Reconciling Individual Probability Forecasts*

Aaron Roth
Department of Computer and Information Sciences, University of Pennsylvania
USA
aaroth@cis.upenn.edu

Alexander Tolbert
Department of Philosophy, University of Pennsylvania
USA
altol25@sas.upenn.edu

Scott Weinstein
Department of Philosophy, University of Pennsylvania
USA
weinstein@cis.upenn.edu

ABSTRACT

Individual probabilities refer to the probabilities of outcomes that are realized only once: the probability that it will rain tomorrow, the probability that Alice will die within the next 12 months, the probability that Bob will be arrested for a violent crime in the next 18 months, etc. Individual probabilities are fundamentally unknowable. Nevertheless, we show that two parties who agree on the data—or on how to sample from a data distribution—cannot agree to disagree on how to model individual probabilities. This is because any two models of individual probabilities that substantially disagree can together be used to empirically falsify *and improve* at least one of the two models. This can be efficiently iterated in a process of "reconciliation" that results in models that both parties agree are superior to the models they started with, and which themselves (almost) agree on the forecasts of individual probabilities (almost) everywhere. We conclude that although individual probabilities are unknowable, they are *contestable* via a computationally and data efficient process that must lead to agreement. Thus we cannot find ourselves in a situation in which we have two equally accurate and unimprovable models that disagree substantially in their predictions—providing an answer to what is sometimes called the predictive or model multiplicity problem.

ACM Reference Format:
Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2023. Reconciling Individual Probability Forecasts. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3593013.3593980

1 INTRODUCTION

Probabilistic modelling in machine learning and statistics predicts "individual probabilities" as a matter of course. In weather forecasting, we speak of the probability of rain tomorrow; in life insurance underwriting we speak of the probability that Alice will die in the next 12 months; in recidivism prediction we speak of the probability that an inmate Bob will commit a violent crime within 18 months of being released on parole; in predictive medicine we speak of the probability that Carol will develop breast cancer before the age of 50 — and so on. But these are not repeated events: we have no way of directly measuring an "individual probability" — and indeed, even the semantics of an individual probability are unclear and have been the subject of deep interrogation within the philosophy of science and statistics [13, 26] and theoretical computer science [17]. Within the philosophy of science, puzzles related to individual probability have been closely identified with "the reference class problem" [26]. This is a close cousin of a concern that has recently arisen in the context of fairness in machine learning called the "predictive multiplicity problem" (a focal subset of "model multiplicity problems") [8, 30] which [9] earlier called the "Rashomon Effect". At the core of both of these problems is the fact that from a data sample that is much smaller than the data universe (i.e. the set of all possible observations), we will have observed at most one individual with a particular set of characteristics, and at most one outcome for the event that an "individual probability" speaks to: It will either rain tomorrow or it will not; Alice will either die within the next year or she will not; etc. We do not have the luxury of observing a large number of repetitions and taking averages.

[13] lays out two broad classes of perspectives on individual probabilities: the *group to individual* perspective and the *individual to group* perspective. The group to individual perspective is roughly as follows: We cannot measure individual probabilities from data, but we *can* measure averages of outcomes within sufficiently large *reference classes* S. A reference class S is just some well defined subset of the observed data: for example (in the weather forecasting setting) the set of days in which there is cloud cover and humidity is above 60%, or (in the life insurance setting) the set of 65 year old women with a history of high blood pressure. Given a reference class that is large enough that we have observed in our data many members of the reference class, we can empirically estimate the prevalence of the outcome we are concerned with forecasting (rain, death within 12 months) for members of the reference class. Then, if we are asked to forecast an individual probability (the probability that *Alice* will die within the next 12 months), we simply pick an appropriate reference class S such that Alice $\in S$ and then respond with the proportion of observed deaths within a 12 month period for individuals from reference class S. The principal problem with this approach (known as the "reference class problem" [26]) is that Alice will simultaneously be a member of many different reference classes S. We cannot condition on *everything* we know about Alice, or we will end up with a reference class that does not contain enough examples for us do statistical inference on: thus we must pick and choose. But should we have conditioned on her age, gender and blood pressure? What about her weight? Her job? Her marital status? Her vaccination history? Defining reference classes with

*This is an abridged version of our paper: the full version can be found at [31].

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06…$15.00
https://doi.org/10.1145/3593013.3593980

respect to different subsets of these attributes will generally lead to different estimates for the probability that Alice will die within the next 12 months: what privileges one of these estimates over another? This is the reference class problem.

