x))^2 \right] + \frac{\eta}{2} \sum_{t=1}^T \mathcal{P}_{\mathcal{X}}(S^t) \\
& \leq \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \mathcal{P}_{\mathcal{X}}(S^t)
\end{aligned}
$$

as desired. The last inequality follows because $\overline{\mu}^1(x), \mu(x)$, and $\overline{\mu}^{T+1}(x)$ all fall in $[0,1]$. ∎

Theorem 7 *Set $T = \frac{1}{\alpha^2} - 1$ and $\eta = \alpha = \frac{1}{\sqrt{T+1}}$ in Algorithm 1. Assume that for every $t \in [T]$, $\lambda^t \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S^t)}$, Then, for every $S \subseteq \mathcal{X}$, we have*

$$\left| \overline{\mu}^{T+1}(S) - \mu(S) \right| \leq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)}.$$

Proof Fix any set $S \subseteq \mathcal{X}$ and imagine extending the sequence by setting $S^{T+1} = S$ and setting $\lambda^{T+1} = \text{sign}(\overline{\mu}^{T+1}(S) - \mu(S))$. By Lemma 6, we would then have:

$$\sum_{t=1}^{T+1} \lambda^t \mathcal{P}_{\mathcal{X}}(S^t) \left(\overline{\mu}^t(S^t) - \mu(S^t) \right) \leq \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T+1} \mathcal{P}_{\mathcal{X}}(S^t)$$

$$\leq \frac{1}{2\eta} + \frac{\eta(T+1)}{2}$$

$$\leq \sqrt{T+1} \qquad \qquad \text{(substituting } \eta = \frac{1}{\sqrt{T+1}})$$

We can then peel off the last term in the sum corresponding to $S^{T+1} = S$ to obtain:

$$\mathcal{P}_{\mathcal{X}}(S) \left| \overline{\mu}^{T+1}(S) - \mu(S) \right| \leq \sqrt{T+1} - \sum_{t=1}^{T} \lambda^t \mathcal{P}_{\mathcal{X}}(S^t) \left(\overline{\mu}^t(S^t) - \mu(S^t) \right)$$

$$\leq \sqrt{T+1} - \alpha T \qquad \qquad \text{(by assumption)}$$

$$= \alpha \qquad \qquad \text{(since } T = \frac{1}{\alpha^2} - 1)$$

which completes the proof. ∎

Appendix B. Details and Proofs from Section 3.2

For intuition, we can think of the pseudo-moment calibration algorithm as playing the following zero sum game using projected online gradient descent against an adversary who plays best responses. Recall that $\overline{\mu}$ is a fixed quantity so that $\widetilde{m}_{k,\overline{\mu}}$ is well defined.

$$\min_{\overline{m}_k} \max_{\substack{R \subseteq \mathcal{X} \\ \psi \in \{-1,1\}}} \psi \mathcal{P}_{\mathcal{X}}(R) \left(\overline{m}_k(R) - \widetilde{m}_{k,\overline{\mu}}(R) \right).$$

Lemma 13 *For any arbitrary $\overline{m}_k^1 : \mathcal{X} \to [0,1]$ and any sequence of $(R^t, \psi^t)_{t=1}^T$, we have that*

$$\sum_{t=1}^{T} \psi^t \mathcal{P}_{\mathcal{X}}(R^t) \left(\overline{m}_k^t(R) - \widetilde{m}_{k,\overline{\mu}}(R) \right) \leq \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \mathcal{P}_{\mathcal{X}}(R^t)$$

Proof

$$\mathbb{E}_{\mathcal{P}} \left[(\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 - (\overline{m}_k^{t+1}(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 \right]$$

$$= \mathcal{P}_{\mathcal{X}}(R^t) \mathbb{E}_{\mathcal{P}} \left[(\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x_i))^2 - (\overline{m}_k^{t+1}(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 | x \in R^t \right]$$

$$\geq \mathcal{P}_{\mathcal{X}}(R^t) \mathbb{E}_{\mathcal{P}} \left[(\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 - (\overline{m}_k^t(x) - \eta \psi^t - \widetilde{m}_{k,\overline{\mu}}(x))^2 | x \in R^t \right]$$

$$= \mathcal{P}_{\mathcal{X}}(R^t) \mathbb{E}_{\mathcal{P}} \left[(\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 - \left((\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 - 2\eta\psi^t(\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x)) + (\eta\psi^t)^2 \right) | x \in R^t \right]$$

$$= 2\eta\psi^t \mathcal{P}_{\mathcal{X}}(R^t) \mathbb{E}_{\mathcal{P}} \left[\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x)) | x \in R^t \right] - \mathcal{P}_{\mathcal{X}}(R^t)(\eta\psi^t)^2$$

18

Here the inequality comes from the fact that projection can only make the ℓ_2 norm smaller. Rearranging terms and observing that $(\psi^t)^2 = 1$ yields

$$\psi^t \mathcal{P}_{\mathcal{X}}(R^t)\mathbb{E}_{\mathcal{P}}\left[\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x)|x \in R^t\right] \leq \frac{1}{2\eta}\mathbb{E}_{\mathcal{P}}\left[(\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 - (\overline{m}_k^{t+1}(x) - \widetilde{m}_{k,\overline{\mu}}(x_i))^2\right] + \frac{\eta\mathcal{P}_{\mathcal{X}}(R^t)}{2}.$$

Plugging this inequality back into the regret, we get

$$\sum_{t=1}^{T}\psi^t\mathcal{P}_{\mathcal{X}}(R^t)\left(\overline{m}_k^t(R) - \widetilde{m}_{k,\overline{\mu}}(R)\right)$$

$$= \sum_{t=1}^{T}\psi^t\mathcal{P}_{\mathcal{X}}(R^t)\mathbb{E}_{\mathcal{P}}\left[\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x)|x \in R^t\right]$$

$$\leq \sum_{t=1}^{T}\left(\frac{1}{2\eta}\mathbb{E}_{\mathcal{P}}\left[(\overline{m}_k^t(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 - (\overline{m}_k^{t+1}(x) - \widetilde{m}_{k,\overline{\mu}}(x_i))^2\right] + \frac{\eta\mathcal{P}_{\mathcal{X}}(R^t)}{2}\right)$$

$$= \frac{1}{2\eta}\left(\sum_{x\in\mathcal{X}}\mathbb{E}_{\mathcal{P}}\left[(\overline{m}_k^1(x) - \widetilde{m}_{k,\overline{\mu}}(x))^2 - (\overline{m}_k^{T+1}(x) - \widetilde{m}_{k,\overline{\mu}}(x_i))^2\right]\right) + \frac{\eta}{2}\sum_{t=1}^{T}\mathcal{P}_{\mathcal{X}}(R^t)$$

$$\leq \frac{1}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathcal{P}_{\mathcal{X}}(R^t)$$

as desired. The last inequality follows because $\overline{m}_k^1(x), \widetilde{m}_{k,\overline{\mu}}(x)$, and $\overline{m}_k^{T+1}(x)$ all fall in $[0,1]$. ∎

Start with an *arbitrary* initial pseudo-moment predictor $\overline{m}_k^1 : \mathcal{X} \to [0,1]$
for $t = 1, \ldots, T$ **do**
　Audit player plays some $R^t \subseteq \mathcal{X}, \psi^t \in \{-1, 1\}$
　$\overline{m}_k^{t+1}(x) = \begin{cases} \text{project}_{[0,1]}\left(\overline{m}_k^t(x) - \eta\psi^t\right) & \text{if } x \in R^t, \\ \overline{m}_k^t(x) & \text{otherwise.} \end{cases}$
Algorithm 2: Projected Gradient Descent(η) for $\overline{m}_k$

Theorem 9 *Let $T = \frac{1}{\beta^2} - 1$ and $\eta = \frac{1}{\sqrt{T+1}} = \beta$ in Algorithm 2, and fix any mean predictor $\overline{\mu}$, which defines the function $\widetilde{m}_{k,\overline{\mu}}(x)$. Assume that for every $t \in [T]$, $\left|\overline{m}_k^t(R^t) - \widetilde{m}_{k,\overline{\mu}}(R^t)\right| \geq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(R^t)}$. Then, for every $R \subseteq \mathcal{X}$, we have $|\overline{m}_k(R) - \widetilde{m}_{k,\overline{\mu}}(R)| \leq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(R)}$ — that is $\overline{m}_k$ is β-pseudo-moment-consistent on every set R.*

Proof Set $R^{T+1} = R$. From Lemma 13, we get

$$\sum_{t=1}^{T+1}\psi^t\mathcal{P}_{\mathcal{X}}(R^t)\left(\overline{m}_k^t(R^t) - \widetilde{m}_{k,\overline{\mu}}(R^t)\right) \leq \frac{1}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T+1}\mathcal{P}_{\mathcal{X}}(R^t) \leq \sqrt{T+1}$$

$$\implies \mathcal{P}_{\mathcal{X}}(R^{T+1})\left|\overline{m}_k^{T+1}(R^{T+1}) - \widetilde{m}_{k,\overline{\mu}}(R^{T+1})\right| \leq \sqrt{T+1} - \beta T = \beta$$

∎

Lemma 14 *Assume $\overline{\mu}$ is such that for all $G \in \mathcal{G}$ and $i, j \in [m]$, $\overline{\mu}$ is α-mean consistent on every set $G(\overline{\mu}, \overline{m}_k, i, j))$:*

$$|\overline{\mu}(G(\overline{\mu}, \overline{m}_k, i, j)) - \mu(G(\overline{\mu}, \overline{m}_k, i, j))| \leq \frac{\alpha}{\mathcal{P}_\mathcal{X}(G(\overline{\mu}, \overline{m}_k, i, j))}.$$

Assume also that $\overline{m}_k$ is β-pseudo-moment-consistent with respect to $\overline{\mu}$ on every set $G(\overline{\mu}, \overline{m}_k, i, j))$ for $G \in \mathcal{G}$ and $i, j \in [m]$:

$$|\overline{m}_k(G(\overline{\mu}, \overline{m}_k, i, j)) - \widetilde{m}_{k,\overline{\mu}}(G(\overline{\mu}, \overline{m}_k, i, j))| \leq \frac{\beta}{\mathcal{P}_\mathcal{X}(G(\overline{\mu}, \overline{m}_k, i, j))}.$$

Then, for every $G \in \mathcal{G}$, $i, j \in [m]$, we have

$$|\overline{m}_k(G(\overline{\mu}, \overline{m}_k, i, j)) - m_k(G(\overline{\mu}, \overline{m}_k, i, j))| \leq \frac{\beta + k\alpha}{\mathcal{P}_\mathcal{X}(G(\overline{\mu}, \overline{m}_k, i, j)} + \frac{k}{m}.$$

This implies in particular that $(\overline{\mu}, \overline{m}_k)$ are $(\alpha, \beta', \epsilon)$-mean-conditioned moment multicalibrated with respect to $\mathcal{G}$, for $\beta' = \beta + k\alpha$ and $\epsilon = \frac{k}{m}$.