On the other hand, the *individual to group* perspective treats individual probabilities as the first class objects. This is the perspective most familiar in machine learning and statistics: models f are learned from data with the goal of mapping individuals (e.g. "Alice") to individual probabilities for the outcome of interest, $f(\text{Alice})$.[1] Models of individual probabilities can also be aggregated over to give predicted probabilities conditional on reference classes. If we want to evaluate the probability of an outcome conditional on some reference class S, we can do so by averaging the model's predictions over individuals in S. We cannot measure individual probabilities, but from data we can measure the average probability of an outcome over a sufficiently large reference class S, which gives us a way to empirically falsify a model f from data: if the prediction implied by f for the average outcome conditional on a large reference class does not match the average outcome we can measure from the data, then the model f must be wrong. Multicalibration, introduced by [27], gives us a way to build models of individual probabilities that are consistent with the data for large numbers of arbitrarily chosen reference classes S — i.e. models that are not empirically falsified by any of the pre-specified reference classes. Nevertheless, multi-calibrated models are not unique: we can have multiple models that have large disagreements in many of their individual predictions that nevertheless are equally consistent with the data on a large collection of reference classes. This is an instance of the *predictive multiplicity* problem [8, 9, 30].

The predictive multiplicity problem is usually not phrased in terms of multicalibration and reference classes, but in terms of *accuracy* or *error.* If a model encodes true individual probabilities, then it will minimize expected squared error[2] amongst all possible models. Moreover, expected squared error is something that we can efficiently estimate from data. Hence, if we have two models f_1 and f_2, and we can infer from data that f_1 has lower expected squared error than f_2, then this is an empirical falsification of the hypothesis that f_2 correctly encodes individual probabilities. This serves as a normative justification for selecting amongst models based on their accuracy, which is a common practice. The predictive multiplicity problem arises when we have two models f_1 and f_2 (and perhaps others) that are equally accurate, but disagree substantially on many of their predictions. More generally the predictive multiplicity problem arises when we have multiple models that differ substantially in their predictions, but are seemingly equally consistent with the data before us.

Despite arising from different conceptions of individual probability, the reference class problem and the predictive multiplicity problem result in the same practical concern: that data do not encode unique estimates for the individual probabilities for many individuals. If this is the case, then what justification do we have in making consequential decisions as a result of predictions that our models make about individual probabilities? How can we justify setting a high rate for Alice's life insurance, denying parole to Bob, or suggesting life-altering preventative surgery to Carol based on the predictions of some model f_1 if we have an equally good (and equally well supported by the data) model f_2 that makes predictions that would lead us to take the opposite course of action?

1.1 Our Results

We show that given a common understanding of the data (or the process of sampling from the data distribution), models of individual probabilities are *contestable* through an efficient model reconciliation process that must lead to broad agreement. Specifically, suppose one party A proposes a model of individual probabilities f_A, that another party B thinks is flawed. B can *contest* f_A by proposing their own model of individual probabilities f_B. There are two possible outcomes:

(1) f_A and f_B agree in their predictions almost everywhere[3]. In this case, it turns out there was no substantial disagreement.
(2) f_A and f_B substantially disagree in their predictions for a large portion of the population.

In the second case, we can efficiently extract from the disagreement region of f_A and f_B a large reference class $S = S(f_A, f_B)$ such that on this reference class, not only do f_A and f_B disagree on individual predictions, they also disagree substantially on their prediction of the average outcome conditional on membership in S. Because S is large, from only a modest amount of data, we can accurately estimate the average outcome conditional on S. But because f_A and f_B have a substantial disagreement about this quantity, our measurement is guaranteed to falsify at least one of the two models.

Suppose it is model f_A that is falsified. Then, using a very simple and efficient model update operation of the same sort used for computing multicalibrated models [27], we can update f_A to produce a new model f'_A that now makes predictions that are correct on average over S. The new model f'_A is guaranteed to have significantly reduced squared error compared to f_A, and so is a better model not only in that it has not yet been falsified, but in that it is more accurate.

After this update, we can then repeat the process: Either f'_A and f_B agree on their predictions almost everywhere, or we can again falsify one of the models and improve it using a large reference class $S' = S(f'_A, f_B)$. The only way for this process to end is with two models that agree in their predictions almost everywhere. Moreover, because each iteration of falsification and improvement improves the expected squared error of at least one of the two models, the process cannot continue for very many iterations — fast agreement of the models is guaranteed.

In Section 4, we formally derive the guarantees of our model reconciliation process under the assumption that we can directly evaluate conditional outcome probabilities conditional on large reference classes S: this makes our analysis more transparent. In the full version, we show that we can run our model reconciliation process on the empirical distribution over a modestly sized dataset that is *sampled* i.i.d. from some unknown underlying distribution,

[1] Here of course what is input to the model is some *representation* of the individual, encoding the information that we have available about them.