Proof Fix $G \in \mathcal{G}$ and $i, j \in [m]$ and let $S \equiv G(\overline{\mu}, \overline{m}_k, i, j)$. Because $\overline{\mu}$ is α-mean consistent on S, we have that:

$$|\mu(S) - \overline{\mu}(S)| \leq \frac{\alpha}{\mathcal{P}_\mathcal{X}(S)}. \tag{1}$$

We can use this to bound the difference between the true moment $m_k(S)$ and the pseudo-moment $\widetilde{m}_{k,\overline{\mu}}(S)$ on S. First, note that:

$$m_k(S) = \mathbb{E}_\mathcal{P}\left[(y - \mu(S))^k \Big| x \in S\right],$$
$$= \mathbb{E}_\mathcal{P}\left[[(y - \overline{\mu}(x)) + (\overline{\mu}(x) - \mu(S))]^k \Big| x \in S\right].$$

We will make use of the following fact:

Lemma 15 *For any $a, b \in [0, 1]$, $|a^k - b^k| \leq k|a - b|$.*

Proof Observe that:

$$|a^k - b^k| = \left|(a - b)\left(\sum_{\ell=0}^{k-1} a^\ell b^{k-1-\ell}\right)\right| \leq |a - b||k(\max(a, b))^{k-1}| \leq k|a - b|.$$

∎

Finally, we conclude that:

$$|m_k(S) - \widetilde{m}_{k,\overline{\mu}}(S)| = \left|\mathbb{E}_\mathcal{P}\left[((y - \overline{\mu}(x)) + (\overline{\mu}(x) - \mu(S)))^k - (y - \overline{\mu}(x))^k \Big| x \in S\right]\right|$$
$$\leq k\mathbb{E}_\mathcal{P}\left[|\overline{\mu}(x) - \mu(S)||x \in S\right]$$
$$\leq k\left(\mathbb{E}_\mathcal{P}\left[|\overline{\mu}(x) - \overline{\mu}(S)||x \in S\right] + |\overline{\mu}(S) - \mu(S)|\right)$$
$$\leq k\left(\frac{1}{m} + \frac{\alpha}{\mathcal{P}_\mathcal{X}(S)}\right).$$

The first inequality follows from Lemma 15 with $a = (y - \overline{\mu}(x)) + (\overline{\mu}(x) - \mu(S))$ and $b = y - \overline{\mu}(x)$. The final inequality follows from (1) (mean consistency) together with the fact that $\overline{\mu}(x)$ falls within a bucket of width $\frac{1}{m}$ for any $x \in S$ (recall that by definition, $S = G(\overline{\mu}, \overline{m}_k, i, j)$), and so does $\overline{\mu}(S)$

Finally, because $\overline{m}_k$ is β-pseudo-moment consistent on S with respect to $\overline{\mu}$ we can invoke the triangle inequality to conclude:

$$|\overline{m}_k(S) - m_k(S)| \leq |\overline{m}_k(S) - \widetilde{m}_{k,\overline{\mu}}(S)| + |\widetilde{m}_{k,\overline{\mu}}(S) - m_k(S)|,$$
$$\leq \frac{\beta}{\mathcal{P}_\mathcal{X}(S)} + k\left(\frac{1}{m} + \frac{\alpha}{\mathcal{P}_\mathcal{X}(S)}\right).$$

∎

Appendix C. Details and Proofs from Section 3.3

$$\overline{\mu}(x) = \begin{cases} \text{project}_{[0,1]}(\overline{\mu}(x) - \alpha\lambda) & \text{if } x \in S, \\ \overline{\mu}(x) & \text{otherwise.} \end{cases}$$

return $\overline{\mu}$

Algorithm 3: MeanConsistencyUpdate$(\overline{\mu}, S, \lambda)$

1: define pseudo-moment labels $\widetilde{m}_{a,\overline{\mu}}(x) = \mathbb{E}\left[(y - \overline{\mu}(x))^a | x\right]$ for all x
2: **while** $\exists R = G(\overline{\mu}, \overline{m}_a, i, j)$ for some $G \in \mathcal{G}$, $i, j \in [m]$ s.t. $|\overline{m}_a(R) - \widetilde{m}_{a,\overline{\mu}}(R)| \geq \frac{\beta}{\mathcal{P}_\mathcal{X}(R)}$ **do**
3: $\quad \psi = \text{sign}(\overline{m}_k(R) - \widetilde{m}_{k,\overline{\mu}}(R))$
4: $\quad \overline{m}_a(x) = \begin{cases} \text{project}_{[0,1]}(\overline{m}_a(x) - \beta\psi) & \text{if } x \in R \\ \overline{m}_a(x) & \text{otherwise.} \end{cases}$
5: return $\overline{m}_k$

Algorithm 4: PseudoMomentConsistency$(a, \beta, \overline{\mu}, \overline{m}_a, \mathcal{G})$

1: initialize $\overline{\mu}^1(x) = 0$ for all x
2: for all $1 < a \leq k$, initialize $\overline{m}_a^1(x) = 0$ for all x
3: $t = 1$
4: **while** $\exists S^t = G(\overline{\mu}^t, i)$ or $S^t = G(\overline{\mu}^t, \overline{m}_a^t, i, j)$ for some $G \in \mathcal{G}$, $i, j \in [m]$, $1 < a \leq k$ s.t. $\left|\overline{\mu}^t(S^t) - \mu(S^t)\right| \geq \frac{\alpha}{\mathcal{P}_\mathcal{X}(S^t)}$ **do**
5: $\quad \lambda^t = \text{sign}(\overline{\mu}(S^t) - \mu(S^t))$
6: $\quad \overline{\mu}^{t+1} = \text{MeanConsistencyUpdate}(\overline{\mu}^t, S^t, \lambda^t)$
7: $\quad$ **for** $a = 2, \ldots, k$ **do**
8: $\quad\quad \overline{m}_a^{t+1} = \text{PseudoMomentConsistency}(a, \beta, \overline{\mu}^{t+1}, \overline{m}_a^t, \mathcal{G})$.
9: $\quad t = t + 1$
10: return $(\overline{\mu}^t, \{\overline{m}_a^t\}_{a=2}^k)$

Algorithm 5: AlternatingGradientDescent$(\alpha, \beta, \mathcal{G})$

Theorem 11 *Let T be the final iterate t of Algorithm 5 (i.e. its output is $(\overline{\mu}^T, \{\overline{m}_a^T\}_{a=2}^k)$. Algorithm 5 has the following guarantees:*

1. **Total Iterations**: *The algorithm halts. The final iterate T is s.t. $T \leq \frac{1}{\alpha^2} - 1$. The total number of gradient descent update operations is at most $\left(\frac{1}{\alpha^2} - 1\right)\left(1 + (k-1)\left(\frac{1}{\beta^2} - 1\right)\right)$.*

2. **Mean multicalibration**: *Output $\overline{\mu}^T$ is α-mean multicalibrated with respect to $\mathcal{G}$.*

3. **Mean Conditioned Moment multicalibration**: *For every $a \in \{2, \ldots, k\}$, the pair $(\overline{\mu}^T, \overline{m}_a^T)$ is $(\alpha, \beta + a\alpha, \frac{a}{m})$-mean-conditioned moment-multicalibrated with respect to $\mathcal{G}$.*

Proof We prove each guarantee in turn.

Total Iterations: Every step t of the while loop in Algorithm 5 performs a gradient descent update using MeanConsistencyUpdate on a pair (λ^t, S^t) such that:

$$\lambda^t\left(\overline{\mu}^t(S^t) - \mu(S^t)\right) \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S^t)}.$$

By Theorem 7, this process can continue for at most $T \leq \frac{1}{\alpha^2} - 1$ many iterations. Within each iteration t of the loop, the algorithm makes one call to PseudoMomentConsistency for each $1 < a \leq k$ for a total of $(k-1)$ calls per iteration. Each of these calls performs at most $\frac{1}{\beta^2} - 1$ iterations of gradient descent, by Theorem 9.

Mean multicalibration: Suppose for a contradiction that Algorithm 5 terminates at $t = T$ with output $\overline{\mu}^T$ which is not mean multicalibrated, i.e. there exists a set $S \equiv G(\overline{\mu}^T, i)$ for some $G \in \mathcal{G}$, $i \in [m]$ such that $|\overline{\mu}^t(S) - \mu(S)| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)}$. Then, by construction of the while loop in Algorithm 5, T cannot be the final iterate of t.

Mean Conditioned Moment multicalibration: The While loop in Algorithm 5 will continue as long as there exists a set $S^t \equiv G(\overline{\mu}^t, \overline{m}_a^t, i, j)$ for some $G \in \mathcal{G}, i, j \in [m]$ such that $|\overline{\mu}^t(S^t) - \mu(S^t)| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S^t)}$. Hence we can conclude that at termination, $\overline{\mu}^T$ is α-mean consistent on every set $G(\overline{\mu}^T, \overline{m}_a^T, i, j)$ for some $G \in \mathcal{G}, i, j \in [m]$. Moreover, during the final iteration, for each $1 < a \leq k$, $\overline{m}_a^T$ was constructed by running PseudoMomentConsistency$(a, \beta, \overline{\mu}^T, \overline{m}_a^{T-1}, \mathcal{G})$. Therefore, by Theorem 9 we know that $\overline{m}_a^T$ is β-pseudo-moment consistent on every set $G(\overline{\mu}^T, \overline{m}_a^T, i, j)$. To see this, note that if PseudoMomentConsistency runs for $\frac{1}{\beta^2} - 1$ many rounds, then it is β-pseudo-moment consistent on *every* set. On the other hand, the only way it can halt before that many rounds (by construction of the halting condition in its While loop) is if $\overline{m}_a^T$ is β-pseudo-moment consistent on every set $G(\overline{\mu}^T, \overline{m}_a^T, i, j)$.

Therefore, $\overline{\mu}^T$ and $\{\overline{m}_a^T\}_{a=2}^k$ jointly satisfy the conditions of Lemma 10. It follows from the Lemma that they are mean-conditioned moment-multicalibrated at the desired parameters. ∎

Appendix D. Implementation with Finite Sample and Runtime Guarantees

In Section 3 we analyzed a version of our algorithm as if we had direct access to the true distribution, $\mathcal{P}$. In particular, in both Algorithm 4 and Algorithm 5, access to $\mathcal{P}$ was needed in (only) two places. First, to identify a set S^t such that $\left|\overline{\mu}^t(S^t) - \mu(S^t)\right| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S^t)}$. Second, to identify a set R such that $|\overline{m}_a(R) - \widetilde{m}_{a,\overline{\mu}}(R)| \geq \frac{\beta}{\mathcal{P}_{\mathcal{X}}(R)}$. In this section, we show how to perform these operations approximately by using a small finite sample of points drawn from $\mathcal{P}$, and hence to obtain a finite sample version of our main result together with sample complexity and running time bounds.

There are two issues at play here: the first issue is purely statistical: how many *samples* are needed to execute the two checks needed to implement our algorithm? Our finite sample algorithm will essentially use a sufficiently large fresh sample of data at every iteration to guarantee uniform convergence of the quantities to be estimated over all of the sets that must be checked at that iteration. The second issue is computational: even if we have enough samples so that we can check in-sample quantities as proxies for the distributional quantities we care about, what is the running time of our algorithm? We are performing gradient descent in a potentially infinite dimensional space, and so we cannot explicitly maintain the weights $\overline{\mu}^t(x), \overline{m}_a^t(x)$ for all x. Instead, we maintain these weights *implicitly* as a weighted linear combination of the indicator functions for each of the sets S^t, R, used to perform updates (recall that we have assumed that each set $G \in \mathcal{G}$ can be represented by an indicator function computed by a polynomially sized circuit, so we have concise implicit representations of every set that our algorithm must manipulate). Ostensibly one must exhaustively enumerate the collection of sets S, R, for which our algorithm must perform some check (in fact their indicator functions), which takes time that scales with $m^2 \cdot |\mathcal{G}|$. We first focus on the statistical problem, showing that the number of *samples* needed to implement our algorithm is small, and then we observe that if we have an *agnostic learning algorithm* for $\mathcal{G}$, we can use it to replace exhaustive enumeration. In both cases—although the details differ—we handle these issues in largely the same way they were handled by Hébert-Johnson et al. (2018), so many of the proofs and calculations will be deferred to the Appendix.

Finally, we remark that it is essential that we draw a fresh sample of n data points each time we try to find a set for consistency violation because $\ell, \overline{\ell}$, as well as the collection of sets $\mathcal{S}$ that we are auditing, are not fixed *a priori* but change adaptively (i.e. as a function of the data) between rounds. Due to the adaptive nature of the statistical tests that need to be performed, we cannot simply union bound over these queries. We remark that we could have applied adaptive data analysis techniques (see e.g. (Dwork et al., 2015; Bassily et al., 2016; Jung et al., 2020)) to partially re-use the data, which would save a quadratic factor in the sample complexity (or for finite data domains, an exponential improvement in some of the existing parameters, at the cost of an additional dependence on $\log |\mathcal{X}|$ by using the private multiplicative weights algorithm of Hardt and Rothblum (2010)). This idea is applied in Hébert-Johnson et al. (2018); it applies here in the same manner; interested readers can refer to Hébert-Johnson et al. (2018).

D.1. Sample Complexity Bounds and a Finite Sample Algorithm via Exhaustive Group Enumeration

First, recall that pseudo-moment consistency is mean consistency with respect to the artificially created label $\widetilde{m}_{k,\overline{\mu}}(x) = (y - \overline{\mu}(x))^k$. To avoid needless repetition, we focus on achieving mean consistency for an arbitrary label defined by a label function $\ell(x,y)$ with a predictor $\overline{\ell}(x)$. Then, auditing for mean consistency for $\overline{\mu}$ and pseudo-moment consistency for $\overline{m}_k$ follows by setting

$$\overline{\ell}(x) = \overline{\mu}(x) \quad \text{and} \quad \ell(x,y) = y$$
$$\overline{\ell}(x) = \overline{m}_k(x) \quad \text{and} \quad \ell(x,y) = (y - \overline{\mu}(x))^k$$

for mean consistency and pseudo-moment consistency respectively. For economy of notation set,

$$\overline{\ell}(S) = \mathbb{E}_{\mathcal{P}}[\overline{\ell}(x)|x \in S] \quad \text{and} \quad \ell(S) = \mathbb{E}_{\mathcal{P}}[\ell(x,y)|x \in S],$$

for all $S \subseteq \mathcal{X}$. For any set $S \subseteq \mathcal{X}$, given a dataset D, we refer to D_S as the subset of the data where the corresponding points lie in S (i.e. $D_S = \{(x, y) \in D : x \in S\}$). If dataset D_S has n' points each drawn independently from $\mathcal{P}$ conditional on $x \in S$, we can appeal to the Chernoff bound (Theorem 29) to argue that empirical averages must be close to their expectations.

We can also appeal to the Chernoff bound to argue that when we sample data points from $\mathcal{P}$, the number of points that fall into some set S (e.g. $G(\overline{\mu}, \overline{m}_k, i, j)$) scales roughly with $n\mathcal{P}_{\mathcal{X}}(S)$ (Lemma 30).

Throughout the execution of our algorithm, we need to audit the current mean and moment estimators for α-mean consistency violations. This is important, because the analysis of the running time of the algorithm (e.g. the fact that it converges after at most $T = \frac{1}{\alpha^2} - 1$ iterations) relies on making a minimum amount of progress guaranteed by α-mean *inconsistency*. In the next lemma, we provide a condition that can be checked using empirical estimates which guarantees α-mean inconsistency (on the true, unknown distribution) whenever the sample is appropriately close to the distribution; it follows from two applications of a Chernoff bound that this approximate closeness condition will occur with high probability. We encapsulate this empirical check in Algorithm 6.

> **if** $n' > 0$ and $\left| \frac{1}{n'} \sum_{b=1}^{n'} \overline{\ell}(x_b) - \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b) \right| - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \geq \frac{\alpha}{\frac{n'}{n} - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}}$ **then**
>
> $\quad \lambda = \text{sign}(\frac{1}{n'} \sum_{b=1}^{n'} \overline{\ell}(x_b) - \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b))$
> $\quad$ Output YES, λ (A Consistency Violation has been found)
> **else**
> $\quad$ Output No

Algorithm 6: Auditor$(\ell, \overline{\ell}, \alpha, \delta, \{(x_b, y_b)\}_{b=1}^{n'})$

Definition 16 *Fix any set $S \subseteq \mathcal{X}$. Given a set of n data points D and its associated $D_S = \{(x_b, y_b)\}_{b=1}^{n'}$, we say that D is approximately close to $\mathcal{P}$ with respect to $(S, \ell, \overline{\ell})$, if the following inequalities hold true:*

$$n' > 0$$

$$\left| \frac{n'}{n} - \mathcal{P}_{\mathcal{X}}(S) \right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \tag{2a}$$

$$\left| \frac{1}{n'} \sum_{b=1}^{n'} \overline{\ell}(x_b) - \overline{\ell}(S) \right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \tag{2b}$$

$$\left| \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b) - \ell(S) \right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \tag{2c}$$

Lemma 17 *Fix any set $S \subseteq \mathcal{X}$. If dataset D is approximately close to $\mathcal{P}$ with respect to $(S, \overline{\ell}, \ell)$, then we have*

$$\text{Auditor}(\ell, \overline{\ell}, \alpha, D_S) = (YES, \lambda) \implies |\overline{\ell}(S) - \ell(S)| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} \quad \text{and} \quad \lambda = \text{sign}(\overline{\ell}(S) - \ell(S))$$

24

Lemma 17 implies that when we find an empirical consistency violation using Algorithm 6, it is indeed a real α-consistency violation with respect to the true distribution, allowing us to make progress — this guarantees that our algorithm will not run for too many iterations. But we need a converse condition, in order to make sure that we don't halt too early: we must show that if there are no empirical α'-consistency violations for some $\alpha' > \alpha$, then there are also no α-consistency violations with respect to the true distribution. This is what we do in Lemma 18. Observe, that without loss of generality, we can restrict attention to sets such that $\mathcal{P}_{\mathcal{X}}(S) \geq \alpha$ because any estimator in the range $[0, 1]$ is trivially α-mean consistent on every set with measure $< \alpha$.

Lemma 18 *Fix any set $S \subseteq \mathcal{X}$ such that $\mathcal{P}_{\mathcal{X}}(S) \geq \alpha$. Assume n is sufficiently large such that $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} < \alpha$ If D is approximately close to $\mathcal{P}$ with respect to $(S, \overline{\ell}, \ell)$, we have*

$$\left|\overline{\ell}(S) - \ell(S)\right| \geq \frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)} \implies \textit{Auditor}(\ell, \overline{\ell}, \alpha, D_S) = \textit{YES},$$

where $\alpha' = \alpha + 4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})} + \left(\alpha - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)^{-2} \left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right).$

The Auditor subroutine above performs a consistency check on a single set. We now use it to audit for mean consistency and pseudo-moment consistency across a collection of sets.

> **for** $S \in \mathcal{S}$ **do**
>   **if** $|D_S| > 0$ and Auditor$(\ell, \overline{\ell}, \alpha, \delta, D_S) = YES, \lambda$ **then**
>   return S, λ
> return $NULL$

Algorithm 7: ConsistencyAuditor$(\overline{\ell}, \ell, \alpha, \delta, D, \mathcal{S})$

Corollary 19 *Fix $\overline{\ell}, \ell, \alpha, \delta$, and a collection of sets $\mathcal{S}$. Given a set of n points D drawn i.i.d. from $\mathcal{P}$ where $\alpha > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}$, ConsistencyAuditor$(\overline{\mu}, \alpha, D, \mathcal{S})$ has the following guarantee with probability $1 - 3\delta|\mathcal{S}|$ over the randomness of D:*

1. *If ConsistencyAuditor does output some set S and λ, then*

$$\left|\overline{\ell}(S) - \ell(S)\right| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} \quad \textit{and} \quad \lambda = sign(\overline{\ell}(S) - \ell(S)).$$

2. *If ConsistencyAuditor outputs $NULL$, then for all $S \in \mathcal{S}$,*

$$\left|\overline{\ell}(S) - \ell(S)\right| \leq \frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)},$$

where $\alpha' = \alpha + 4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})} + \left(\alpha - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)^{-2} \left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right).$

Thus, to detect a set S with α-mean consistency violation in line 4 of Algorithm 5, we can leverage Algorithm 7 by drawing a fresh sample of size n and setting $\bar{\ell}(x) = \bar{\mu}(x), \ell(x, y) = y$, and $\mathcal{S} = \{G(\bar{\mu}, i) : G \in \mathcal{G}, i \in [m]\} \cup \{G(\bar{\mu}, \overline{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m], a \in \{2, \ldots, k\}\}$. Likewise, for every $a \in \{2, \ldots, k\}$, to detect a set S with β-pseudo-moment consistency violation in line 2 of Algorithm 4, we can leverage Algorithm 7 by drawing a fresh sample of size n and setting $\bar{\ell}(x) = \overline{m}_k(x), \ell(x, y) = (y - \bar{\mu}(x))^k$, and $\mathcal{S} = \{G(\bar{\mu}, \overline{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m]\}$. We write out the pseudocode of this process below.

$D = \{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n$
$\mathcal{S} = \{G(\bar{\mu}, \overline{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m]\}$
$\bar{\ell}(x) = \overline{m}_a(x)$
$\ell(x, y) = (y - \bar{\mu}(x))^a$
$R, \psi = \text{ConsistencyAuditor}(\bar{\ell}, \ell, \beta, \delta, D, \mathcal{S})$
while $R, \psi \neq NULL$ **do**

$$\overline{m}_a(x) = \begin{cases} \text{project}_{[0,1]}(\overline{m}_a(x) - \beta\psi) & \text{if } x \in R \\ \overline{m}_a(x) & \text{otherwise.} \end{cases}$$

    $D = \{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n$
    $\mathcal{S} = \{G(\bar{\mu}, \overline{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m]\}$
    $\bar{\ell}(x) = \overline{m}_a(x)$
    $R, \psi = \text{ConsistencyAuditor}(\bar{\ell}, \ell, \beta, \delta, D, \mathcal{S})$
return $\overline{m}_k$

Algorithm 8: PseudoMomentConsistencyFinite$(a, \beta, \delta, \bar{\mu}, \overline{m}_a, n, \mathcal{G})$

initialize $\bar{\mu}^1(x) = 0$ for all x
for all $1 < a \leq k$, initialize $\overline{m}_a^1(x) = 0$ for all x
$t = 1$
$\bar{\ell}^t(x) = \bar{\mu}^1(x)$
$\ell(x, y) = y$
$D^t = \{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n$
$\mathcal{S}^t = \{G(\bar{\mu}^t, i) : G \in \mathcal{G}, i \in [m]\} \cup \{G(\bar{\mu}^t, \overline{m}_a^t, i, j) : G \in \mathcal{G}, i, j \in [m], a \in \{2, \ldots, k\}\}$
$S^t, \lambda^t = \text{ConsistencyAuditor}(\bar{\ell}^t, \ell, \alpha, \delta, D, \mathcal{S})$
while $S^t, \lambda^t \neq NULL$ **do**
    $\bar{\mu}^{t+1} = \text{MeanConsistencyUpdate}(\bar{\mu}^t, S^t, \lambda^t)$
    for $a = 2, \ldots, k$ **do**
        $\overline{m}_a^{t+1} = \text{PseudoMomentConsistencyFinite}(a, \beta, \delta, \bar{\mu}^{t+1}, \overline{m}_a^t, n, \mathcal{G})$.
    $t = t + 1$
    $\bar{\ell}^t(x) = \bar{\mu}^t(x)$
    $D^t = \{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n$
    $\mathcal{S}^t = \{G(\bar{\mu}^t, i) : G \in \mathcal{G}, i \in [m]\} \cup \{G(\bar{\mu}^t, \overline{m}_a^t, i, j) : G \in \mathcal{G}, i, j \in [m], a \in \{2, \ldots, k\}\}$
    $S^t, \lambda^t = \text{ConsistencyAuditor}(\bar{\ell}^t, \ell, \alpha, \delta, D^t, \mathcal{S}^t)$
return $(\bar{\mu}^t, \{\overline{m}_a^t\}_{a=2}^k)$

Algorithm 9: AlternatingGradientDescentFinite$(\alpha, \beta, \delta, n, \mathcal{G})$

Theorem 20 *Let T be the final iterate of Algorithm 9. If $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \leq \alpha$ and $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \leq \beta$, we have the following guarantees:*

1. **Total Iterations**: *With probability $1 - 3\delta|\mathcal{G}|Q_\alpha\left((m^2 + m) + m^2 Q_\beta\right)$ over the randomness of our samples, the final iterate T is s.t. $T \leq \frac{1}{\alpha^2} - 1$ and the total number of gradient descent update operations will be at most Q, where*

$$Q_\alpha = \frac{1}{\alpha^2} - 1, Q_\beta = (k-1)\left(\frac{1}{\beta^2} - 1\right), Q = Q_\alpha(1 + Q_\beta).$$

   In particular, the algorithm uses at most nQ samples from $\mathcal{P}$.

2. **Mean multicalibration**: *With probability $1 - 3\delta(m^2 + m)|\mathcal{G}|$, output $\overline{\mu}^T$ is α'-mean multi-calibrated with respect to $\mathcal{G}$ where*

$$\alpha' = \alpha + 4\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} + \left(\alpha - 2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right).$$

3. **Mean Conditioned Moment multicalibration**: *With probability $1 - 3\delta|\mathcal{G}|(km^2 + m)$, for any $a \in \{2, \ldots, k\}$, pair $(\overline{\mu}^T, \overline{m}_a^T)$ is $(\alpha', a\alpha' + \beta', \frac{a}{m})$-mean-conditioned-moment calibrated where*

$$\alpha' = \alpha + 4\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} + \left(\alpha - 2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)$$

$$\beta' = \beta + 4\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} + \left(\beta - 2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)$$

The following corollary derives the sample complexity of our algorithm, implied by Theorem 20, for a target set of parameters. Observe that the sample complexity is polynomial in $k, m, 1/\alpha, 1/\beta, 1/\epsilon, \log(1/\delta)$, and $\log|\mathcal{G}|$.

Corollary 21 *Fix target parameters α', β', δ' and $\epsilon > 0$ such that $\epsilon < \alpha'$ and $\epsilon < \beta'$. Define*

$$\overline{Q} = \frac{6|\mathcal{G}|km^2}{\left(\frac{\alpha'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2\left(\frac{\beta'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}, \quad \delta = \frac{\delta'}{\max(3|\mathcal{G}|(km^2+m), \overline{Q})}, \quad n_\alpha = \frac{\ln(\frac{2\overline{Q}}{\delta})}{2\left(\frac{\alpha'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}, \quad n_\beta = \frac{\ln(\frac{2\overline{Q}}{\delta})}{2\left(\frac{\beta'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}.$$

Then, AlternatingGradientDescentFinite$(\alpha, \beta, \delta, n, \mathcal{G})$ where

$$\alpha = 2\sqrt{\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\alpha}} + \epsilon, \beta = 2\sqrt{\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\beta}} + \epsilon,$$

$$n = \max\left(\frac{\ln(\frac{2\overline{Q}}{\delta})}{\ln(\frac{2}{\delta})}n_\alpha, \frac{\ln(\frac{2\overline{Q}}{\delta})}{\ln(\frac{2}{\delta})}n_\beta, \frac{2\ln(\frac{2}{\delta})}{\alpha^2}, \frac{2\ln(\frac{2}{\delta})}{\beta^2}\right)$$

has the following guarantees with probability $1 - \delta'$:

1. *The total number of gradient descent updates will be at most Q, where Q is as defined in Theorem 20.*

2. *$\overline{\mu}^T$ is α'-mean-calibrated.*

3. *For every $a \in \{2, \ldots, k\}$, $(\overline{\mu}^T, \overline{m}_a^T)$ is $(\alpha', a\alpha' + \beta', \frac{a}{m})$-mean-conditioned-moment calibrated.*

Finally, we state the running time of the algorithm in the following Theorem.

Theorem 22 *With probability $1 - 3\delta|\mathcal{G}|Q_\alpha\left((m^2 + m) + m^2 Q_\beta\right)$, the running time of Algorithm 9 is $O\left(Q|\mathcal{G}|m^2 n\right) = O\left(\frac{k|\mathcal{G}|m^2 n}{\alpha^2 \beta^2}\right)$ where Q_α, Q_β, and Q are as defined in Theorem 20.*

D.2. Oracle Efficient Implementation

In Section D.1 we analyzed an algorithm that had favorable sample-complexity bounds, but was computationally expensive when $\mathcal{G}$ was large: although it ran for only a small number of iterations, each iteration required a complete enumeration of every set in $\mathcal{G}$. In this section, we show how to replace this expensive step with a call to an algorithm which can solve learning problems over $\mathcal{G}$, if one is available. Because the remaining portion of the algorithm is computationally efficient — even if $\mathcal{G}$ is very large — this yields what is sometimes known as an "oracle efficient algorithm". Similar reductions have been given in Hébert-Johnson et al. (2018); Kearns et al. (2018); Kim et al. (2019).

Definition 23 *For some $\rho \in [0, 1]$ and non-increasing function $p : \mathbb{N} \to [0, 1]$, A is a (ρ, p)-agnostic learning oracle for hypothesis class $\mathcal{H} \subseteq 2^{\mathcal{X}}$ with respect to a label function $r(x, y) \in [-1, 1]$, if for any distribution $\mathcal{P}$, given n random samples from $\mathcal{P}$, it outputs $f : \mathcal{X} \to \{0, 1\}$ such that with probability $1 - p(n)$,*

$$\mathbb{E}_{(x,y)\sim\mathcal{P}}[f(x) \cdot r(x, y)] + \rho \geq \sup_{h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{P}}[h(x) \cdot r(x, y)].$$

We write $\tau(n)$ to denote the running time of the oracle A when n data points are used, which we assume is at least $\Omega(n)$.

Remark 24 *A more common definition of an agnostic learning oracle would use hypotheses with range $\{-1, 1\}$ rather than $\{0, 1\}$. But this definition will be more convenient for us, and is equivalent (up to a constant factor in the parameters) via a linear transformation.*

We will use a learning algorithm for any class $\mathcal{H}$ such that $\mathcal{G} \subseteq \mathcal{H}$ to replace the set enumeration steps of our algorithm. In particular, to find a set of the form $G(\overline{\mu}, i)$ on which our existing predictor $\overline{\mu}$ fails to be mean consistent, we run our learning algorithm on the subset of our sample that intersects with $\mathcal{X}(\overline{\mu}, i)$, labelled with the positive and negative *residuals* of our predictor — i.e. on the labels $r_R^+(x, y) = \overline{\mu}(x) - y$ and $r_R^- = y - \overline{\mu}(x)$. Similarly, to find a set of the form $G(\overline{\mu}, \overline{m}_a, i, j)$, we run our learning algorithm on the sets $\mathcal{X}(\overline{\mu}, \overline{m}_a, i, j)$ labeled with both the positive and negative residuals. Finding sets on which we fail to be moment pseudo-consistent with respect to $\overline{\mu}$ is similar. All in all, this requires $O(k \cdot m^2)$ runs of our learning algorithm per gradient descent step,

replacing the complete enumeration of the collection of sets $\mathcal{G}$. We make this process more precise in Algorithm 11 and state the guarantees in Theorem 26. We also include the pseudocode for the correspondingly updated AlternatingGradientDescentFinite using Algorithm 11 as the auditing subroutine in the appendix – see Algorithm 13.

$$r_R^+(x,y) = \begin{cases} \overline{\ell}(x) - \ell(x,y) & \text{if } x \in R \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad D_R^+ = \{(x_b, r_R^+(x,y))\}_{b=1}^n$$

$$r_R^-(x,y) = \begin{cases} \ell(x,y) - \overline{\ell}(x) & \text{if } x \in R \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad D_R^- = \{(x_b, r_R^-(x,y))\}_{b=1}^n$$

$\chi_{S^+} = A(D_R^+, \mathcal{H})$
$\chi_{S^-} = A(D_R^-, \mathcal{H})$
return (S^+, S^-)

Algorithm 10: LearningOracleConsistencyAuditor($\overline{\ell}, \ell, \alpha, \delta, D, R, A$)

$\mathcal{V} = \{\}$
for $R \in \mathcal{R}$ **do**
  if $|D_R| > 0$ **then**
    $S^+, S^- = $ LearningOracleConsistencyAuditor($\overline{\ell}, \ell, \alpha, \delta, D, D^{check}, R, A$)
    $\mathcal{V} = \mathcal{V} \cup \{(S^+ \cap R), (S^- \cap R)\}$
return ConsistencyAuditor($\overline{\ell}, \ell, \alpha, D^{\text{check}}, \mathcal{V}$)

Algorithm 11: LearningOracleConsistencyAuditorWrapper($\overline{\ell}, \ell, \alpha, \delta, D, D^{\text{check}}, \mathcal{R}, A$)

First we observe that the objective of the agnostic learning oracle on the sets we run it on corresponds directly to the (positive and negative) violation of mean consistency on these sets, weighted by the measure of the sets.

Lemma 25 *For each $R \in \mathcal{R}$ and any χ_S:*

$$\mathbb{E}_{(x,y)}[\chi_S(x) \cdot r_R^+(x,y)] = \mathcal{P}_{\mathcal{X}}(R \cap S)\left(\overline{\ell}(R \cap S) - \ell(R \cap S)\right)$$

$$\mathbb{E}_{(x,y)}[\chi_S(x) \cdot r_R^-(x,y)] = \mathcal{P}_{\mathcal{X}}(R \cap S)\left(\ell(R \cap S) - \overline{\ell}(R \cap S)\right)$$

Using this, we can show that our learning oracle based consistency auditor has comparable guarantees to the consistency auditor that operated via set enumeration:

Theorem 26 *Assume n is sufficiently large such that $\alpha > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}$. Algorithm 11 has the following guarantees:*

1. *If it returns some S and λ, then with probability $1 - 3\delta|\mathcal{R}|$ over the randomness of D^{check},*

$$|\overline{\ell}(S) - \ell(S)| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)}.$$

2. *If it returns $NULL$, then with probability $1 - |\mathcal{R}|(3\delta + 2p(n))$ over the randomness of D and D^{check}, for all $\chi_S \in \mathcal{H}$ and $R \in \mathcal{R}$,*

$$|\bar{\ell}(R \cap S) - \ell(R \cap S)| \leq \frac{\alpha' + \rho}{\mathcal{P}_{\mathcal{X}}(R \cap S)},$$

where α' is as defined in Corollary 19.

Observe that when $\mathcal{G} \subseteq \mathcal{H}$ and $\mathcal{R} = \{\mathcal{X}(\bar{\mu}, \bar{m}_a, i, j) : i, j \in [m]\}$, then, the collection of intersections $R \cap S$ over all $\chi_S \in \mathcal{H}$ and $R \in \mathcal{R}$ contains $\{G(\bar{\mu}, \bar{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m]\}$. The same observation applies when $\mathcal{R} = \{\mathcal{X}(\bar{\mu}, i) : i \in [m]\} \cup \{\mathcal{X}(\bar{\mu}, \bar{m}_a, i, j) : i, j \in [m]\}$ – the collection of intersections includes $\{G(\bar{\mu}, i) : G \in \mathcal{G}, i \in [m]\} \cup \{\mathcal{X}(\bar{\mu}, \bar{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m]\}$.

We now present the guarantees of a version of AlternatingGradientDescent that uses Algorithm 11 as the auditor. Its pseudo-code can be found as Algorithm 13 in the appendix. We elide the proof as it is almost identical to that of Theorem 11 and Theorem 20.

Theorem 27 *Assume $\mathcal{G} \subseteq \mathcal{H}$. Let T be the final iterate of Algorithm 13. If $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \leq \alpha$, $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \leq \beta$, and $\mathcal{G} \subseteq \mathcal{H}$, we have the following guarantees:*

1. **Total Iterations**: *With probability $1 - 3\delta Q_\alpha \left((m^2 + m) + m^2 Q_\beta\right)$ over the randomness of our samples, the final iterate T is such that $T \leq \frac{1}{\alpha^2} - 1$ and the total number of gradient descent update operations will be at most Q, where Q_α, Q_β, and Q are all as defined in Theorem 20.*

   In particular, the algorithm uses at most $O(nQ)$ samples from $\mathcal{P}$.

2. **Mean multicalibration**: *With probability $1 - (m^2 + m)(3\delta + 2p(n))$, output $\bar{\mu}^T$ is α''-mean multicalibrated with respect to $\mathcal{G}$ where $\alpha'' = \alpha' + \rho$ and α' is as defined in Theorem 20.*

3. **Mean Conditioned Moment multicalibration**: *With probability $1 - (km^2 + m)(3\delta + 2p(n))$, for any $a \in \{2, \ldots, k\}$, pair $(\bar{\mu}^T, \bar{m}_a^T)$ is $(\alpha'', a\alpha'' + \beta'', \frac{a}{m})$-mean-conditioned-moment calibrated where $\beta'' = \beta' + \rho$ and β' is as defined in Theorem 20.*

Finally, we state the running time of Algorithm 13.

Theorem 28 *With probability at least $1 - 3\delta Q_\alpha \left((m^2 + m) + m^2 Q_\beta\right)$, the running time of Algorithm 13 is bounded by $O(Qm^2\tau(n))$, where Q is the total number of gradient descent operations as defined in Theorem 20.*

Appendix E. Details and Proofs from Section D

Theorem 29 (Chernoff Bound) *Fix distribution $\mathcal{P}$ and some function $f(x, y) \in [0, 1]$. Let $\{(x_b, y_b)\}_{b=1}^n$ be n points sampled i.i.d. from $\mathcal{P}$. Then, we have for any $\delta \in [0, 1]$,*

$$\Pr_{\{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n} \left[\left| \frac{1}{n} \sum_{b=1}^n f(x_b, y_b) - \mathbb{E}_{(x,y) \sim \mathcal{P}}[f(x, y)] \right| \geq \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \right] \leq \delta.$$

Lemma 30 *For any set $S \subseteq \mathcal{X}$,*

$$
\Pr_{D \sim \mathcal{P}^n} \left[\left| \frac{1}{n} \sum_{b=1}^{n} \mathbb{1}(x_b \in S) - \mathcal{P}_{\mathcal{X}}(S) \right| > \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \right] \leq \delta
$$

Proof We apply a Chernoff bound (Theorem 29) with $f(x,y) = \mathbb{1}(x \in S)$. Observe that $\mathbb{E}[f(x,y)] = \mathcal{P}_{\mathcal{X}}(S)$. ∎

Lemma 31 *Fix any set $S \subseteq \mathcal{X}$. If dataset D is approximately close to $\mathcal{P}$ with respect to $(S, \bar{\ell}, \ell)$, then we have*

$$
\text{Auditor}(\ell, \bar{\ell}, \alpha, D_S) = (\text{YES}, \lambda) \implies \left| \bar{\ell}(S) - \ell(S) \right| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} \quad \text{and} \quad \lambda = sign(\bar{\ell}(S) - \ell(S))
$$

Proof To see this, observe that

$$
\begin{aligned}
&\left| \bar{\ell}(S) - \ell(S) \right| \\
&\geq \left| \frac{1}{n'} \sum_{b=1}^{n'} \bar{\ell}(x_b) - \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b) \right| - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \\
&\geq \frac{\alpha}{\frac{n'}{n} - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}} \\
&\geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)}
\end{aligned}
$$

Here, the first inequality follows from the (2b) and (2c), the second from the condition of Algorithm 6, and the last inequality follows from (2a).

Finally, if $\frac{1}{n'} \sum_{b=1}^{n'} \bar{\ell}(x_b) \geq \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b)$, then

$$
\bar{\ell}(S) \geq \frac{1}{n'} \sum_{b=1}^{n'} \bar{\ell}(x_b) - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \geq \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b) + \sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \geq \ell(S).
$$

The same argument applies when $\frac{1}{n'} \sum_{b=1}^{n'} \bar{\ell}(x_b) < \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b)$. Therefore, $sign(\frac{1}{n'} \sum_{b=1}^{n'} \bar{\ell}(x_b) - \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b)) = sign(\bar{\ell}(S) - \ell(S))$. ∎

Lemma 32 *Fix any set $S \subseteq \mathcal{X}$ such that $\mathcal{P}_{\mathcal{X}}(S) \geq \alpha$. Assume n is sufficiently large such that $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} < \alpha$ If D is approximately close to $\mathcal{P}$ with respect to $(S, \bar{\ell}, \ell)$, we have*

$$
\left| \bar{\ell}(S) - \ell(S) \right| \geq \frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)} \implies \text{Auditor}(\ell, \bar{\ell}, \alpha, D_S) = \text{YES},
$$

where $\alpha' = \alpha + 4\sqrt{\frac{1}{2n} \ln(\frac{2}{\delta})} + \left(\alpha - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \right)^{-2} \left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \right).$

Proof The pre-condition implies that

$$\left| \frac{1}{n'} \sum_{b=1}^{n'} \overline{\ell}(x_b) - \frac{1}{n'} \sum_{b=1}^{n'} \ell(x_b, y_b) \right| - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}}$$

$$\geq \left| \overline{\ell}(S) - \ell(S) \right| - 4\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}}$$

$$\geq \frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)} - 4\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}}.$$

Therefore it is sufficient to show that $\frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)} - 4\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \geq \frac{\alpha}{\frac{n'}{n} - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}}.$

$$\frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)} = \frac{\alpha + 4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})} + \left(\alpha - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)}{\mathcal{P}_{\mathcal{X}}(S)}$$

$$\geq \frac{\alpha + 4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})}}{\mathcal{P}_{\mathcal{X}}(S) - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}} \tag{3}$$

$$\geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S) - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}} + \frac{4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})}}{\mathcal{P}_{\mathcal{X}}(S) - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}}$$

$$\geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S) - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}} + 4\sqrt{\frac{\ln(\frac{2}{\delta})}{2n(\mathcal{P}_{\mathcal{X}}(S) - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}})}} \qquad \forall x \in [0,1] : x \leq \sqrt{x}$$

$$\geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S) - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}} + 4\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \qquad\qquad \text{by (2a)}$$

$$\geq \frac{\alpha}{\frac{n'}{n} - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}} + 4\sqrt{\frac{\ln(\frac{2}{\delta})}{2n'}} \qquad\qquad \text{by (2a)}$$

Inequality (3) comes from Lemma 33, where we plug in $x = \mathcal{P}_{\mathcal{X}}(S)$, $c = \alpha + 4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})}$ and $\epsilon = 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}$. ∎

Lemma 33 *For any $0 < \epsilon \leq \alpha \leq x \leq 1$ and $0 < c \leq 1$,*

$$\frac{c + \frac{\epsilon}{(\alpha - \epsilon)^2}}{x} \geq \frac{c}{x - \epsilon}$$

Proof Because

$$\frac{c + \frac{\epsilon}{(\alpha-\epsilon)^2}}{x} \geq \frac{c}{x} + \frac{\epsilon}{(\alpha-\epsilon)^2},$$

it is sufficient to show that

$$\frac{c}{x} + \frac{\epsilon}{(\alpha-\epsilon)^2} \geq \frac{c}{x-\epsilon}.$$

Because $f(x) = \frac{c}{x}$ is convex, it's easy to see that:

$$f(x-\epsilon) + \epsilon f'(x-\epsilon) \leq f(x)$$

$$\frac{c}{x-\epsilon} - \frac{c\epsilon}{(x-\epsilon)^2} \leq \frac{c}{x}.$$

Now, because $\epsilon \leq \alpha \leq x$ and $0 < c \leq 1$, we have

$$\frac{c}{x-\epsilon} - \frac{\epsilon}{(\alpha-\epsilon)^2} \leq \frac{c}{x}.$$

∎

Corollary 34 *Fix $\bar{\ell}, \ell, \alpha, \delta$, and a collection of sets $\mathcal{S}$. Given a set of n points D drawn i.i.d. from $\mathcal{P}$ where $\alpha > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}$, ConsistencyAuditor($\bar{\mu}, \alpha, D, \mathcal{S}$) has the following guarantee with probability $1 - 3\delta|\mathcal{S}|$ over the randomness of D:*

1. If ConsistencyAuditor does output some set S and λ, then

$$\left|\bar{\ell}(S) - \ell(S)\right| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} \quad \text{and} \quad \lambda = sign(\bar{\ell}(S) - \ell(S)).$$

2. If ConsistencyAuditor outputs $NULL$, then for all $S \in \mathcal{S}$,

$$\left|\bar{\ell}(S) - \ell(S)\right| \leq \frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)},$$

where $\alpha' = \alpha + 4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})} + \left(\alpha - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right).$

Proof For each set $S \in \mathcal{S}$, we write $D_S = \{(x_b^S, y_b^S)\}_{b=1}^{n'_S}$ to denote the points from D that fall in S.

First, by union bounding the failure probabilities of Lemma 30 over every $S \in \mathcal{S}$, we have with probability $1 - \delta|\mathcal{S}|$,

$$\left|\frac{n'_S}{n} - \mathcal{P}_{\mathcal{X}}(S)\right| > \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}.$$

We apply the Chernoff bound again for every set S where $n'_S > 0$ and take the union bound to argue that with probability at least $1 - 2|\mathcal{S}|\delta$, for all such sets S where $n'_S > 0$,

$$\left|\frac{1}{n'_S}\sum_{b=1}^{n'_S}\bar{\ell}(x_b^S) - \bar{\ell}(S)\right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{2n'_S}}$$

$$\left|\frac{1}{n'_S}\sum_{b=1}^{n'_S}\ell(x_b^S, y_b^S) - \ell(S)\right| \leq \sqrt{\frac{\ln(\frac{2}{\delta})}{2n'_S}}.$$

33

Observe that despite the fact that n'_S is not fixed before we draw the sample, we can still apply a Chernoff bound here because for *every* realized value of n'_S, the distribution, conditional on the value of n'_S, of points (x, y) such that $(x, y) \in S$ remains a product distribution, with individual such points distributed as $\mathcal{P}|x \in S$. Now, we go through each scenario:

1. **ConsistencyAuditor outputs some set S and λ**: In this case, S would have been returned only if $n'_S > 0$ due to the if condition in Algorithm 6. Therefore, D must be approximately close to $\mathcal{P}$ with respect to $(S, \bar{\ell}, \ell)$. By Lemma 18, we have

$$|\bar{\ell}(S) - \ell(S)| \geq \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} \quad \text{and} \quad \lambda = \text{sign}(\bar{\ell}(S) - \ell(S))$$

2. **ConsistencyAuditor outputs $NULL$**: For any set S, $\mathcal{P}_{\mathcal{X}}(S) < \alpha$ directly implies that

$$|\bar{\ell}(S) - \ell(S)| \leq 1 < \frac{\alpha}{\mathcal{P}_{\mathcal{X}}(S)} \leq \frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)}.$$

Therefore, we focus only on sets S where $\mathcal{P}_{\mathcal{X}}(S) \geq \alpha$. For these sets, we have $n'_S > 0$ because

$$\frac{n'_S}{n} \geq \mathcal{P}_{\mathcal{X}}(S) - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \geq \alpha - \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} > \sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} > 0,$$

as we assumed $\alpha > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}$. Therefore, for every set S where $\mathcal{P}_{\mathcal{X}}(S) \geq \alpha$, we must have that D must be approximately close to $\mathcal{P}$ with respect to $(S, \bar{\ell}, \ell)$. Thus, by applying Lemma 18 to these sets S, we have

$$|\bar{\ell}(S) - \ell(S)| \leq \frac{\alpha'}{\mathcal{P}_{\mathcal{X}}(S)},$$

where $\alpha' = \alpha + 4\sqrt{\frac{1}{2n}\ln(\frac{2}{\delta})} + \left(\alpha - 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}\right).$

∎

Theorem 20 *Let T be the final iterate of Algorithm 9. If $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \leq \alpha$ and $2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \leq \beta$, we have the following guarantees:*

1. *$Total Iterations$: With probability $1 - 3\delta|\mathcal{G}|Q_\alpha\left((m^2 + m) + m^2 Q_\beta\right)$ over the randomness of our samples, the final iterate T is s.t. $T \leq \frac{1}{\alpha^2} - 1$ and the total number of gradient descent update operations will be at most Q, where*

$$Q_\alpha = \frac{1}{\alpha^2} - 1, Q_\beta = (k - 1)\left(\frac{1}{\beta^2} - 1\right), Q = Q_\alpha(1 + Q_\beta).$$

In particular, the algorithm uses at most nQ samples from $\mathcal{P}$.

2. **Mean multicalibration**: *With probability $1 - 3\delta(m^2 + m)|\mathcal{G}|$, output $\overline{\mu}^T$ is α'-mean multicalibrated with respect to $\mathcal{G}$ where*

$$\alpha' = \alpha + 4\sqrt{\frac{\ln\left(\frac{2}{\tilde{\delta}}\right)}{2n}} + \left(\alpha - 2\sqrt{\frac{\ln\left(\frac{2}{\tilde{\delta}}\right)}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\tilde{\delta}})}{2n}}\right).$$

3. **Mean Conditioned Moment multicalibration**: *With probability $1 - 3\delta|\mathcal{G}|(km^2 + m)$, for any $a \in \{2, \ldots, k\}$, pair $(\overline{\mu}^T, \overline{m}_a^T)$ is $(\alpha', a\alpha' + \beta', \frac{a}{m})$-mean-conditioned-moment calibrated where*

$$\alpha' = \alpha + 4\sqrt{\frac{\ln\left(\frac{2}{\tilde{\delta}}\right)}{2n}} + \left(\alpha - 2\sqrt{\frac{\ln\left(\frac{2}{\tilde{\delta}}\right)}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\tilde{\delta}})}{2n}}\right)$$

$$\beta' = \beta + 4\sqrt{\frac{\ln\left(\frac{2}{\tilde{\delta}}\right)}{2n}} + \left(\beta - 2\sqrt{\frac{\ln\left(\frac{2}{\tilde{\delta}}\right)}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln(\frac{2}{\tilde{\delta}})}{2n}}\right)$$

Proof We prove each guarantee in turn.

Total Iterations: As argued in Theorem 11, if the auditor can successfully find a set S on which there is α-mean inconsistency and β-pseudo-moment inconsistency respectively in AlternatingGradientDescentFinite (Algorithm 9) and PseudoMomentConsistencyFinite (Algorithm 8), Theorem 7 guarantees that T will be at most $\frac{1}{\alpha^2} - 1$ and Theorem 9 guarantees that the total number of gradient descent operations in each PseudoMomentconsistencyFinite will be at most $\frac{1}{\beta^2} - 1$. Then, because in each iteration of AlternatingGradientDescentFinite, there are $k - 1$ calls to PseudoMomentconsistencyFinite, the total number of number of gradient descent operations will be at most $Q = Q_\alpha(1 + Q_\beta)$ where $Q_\alpha = \frac{1}{\alpha^2} - 1$ and $Q_\beta = (k-1)(\frac{1}{\beta^2} - 1)$.

Therefore, it is sufficient for us to show that there is α-mean inconsistency and β-pseudo-moment inconsistency on every S^t and R returned by ConsistencyAuditor (Algorithm 7) for AlternatingGradientDescentFinite and PseudoMomentconsistencyFinite respectively.

For AlternatingGradientDescentFinite, because we set $\overline{\ell}^t(x) = \overline{\mu}^t(x)$, $\ell(x, y) = y$, and $\mathcal{S}^t = \{G(\overline{\mu}, i) : G \in \mathcal{G}, i \in [m]\} \cup \{G(\overline{\mu}, \overline{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m], a \in \{2, \ldots, k\}\}$, Corollary 19 guarantees that with probability $1 - 3\delta|\mathcal{G}|(m^2 + m)$, $\overline{\mu}^t$ is α-mean inconsistent on S^t with n $\lambda^t = \text{sign}(\overline{\mu}^t(S^t) - \mu(S^t))$ as desired. Because T is at most Q_α, by a union bound, $\overline{\mu}^t$ is α-mean-inconsistent on S^t for every $t \in [T]$ with probability $1 - 3\delta|\mathcal{G}|(m^2 + m)Q_\alpha$.

Likewise, for PseudoMomentconsistencyFinite, we set $\overline{\ell}(x) = \overline{m}_k(x)$, $\ell(x, y) = (y - \overline{\mu}(x))^a$, and $\mathcal{S} = \{G(\overline{\mu}, \overline{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m], a \in \{2, \ldots, k\}\}$. Hence, by union bounding over every $a \in \{2, \ldots, k\}$, Corollary 19 promises us that with probability $1 - 3\delta|\mathcal{G}|m^2 Q_\alpha Q_\beta$, $\overline{m}_a$ is β-pseudo-moment inconsistent on R throughout every iteration of PseudoMomentConsistencyFinite for every $a \in \{2, \ldots, k\}$ and $\psi = \text{sign}(\overline{m}_a(S) - \widetilde{m}_{a, \overline{\mu}}(S))$ as desired. Note that there are a total of Q_β calls to ConsistencyAuditor from each PseudoMomentConsistencyFinite, which is invoked a total of Q_α many times.

Mean Multi-Calibration: Our algorithm halts only if ConsistencyAuditor doesn't find S in AlternatingGradientDescentFinite. Corollary 19 promises us that with probability $1 - 3\delta(m^2 + m)|\mathcal{G}|$, $\overline{\mu}^T$

must be α'-mean-consistent on every set $S \in \mathcal{S}^T$. Because $\mathcal{S}$ includes $\{G(\overline{\mu}^T, i) : G \in \mathcal{G}, i \in [m]\}$, it must be that $\overline{\mu}^T$ is α'-mean multi-calibrated with respect to $\mathcal{G}$.

Mean Conditioned Moment Multi-Calibration: In the last round T, consider each $\overline{m}_a^T$ for $a \in \{2, \ldots, k\}$. PseudoMomentConsistencyFinite returns $\overline{m}_a^T$ only if ConsistencyAuditor doesn't return any R. Corollary 19 guarantees us that with probability $1 - 3\delta m^2 |\mathcal{G}|$, $\overline{m}_a^T$ must be β'-pseudo-moment-consistent. Because $\overline{\mu}^T$ is α'-mean consistent and $\overline{m}_a^T$ is pseudo-moment-consistent on $\{G(\overline{\mu}, \overline{m}_a, i, j) : G \in \mathcal{G}, i, j \in [m], a \in \{2, \ldots, k\}\}$, Lemma 10 tells us that $(\overline{\mu}^T, \overline{m}_a^T)$ must be $(\alpha', a\alpha + \beta', \frac{a}{m})$-mean-conditioned-moment multicalibrated. By union bounding over each $a \in \{2, \ldots, k\}$ the total failure probability is $1 - 3\delta|\mathcal{G}|(km^2 + m)$. ∎

Corollary 35 *Fix target parameters α', β', δ' and $\epsilon > 0$ such that $\epsilon < \alpha'$ and $\epsilon < \beta'$. Define*

$$\overline{Q} = \frac{6|\mathcal{G}|km^2}{\left(\frac{\alpha'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2 \left(\frac{\beta'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}, \quad \delta = \frac{\delta'}{\max(3|\mathcal{G}|(km^2+m), \overline{Q})}, \quad n_\alpha = \frac{\ln(\frac{2\overline{Q}}{\delta})}{2\left(\frac{\alpha'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}, \quad n_\beta = \frac{\ln(\frac{2\overline{Q}}{\delta})}{2\left(\frac{\beta'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}.$$

Then, AlternatingGradientDescentFinite$(\alpha, \beta, \delta, n, \mathcal{G})$ where

$$\alpha = 2\sqrt{\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\alpha}} + \epsilon, \beta = 2\sqrt{\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\beta}} + \epsilon,$$

$$n = \max\left(\frac{\ln(\frac{2\overline{Q}}{\delta})}{\ln(\frac{2}{\delta})}n_\alpha, \frac{\ln(\frac{2\overline{Q}}{\delta})}{\ln(\frac{2}{\delta})}n_\beta, \frac{2\ln(\frac{2}{\delta})}{\alpha^2}, \frac{2\ln(\frac{2}{\delta})}{\beta^2}\right)$$

has the following guarantees with probability $1 - \delta'$:

1. *The total number of gradient descent updates will be at most Q, where Q is as defined in Theorem 20.*

2. *$\overline{\mu}^T$ is α'-mean-calibrated.*

3. *For every $a \in \{2, \ldots, k\}$, $(\overline{\mu}^T, \overline{m}_a^T)$ is $(\alpha', a\alpha' + \beta', \frac{a}{m})$-mean-conditioned-moment calibrated.*

Proof Note that by construction, we have

$$\alpha > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}} \quad \text{and} \quad \beta > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}.$$

36

Therefore, in Theorem 20, the level of mean calibration for $\overline{\mu}^T$ will be

$$
\alpha + 4\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} + \left(\alpha - 2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\right)^{-2}\left(2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}\right)
$$

$$
\leq \alpha + 4\sqrt{\frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{2n_\alpha}} + \left(\alpha - 2\sqrt{\frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{2n_\alpha}}\right)^{-2}\left(2\sqrt{\frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{2n_\alpha}}\right)
$$

$$
= 2\sqrt{\frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{2n_\alpha}} + \epsilon + 4\sqrt{\frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{2n_\alpha}} + \frac{\left(2\sqrt{\frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{2n_\alpha}}\right)}{\epsilon^2}
$$

$$
= \sqrt{\frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{2n_\alpha}}\left(6 + \frac{2}{\epsilon^2}\right) + \epsilon
$$

$$
= \alpha',
$$

where the first inequality follows because $n \geq \frac{\ln\left(\frac{2\overline{Q}}{\delta}\right)}{\ln\left(\frac{2}{\delta}\right)}n_\alpha$ and the last equality from the definition of n_α.

Applying the same analysis, we can show that we satisfy pseudo-moment-consistency at level β'. Therefore, for any $a \in \{2, \ldots, k\}$, $(\overline{\mu}^t, \overline{m}_a^T)$ satisfy $(\alpha', a\alpha' + \beta', \frac{a}{m})$-mean-conditioned-moment multicalibration.

The failure probabilities for mean muticalibration and that of mean-conditioned-moment multicalibration are both less than δ', as $3\delta(m^2+m)|\mathcal{G}| \leq 3\delta|\mathcal{G}|(km^2+m) \leq \delta'$ and $3\delta|\mathcal{G}|(km^2+m) \leq \delta'$.

The failure probability for termination is

$$3\delta|\mathcal{G}|\left(\frac{1}{\alpha^2}-1\right)\left((m^2+m)+m^2(k-1)\left(\frac{1}{\beta^2}-1\right)\right)$$

$$\leq 3\delta|\mathcal{G}|\left(\frac{1}{\alpha^2}\right)\frac{2km^2}{\beta^2}$$

$$= 6|\mathcal{G}|km^2\delta\cdot\frac{1}{\left(\sqrt{\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\alpha}}+\epsilon\right)^2\left(\sqrt{\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\beta}}+\epsilon\right)^2}$$

$$\leq 6|\mathcal{G}|km^2\delta\cdot\frac{1}{\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\alpha}\frac{\ln(\frac{2\overline{Q}}{\delta})}{2n_\beta}}$$

$$\leq 24|\mathcal{G}|km^2\delta\cdot\frac{1}{\left(\ln(\frac{2\overline{Q}}{\delta})\right)^2}n_\alpha n_\beta$$

$$= 24|\mathcal{G}|km^2\delta\cdot\frac{1}{\left(\ln(\frac{2\overline{Q}}{\delta})\right)^2}\frac{\ln(\frac{2\overline{Q}}{\delta})}{2\left(\frac{\alpha'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}\frac{\ln(\frac{2\overline{Q}}{\delta})}{2\left(\frac{\beta'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}$$

$$\leq 6|\mathcal{G}|km^2\delta\cdot\frac{1}{\left(\frac{\alpha'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2\left(\frac{\beta'-\epsilon}{6+\frac{2}{\epsilon^2}}\right)^2}$$

$$= \overline{Q}\delta$$

$$\leq \delta'$$

∎

Theorem 22 *With probability* $1-3\delta|\mathcal{G}|Q_\alpha\left((m^2+m)+m^2Q_\beta\right)$*, the running time of Algorithm* 9 *is* $O\left(Q|\mathcal{G}|m^2n\right)=O\left(\frac{k|\mathcal{G}|m^2n}{\alpha^2\beta^2}\right)$ *where* Q_α, Q_β*, and* Q *are as defined in Theorem* 20*.*

Proof Theorem 20 tells us that except with probability $1-3\delta|\mathcal{G}|Q_\alpha\left((m^2+m)+m^2Q_\beta\right)$, the algorithm will halt after at most Q many gradient descent updates. For each gradient descent update, it must have been that Algorithm 7 was invoked with either $\mathcal{S}=\{G(\overline{\mu},i):G\in\mathcal{G},i\in[m]\}\cup\{G(\overline{\mu},\overline{m}_a,i,j):G\in\mathcal{G},i,j\in[m],a\in\{2,\ldots,k\}\}$ or $\mathcal{S}=\{G(\overline{\mu},\overline{m}_a,i,j):G\in\mathcal{G},i,j\in[m],a\in\{2,\ldots,k\}\}$. Note that Algorithm 7 needs to iterate through each set S in $\mathcal{S}$, whose size is at most $O(|\mathcal{G}|m^2)$ in either case. And processing each set S through Algorithm 6 requires finding the average of at most $O(n)$ elements twice. Therefore, the algorithm will take time $O\left(Q|\mathcal{G}|m^2n\right)$ with probability $1-3\delta|\mathcal{G}|Q_\alpha\left((m^2+m)+m^2Q_\beta\right)Q$. ∎

$D = \{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n$
$D^{\text{check}} \sim \mathcal{P}^n$
$\mathcal{S} = \{\mathcal{X}(\overline{\mu}, \overline{m}_a, i, j) : i, j \in [m]\}$
$\overline{\ell}(x) = \overline{m}_a(x)$
$\ell(x, y) = (y - \overline{\mu}(x))^a$
$R, \psi = \text{LearningOracleConsistencyAuditorWrapper}(\overline{\ell}, \ell, \beta, \delta, D, D^{\text{check}}, \mathcal{S}, A)$
while $R, \psi \neq NULL$ **do**

$\qquad \overline{m}_a(x) = \begin{cases} \text{project}_{[0,1]}(\overline{m}_a(x) - \beta\psi) & \text{if } x \in R \\ \overline{m}_a(x) & \text{otherwise.} \end{cases}$

$\qquad D = \{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n$
$\qquad \mathcal{S} = \{\mathcal{X}(\overline{\mu}, \overline{m}_a, i, j) : i, j \in [m]\}$
$\qquad \overline{\ell}(x) = \overline{m}_a(x)$
$\qquad D = \{(x_b, y_b)\}_{b=1}^n \sim \mathcal{P}^n$
$\qquad D^{\text{check}} \sim \mathcal{P}^n$
$\qquad R, \psi = \text{LearningOracleConsistencyAuditorWrapper}(\overline{\ell}, \ell, \beta, \delta, D, D^{\text{check}}, \mathcal{S}, A)$
return $\overline{m}_k$

Algorithm 12: PseudoMomentConsistencyWithOracle$(a, \beta, \delta, \overline{\mu}, \overline{m}_a, n, \mathcal{G})$

initialize $\overline{\mu}^1(x) = 0$ for all x
for all $1 < a \leq k$, initialize $\overline{m}_a^1(x) = 0$ for all x
$t = 1$
$\overline{\ell}^t(x) = \overline{\mu}(x)$
$\ell(x, y) = y$
$D^t = \sim \mathcal{P}^n$
$D^{\text{check}^t} \sim \mathcal{P}^n$
$\mathcal{S}^t = \{\mathcal{X}(\overline{\mu}^t, i) : i \in [m]\} \cup \{\mathcal{X}(\overline{\mu}^t, \overline{m}_a^t, i, j) : i, j \in [m], a \in \{2, \ldots, k\}\}$
$S^t, \lambda^t = \text{LearningOracleConsistencyAuditorWrapper}(\overline{\ell}^t, \ell, \beta, \delta, D^t, D^{\text{check}^t}, \mathcal{S}^t, A)$
while $S^t, \lambda^t \neq NULL$ **do**
$\quad \overline{\mu}^{t+1} = \text{MeanConsistencyUpdate}(\overline{\mu}^t, S^t, \lambda^t)$
$\quad$ **for** $a = 2, \ldots, k$ **do**
$\quad\quad \overline{m}_a^{t+1} = \text{PseudoMomentConsistencyFinite}(a, \beta, \delta, \overline{\mu}^{t+1}, \overline{m}_a^t, n, \mathcal{G}).$
$\quad t = t + 1$
$\quad \overline{\ell}^t(x) = \overline{\mu}(x)$
$\quad D^t \sim \mathcal{P}^n$
$\quad D^{\text{check}^t} \sim \mathcal{P}^n$
$\quad \mathcal{S}^t = \{\mathcal{X}(\overline{\mu}^t, i) : i \in [m]\} \cup \{\mathcal{X}(\overline{\mu}^t, \overline{m}_a^t, i, j) : i, j \in [m], a \in \{2, \ldots, k\}\}$
$\quad S^t, \lambda^t = \text{LearningOracleConsistencyAuditorWrapper}(\overline{\ell}^t, \ell, \beta, \delta, D^t, D^{\text{check}^t}, \mathcal{S}^t, A)$
return $(\overline{\mu}^t, \{\overline{m}_a^t\}_{a=2}^k)$

Algorithm 13: AlternatingGradientDescentWithOracle$(\alpha, \beta, \delta, n, \mathcal{G}, A)$

Lemma 36 *For each $R \in \mathcal{R}$ and any χ_S:*

$$\mathbb{E}_{(x,y)}[\chi_S(x) \cdot r_R^+(x, y)] = \mathcal{P}_{\mathcal{X}}(R \cap S)\left(\overline{\ell}(R \cap S) - \ell(R \cap S)\right)$$

$$\mathbb{E}_{(x,y)}[\chi_S(x) \cdot r_R^-(x,y)] = \mathcal{P}_\mathcal{X}(R \cap S)\left(\ell(R \cap S) - \bar{\ell}(R \cap S)\right)$$

Proof

$$\mathbb{E}_{(x,y)}[\chi_S(x) \cdot r_R^+(x,y)] = \sum_{x,y} \mathcal{P}(x,y)\chi_S(x)r_R^+(x,y)$$

$$= \sum_{x,y: x \in S, x \in R} \mathcal{P}(x,y)(\bar{\ell}(x) - \ell(x,y))$$

$$= \mathcal{P}_\mathcal{X}(R \cap S)\left(\bar{\ell}(R \cap S) - \ell(R \cap S)\right)$$

The same argument applies for r_R^- as well. ■

Theorem 26 *Assume n is sufficiently large such that $\alpha > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}$. Algorithm 11 has the following guarantees:*

1. *If it returns some S and λ, then with probability $1 - 3\delta|\mathcal{R}|$ over the randomness of D^{check},*

$$|\bar{\ell}(S) - \ell(S)| \geq \frac{\alpha}{\mathcal{P}_\mathcal{X}(S)}.$$

2. *If it returns $NULL$, then with probability $1 - |\mathcal{R}|(3\delta + 2p(n))$ over the randomness of D and D^{check}, for all $\chi_S \in \mathcal{H}$ and $R \in \mathcal{R}$,*

$$|\bar{\ell}(R \cap S) - \ell(R \cap S)| \leq \frac{\alpha' + \rho}{\mathcal{P}_\mathcal{X}(R \cap S)},$$

where α' is as defined in Corollary 19.

Proof With probability at least $1 - 3|\mathcal{R}|\delta$ (since $|\mathcal{R}| \geq |\mathcal{V}|$), Corollary 19 gives us the following guarantees:

1. If S, λ is returned, then

$$\left|\bar{\ell}(S) - \ell(S)\right| \geq \frac{\alpha}{\mathcal{P}_\mathcal{X}(S)} \quad \text{and} \quad \lambda = \text{sign}(\bar{\ell}(S) - \ell(S)).$$

2. If $NULL$ is returned, then for all $V \in \mathcal{V}$,

$$|\bar{\ell}(V) - \ell(V)| \leq \frac{\alpha'}{\mathcal{P}_\mathcal{X}(V)} \tag{4}$$

Whenever the distributional closeness conditions of Definition 16 hold (which occur with the same $1 - 3|\mathcal{R}|\delta$ success probability of Corollary 19), and $\alpha > 2\sqrt{\frac{\ln(\frac{2}{\delta})}{2n}}$, it must be that if $|D_R| = 0$ then $\mathcal{P}_\mathcal{X}(R) \leq \alpha$. And for any such R we have that $\mathcal{P}_\mathcal{X}(R \cap S') \leq \alpha$ for any other set S', which implies that we trivially satisfy $(\alpha' + \rho)$-mean consistency for $R \cap S'$. More precisely, if $\mathcal{P}_\mathcal{X}(R) \leq \alpha$, then

$$\sup_{\chi_{S'} \in \mathcal{H}} |\bar{\ell}(R \cap S') - \ell(R \cap S')| \leq 1 \leq \frac{\alpha}{\mathcal{P}_\mathcal{X}(R \cap S')} \leq \frac{\alpha' + \rho}{\mathcal{P}_\mathcal{X}(R \cap S')}.$$

We can therefore restrict our attention to those $R \in \mathcal{R}$ that satisfy $|D_R| > 0$, since we only have a non-trivial statement to prove for sets R with $\mathcal{P}_{\mathcal{X}}(R) > \alpha$. Using Lemma 25 and the definition of an agnostic learning oracle, we know that for each $V = R \cap S^+$, with probability $1 - p(n)$,

$$
\begin{aligned}
&\mathcal{P}_{\mathcal{X}}(R \cap S^+) \left(\overline{\ell}(R \cap S^+) - \ell(R \cap S^+)\right) + \rho \\
&= \mathbb{E}_{(x,y)}[\chi_{S^+}(x) \cdot r_R^+(x,y)] + \rho \\
&\geq \sup_{\chi_{S'} \in \mathcal{H}} \mathbb{E}_{(x,y)}[\chi_{S'}(x) \cdot r_R^+(x,y)] \\
&= \sup_{\chi_{S'} \in \mathcal{H}} \mathcal{P}_{\mathcal{X}}(R \cap S') \left(\overline{\ell}(R \cap S') - \ell(R \cap S')\right)
\end{aligned}
\tag{5}
$$

The same argument applies for $V = R \cap S^-$, and we obtain

$$
\mathcal{P}_{\mathcal{X}}(R \cap S^+) \left(\ell(R \cap S^-) - \overline{\ell}(R \cap S^-)\right) + \rho \geq \sup_{\chi_{S'} \in \mathcal{H}} \mathcal{P}_{\mathcal{X}}(R \cap S') \left(\ell(R \cap S') - \overline{\ell}(R \cap S')\right).
\tag{6}
$$

Combining (4), (5), and (6), we get that with probability $1 - 2|\mathcal{R}|p(n)$,

$$
\sup_{\chi_{S'} \in \mathcal{H}} |\overline{\ell}(R \cap S') - \ell(R \cap S')| \leq \frac{\alpha' + \rho}{\mathcal{P}_{\mathcal{X}}(R \cap S')}
$$

■

Theorem 28 *With probability at least $1 - 3\delta Q_\alpha \left((m^2 + m) + m^2 Q_\beta\right)$, the running time of Algorithm 13 is bounded by $O(Qm^2\tau(n))$, where Q is the total number of gradient descent operations as defined in Theorem 20.*

Proof The running time of Algorithm 11 is $O(m^2\tau(n))$, as we always call it with $|\mathcal{R}| = O(m^2)$ and we assumed $\tau(n) = \Omega(n)$, meaning the running time of the learning oracle dominates the calculations in the empirical check. And Theorem 27 gives that with probability $1 - 3\delta Q_\alpha \left((m^2 + m) + m^2 Q_\beta\right)$, there will be at most Q gradient descent operations. Because the number of gradient descent operations is equal to the number of subroutine calls to Algorithm 11, the overall running time is $O(Qm^2\tau(n))$. ■

Appendix F. Marginal Prediction Intervals

We now present an application of our results. Given subgroups of interest $\mathcal{G}$, we have shown how to to construct a multicalibrated mean predictor $\overline{\mu}$ and moment predictors $(\overline{m}_a)_{a=2}^k$ that are simultaneously mean-conditioned moment-multicalibrated. A key question is whether we can use mean-conditioned moment multicalibrated predictors in applications in which we would use real distributional moments, were they available.

In this section, we show that the answer is *yes* in an important application. Mean-conditioned moment multicalibrated predictors can be used in tail bounds just as real moments could be to compute prediction intervals. Where real moments would yield prediction intervals conditioned on an individual vector of features x, mean-conditioned moment-multicalibrated predictors when used

in the same computations yield marginal prediction intervals that are simultaneously valid for every sufficiently large subgroup. In particular, given a coverage failure probability δ and a group size γ we show how to construct just from mean and moment predictions, for every $x \in X$, an interval $I(x, \gamma)$ such that for every $G \in \mathcal{G}$ and for every pair of predictions i, j such that $G(\overline{\mu}, \overline{m}_a, i, j)$ has mass at least γ we have: $\Pr_{(x,y)}[y \in I(x, \gamma) | x \in G(\overline{\mu}, \overline{m}_a, i, j)] \geq 1 - \delta$.

Recall the following tail inequality (a simple consequence of Markov's inequality: when $k = 2$, it is known as Chebyshev's inequality):

Lemma 37 *Let X be a random variable with mean μ. Then for even k, $t > 0$:*

$$\Pr[|X - \mu| \geq t] \leq \frac{\mathbb{E}[(X - \mu)^k]}{t^k}.$$

Suppose we knew the *real* moments $m_k(x)$ of the distribution on y conditional on features x: A direct application of the above lemma would allow us to conclude that for any even moment k:

$$\Pr\left[y \notin \left[\mu(x) - \left(\frac{m_k(x)}{\delta}\right)^{\frac{1}{k}}, \mu(x) + \left(\frac{m_k(x)}{\delta}\right)^{\frac{1}{k}}\right] \middle| x\right] \leq \delta.$$

Bounds of this form are simple, but also strong: there is always an integer moment k such that the above bound is at least as tight as a generalized Chernoff bound[7] (Philips and Nelson, 1995), and only the first $k \leq O(\log(1/\delta))$ moments are necessary to match Chernoff bounds at coverage probability $1 - \delta$ (Schmidt et al., 1995).

If we had *exactly* mean-conditioned moment-calibrated predictors $(\overline{\mu}, \overline{m}_k)$ for some k even, over a set of groups G, we would obtain *exactly* the same bound using these predictors as a marginal prediction interval: i.e. we would obtain for every $G \in \mathcal{G}$, and every i, j:

$$\Pr_{(x,y)}\left[y \notin \left[\overline{\mu}(x) - \left(\frac{\overline{m}_k(x)}{\delta}\right)^{1/k}, \overline{\mu}(x) + \left(\frac{\overline{m}_k(x)}{\delta}\right)^{1/k}\right] \middle| x \in G(\overline{\mu}, \overline{m}_k, i, j)\right] \leq \delta.$$

This is because mean-conditioned moment-multicalibrated predictors actually do provide real distributional moments, over the selection of a random point within any set $G(\overline{\mu}, \overline{m}_k, i, j)$. Of course we only have approximately mean-conditioned-moment multicalibrated predictors. Given $(\alpha, \beta, \epsilon)$-mean-conditioned-moment multicalibrated predictors $(\overline{\mu}, \overline{m}_k)$, k even, with respect to some collection of groups $\mathcal{G}$, we can endow our predictions with (marginal) prediction intervals that have coverage probability $1 - \delta$ as follows. The *width* of our prediction interval for a point x will be:

$$\Delta_{\gamma, k}(x) = \frac{\alpha}{\gamma} + \epsilon + \frac{1}{m} + \left(\frac{\overline{m}_k(x) + \epsilon + \frac{1}{m} + \frac{\beta}{\gamma}}{\delta}\right)^{\frac{1}{k}},$$

Our prediction interval for x will be centered at its predicted mean, and is defined as follows:

$$I_{\gamma, k}(x) = [\overline{\mu}(x) - \Delta_{\gamma, k}(x), \overline{\mu}(x) + \Delta_{\gamma, k}(x)].$$

These are valid marginal prediction intervals as averaged over *any* set of the form $G(\overline{\mu}, \overline{m}_k, i, j)$ that has measure larger than γ. Note that all of the approximation terms $\alpha, \beta, \epsilon, 1/m$ are terms that we can drive to zero at polynomial cost in running time and sample complexity.

---

7. Chernoff's bound is $\Pr[X \geq t] \leq \inf_{\theta \geq 0} M_X(\theta) e^{-\theta t}$, where $M_X(\theta)$ is the moment generating function for X

Theorem 38 *Assume that $\overline{\mu}, \overline{m}_k$ is $(\alpha, \beta, \epsilon)$-mean-conditioned moment multicalibrated with respect to $\mathcal{G}$, with k even. Then for any group $G \in \mathcal{G}$ and any set $G(\overline{\mu}, \overline{m}_k, i, j)$ such that $\mathcal{P}_{\mathcal{X}}[G(\overline{\mu}, \overline{m}_k, i, j)] \geq \gamma$, we have:*

$$\mathcal{P}[y \notin I_{\gamma,k}(x) | x \in G(\overline{\mu}, \overline{m}_k, i, j)] \leq \delta$$

Proof To see this note that:

$$\mathcal{P}[y \notin I_{\gamma,k}(x) | x \in G(\overline{\mu}, \overline{m}_k, i, j)]$$

$$= \mathcal{P}\left[|y - \overline{\mu}(x)| \geq \frac{\alpha}{\gamma} + \frac{1}{m} + \epsilon + \left(\frac{\overline{m}_k(x) + \frac{1}{m} + \epsilon + \frac{\beta}{\gamma}}{\delta} \right)^{\frac{1}{k}} \middle| x \in G(\overline{\mu}, \overline{m}_k, i, j) \right]$$

$$\leq \mathcal{P}\left[|y - \overline{\mu}(G(\overline{\mu}, \overline{m}_k, i, j))| \geq \frac{\alpha}{\gamma} + \epsilon + \left(\frac{\overline{m}_k(G(\overline{\mu}, \overline{m}_k, i, j)) + \frac{\beta}{\gamma} + \epsilon}{\delta} \right)^{\frac{1}{k}} \middle| x \in G(\overline{\mu}, \overline{m}_k, i, j) \right]$$

$$\leq \mathcal{P}\left[|y - \mu(G(\overline{\mu}, \overline{m}_k, i, j))| \geq \left(\frac{m_k(G(\overline{\mu}, \overline{m}_k, i, j))}{\delta} \right)^{\frac{1}{k}} \middle| x \in G(\overline{\mu}, \overline{m}_k, i, j) \right]$$

$$\leq \delta$$

Here, the first inequality follows from the fact that all $x \in G(\overline{\mu}, \overline{m}_k, i, j)$ are (by definition) such that $|\overline{\mu}(x) - \frac{i}{m}| \leq \frac{1}{2m}$ and $|\overline{m}_k(x) - \frac{j}{m}| \leq \frac{1}{2m}$, and hence $|\overline{\mu}(x) - \overline{\mu}(G(\overline{\mu}, \overline{m}_k, i, j))| \leq \frac{1}{m}$ and $|\overline{m}_k(x) - \overline{m}_k(G(\overline{\mu}, \overline{m}_k, i, j)))| \leq \frac{1}{m}$. The second inequality follows from the definition of (α, β)-mean conditioned moment multicalibration and the fact that $\mathcal{P}[G(\overline{\mu}, \overline{m}_k, i, j)] \geq \gamma$. Finally, once we have replaced our mean and moment estimates with the true mean and moment of $G(\overline{\mu}, \overline{m}_k, i, j)$, the final inequality follows as an application of Lemma 37. ∎

This theorem shows how—given $(\alpha, \beta, \epsilon)$ mean-conditioned moment-multicalibrated predictors $\overline{\mu}, \overline{m}_k$—we can construct prediction intervals for any set $G(\overline{\mu}, \overline{m}_k, i, j)$ with probability larger than γ.[8] However, we have more information available to us: We have mean conditioned moment-calibrated predictors for all moments 2 thru k, $(\overline{m}_a)_{a=2}^k$. A straightforward valid solution is to pick some even moment a s.t. $1 < a \leq k$, and then construct prediction intervals as above. We could optimize our choice of a so as to minimize e.g. the expected width of the prediction intervals over a random choice of x. But this leads to the question of whether we can do better by using more than one moment estimator at a time. In Appendix G we show that this problem reduces to the venerable submodular-cost set-cover problem. Known approximation guarantees for this problem are relatively weak in this context (scaling with $\log |\mathcal{X}|$, which will typically be linear in the *dimension* of the data). We leave the question of how to optimally use multiple mean-conditioned moment multicalibrated predictors—taking advantage of multiple moments simultaneously—to future research.

---

8. We showed this just for *even* moments k — but a version of Lemma 37 also holds for k odd, i.e. for any r.v. X with mean μ, any number k, and any $t > 0$, we have $\Pr[|X - \mu| \geq t] \leq \frac{\mathbb{E}[|(X - \mu)^k|]}{t^k}$. We can use this to construct valid confidence intervals using "absolute central moments" of any degree, even or odd. Note also that our algorithms and analysis apply identically if the goal was to provide mean-conditioned, multicalibrated predictors of absolute central moments (i.e. the analog of Definition 4 but where instead of $m_k(\cdot)$, we calibrate our predictor to the analogous absolute central moment).

Appendix G. A Submodular Set-Cover Formulation

We can define the following problem. Theorem 38 shows us that for every even a, and every $G \in \mathcal{G}$, $i, j \in [m]$, $I_{\gamma,a}(x)$ forms a valid marginal prediction interval for every set $G(\overline{\mu}, \overline{m}_a, i, j)$ with probability at least γ under $\mathcal{P}_{\mathcal{X}}$. Can we construct tighter prediction intervals using all $\lfloor \frac{k}{2} \rfloor$ moments?

We make the following simplifying assumptions in this section:

1. $\mathcal{X}$ is a set with finite cardinality.

2. $\mathcal{P}_{\mathcal{X}}$ is known exactly (note that we do not assume we know the distribution on *labels* y, which preserves the core motivation of the problem).

3. For every $x \in \mathcal{X}$, there exists $G \in \mathcal{G}$, a even s.t. $1 < a \le k$, and $i, j \in [m]$ such that $x \in G(\overline{\mu}, \overline{m}_a, i, j)$ and $\mathcal{P}_{\mathcal{X}}(G(\overline{\mu}, \overline{m}_a, i, j)) \ge \gamma$ (otherwise there is no way to give a valid marginal prediction interval for such an x).

Let us define the set of all relevant sets as

$$\mathcal{S} \equiv \{G(\overline{\mu}, \overline{m}_a, i, j) : \forall G \in \mathcal{G}, i, j \in [m], 1 < a \le k, a \text{ even s.t. } \mathcal{P}_{\mathcal{X}}(G(\overline{\mu}, \overline{m}_a, i, j)) \ge \delta\}.$$

With each set $S \in \mathcal{S}$, we associate the width $\Delta_S(\cdot)$ in the obvious way.

Given any $\mathcal{S}' \subseteq \mathcal{S}$ we say that $\mathcal{S}'$ covers $\mathcal{X}$ if $\forall x \in \mathcal{X}$, $\exists S \in \mathcal{S}'$ s.t. $x \in S$. Given any $\mathcal{S}' \subseteq \mathcal{S}$ that covers $\mathcal{X}$, we can construct valid marginal prediction intervals for all $x \in X$:

$$\Delta_{\mathcal{S}'}(x) \equiv \max_{S \in \mathcal{S}'|x \in S} \Delta_S(x),$$
$$I_{\mathcal{S}'}(x) = [\overline{\mu}(x) - \Delta_{\mathcal{S}'}(x), \overline{\mu}(x) + \Delta_{\mathcal{S}'}(x)].$$

To see that this will result in a valid prediction interval, observe that for any $x \in X$, it is covered by some $S \in \mathcal{S}'$. By definition of $\mathcal{S}'$, $S = G(\overline{\mu}, \overline{m}_a, i, j)$ for some a even, $i, j \in [m]$, $G \in \mathcal{G}$. Note that $I_{\gamma,a}(x) \subseteq I_{\mathcal{S}'}(x)$ by construction of $I_{\mathcal{S}'}(\cdot)$. Therefore Theorem 38 ensures that these prediction intervals are valid for any $S \in \mathcal{S}'$, and indeed, therefore valid for any group $G \in \mathcal{G}$.

A natural optimization problem is to find a subset $\mathcal{S}'$ that covers $\mathcal{X}$ so as to minimize the expected width of the marginal prediction intervals that can be produced in this way, i.e. solves (exactly or approximately)

$$\min_{\mathcal{S}' \subseteq \mathcal{S}} \mathbb{E}_{x \sim \mathcal{P}_{\mathcal{X}}}[\Delta_{\mathcal{S}'}(x)]$$
$$\text{s.t. } \mathcal{S}' \text{ covers } \mathcal{X}.$$

We can rewrite the problem in the following way. Let A be a $0-1$ matrix of dimension $|\mathcal{X}| \times |\mathcal{S}|$. The columns correspond to sets $S \in \mathcal{S}$ and the rows to elements $x \in \mathcal{X}$. If $A_{xS} = 1$ this means that element $x \in \mathcal{X}$ is contained in set $S \in \mathcal{S}$. Associated with each column S there is a function Δ_S. Recall that $\mathcal{P}_{\mathcal{X}}(x)$ denotes the probability of x.

We can denote any subset of $\mathcal{S}' \subseteq \mathcal{S}$ by a $0/1$ vector $w \in \{0,1\}^{|\mathcal{S}|}$ such that $w_S = 1$ if $S \in \mathcal{S}'$. We can therefore recast the optimization problem:

$$
\begin{aligned}
&\min_{z,w} \sum_{x \in \mathcal{X}} \mathcal{P}_\mathcal{X}(x) z_x \\
&\text{s.t. } z_x - \Delta_S(x) A_{xS} w_S \geq 0 && \forall x \in \mathcal{X} \\
&\qquad \sum_{S \in \mathcal{S}} A_{xS} w_S \geq 1 && \forall x \in \mathcal{X}.
\end{aligned}
$$

For any subset $\mathcal{S}' \subseteq \mathcal{S}$ let $f_x(\mathcal{S}') = \max_{S \in \mathcal{S}'} \Delta_S(x) A_{xS}$. Notice $f_x(\mathcal{S}')$ is a non-decreasing and submodular function of $\mathcal{S}'$. Let $f(\mathcal{S}') = \sum_x \mathcal{P}_\mathcal{X}(x) f_x(\mathcal{S}')$, clearly $f(\cdot)$ is non-decreasing and submodular. Similarly, for any $\mathcal{S}' \subseteq \mathcal{S}$ let w be the associated $0/1$ vector and define $g(\mathcal{S}') = |\{x : \sum_{S \in \mathcal{S}} A_{xS} w_S \geq 1\}|$. Again $g(\cdot)$ is a non-decreasing and submodular function of $\mathcal{S}'$. Observe that can write our problem as:

$$
\begin{aligned}
&\min_{\mathcal{S}' \subseteq \mathcal{S}} f(\mathcal{S}'), \\
&\text{s.t. } g(\mathcal{S}') \geq |\mathcal{X}|.
\end{aligned}
$$

Therefore our problem is the submodular cost submodular cover problem.

We can now hope to apply known results to solve it. For example, Wan et al. (2010) show that the greedy solution (to iteratively add the set with the smallest average width) is approximately optimal. In particular, their Theorem 2.1 guarantees that the greedy solution provides a $\frac{k}{2}H$-approximate to our optimization problem where H is the ℓth-harmonic number, $\ell = \max\{|S| : S \in \mathcal{S}\}$, and k is the number of moments we have access to.

Unfortunately in this context, these guarantees are unsatisfactory: the approximation grows with the number of moments k we have access to, and with $\log |\mathcal{X}|$, which will typically be linear in the data dimension. Note that Wan et al. (2010) study general submodular objective functions and does not exploit the specific structure of the objective function here. We leave the question of whether guarantees can be offered for this problem to future research. Another natural question is how to approximate this optimization when $\mathcal{P}_\mathcal{X}$ is not known, i.e. we only have a finite sample from $\mathcal{P}$.