[2] or any other proper scoring rule.

[3] We use the expressions "almost everywhere" and "almost agree" as shorthand for quantitative statements that are made explicit in the formal presentation of our results. Insofar as we are working in the context of discrete distributions, it should be clear that we are not using these expressions in their usual measure-theoretic sense. We note that our focus on discrete distributions is merely to avoid dealing with measure-theoretic niceties, and is not essential to any of our results

and that its guarantees carry over to the unknown distribution of interest. Here "modestly sized" means a number of samples that is *independent* of the complexity of the models to be reconciled or the dimension (or any other property) of the underlying distribution, and that depends only polynomially on the quantitative parameters controlling how closely we want the models resulting from the reconciliation process to agree. In Section 5 we show how to affix a single, modestly sized sample to a model (which we call a *contestable* model) which can be used to reconcile that model with an exponentially large sequence of models that might be used to contest it in the future.

2 DISCUSSION

Are "Individual Probabilities" Coherent? What are we assuming when we model the world using "individual probabilities"? Are we assuming some kind of idealized, unrealized randomness, which is simply a poor stand-in for our ignorance of the relevant processes? No. Our modelling choices do not preclude a deterministic world: all true individual "probabilities" could be 0 or 1, simply recording the outcomes of interest. We still allow for *models* to predict non-integer probabilities; we can view these as simply expressing uncertainty about the outcomes, or as encoding objective features of a stochastic universe. Our results imply that after reconciliation, two parties must agree about their assignments of individual probabilities, regardless of the philosophical commitments each may have about the nature of probability.

Agreement is Guaranteed; Not Truth. We emphasize that individual probabilities are not uniquely determined from observed data. Consider a toy model of weather forecasting in which features x encode the date, and outcomes y encode whether or not it rains on that date. The following two situations are observationally indistinguishable:

(1) On every day x, the individual probability of rain $p(x) = 1/2$, and
(2) Before the start of time, God selected a subset of days uniformly at random to have individual probability of rain $p(x) = 1$ and the remaining set to have individual probability of rain $p(x) = 0$.

There is no hope of distinguishing these two situations from data about outcomes alone, and so it is plainly impossible to learn a model that is guaranteed to accurately encode individual probabilities from such data[4]— in fact, it is not clear that this goal is meaningful, as they are not uniquely determined.[5] Nevertheless, suppose we believed that the individual probability of rain was $p(x) = 1/2$ every day. If we met a forecaster who was able to make more accurate predictions (i.e. predictions that had lower squared error) on previously unobserved data, we would be forced to recognize that our model was incorrect — because we could *compare* the performance of the two models on data. This drives our result (and similar work on multiple expert testing [2, 20] — see Section 2.1), and is the reason that we can guarantee *agreement* rather than *truth*. Nevertheless, the updates that result from our reconciliation process always move towards truth—because they are error improving—but they stop when the available models agree, which might be well before truth is attained.

Predictive Multiplicity Comes from Restricting Model Classes. Previous work has empirically noted and quantified the phenomenon of predictive multiplicity—i.e. that solving an error minimization problem over some class of models can result in multiple solutions of (roughly) equivalent error [10, 30]. How do these results square with our contention that the predictive multiplicity problem cannot arise, because two equally accurate *but substantially different* models constructively imply the existence of a more accurate model?

The answer is that predictive multiplicity can arise when models are restricted to lie within some pre-specified hypothesis class, like linear threshold functions, bounded depth decision trees, or neural networks with a particular architecture. Traditionally machine learning is done by optimizing a model within a fixed model class, and this is the setting in which predictive multiplicity has been empirically observed and quantified. In contrast, our algorithm for reconciling pairs of models f_1 and f_2 produces a model f_3 that need not lie in the same model class as f_1 and f_2. This is key to sidestepping the predictive multiplicity problem. Traditional methods in machine learning and statistics optimize over models from restricted classes to avoid the problem of overfitting. In contrast, we avoid overfitting despite not restricting our model classes *a priori* by bounding the *number* of updates that can occur through our reconciliation process.

Do models really predict individual probabilities? Another objection we can imagine is that in the settings we discuss—weather prediction, life insurance underwriting, recidivism prediction, etc.—it is logically impossible to observe repeated trials, because tomorrow will only occur once, Alice has only one life to live, and so on. In contrast, when we move to the formalism of a probability distribution over representations of individuals, it may be extremely unlikely (or even a measure 0 event) to observe the same representation of an individual multiple times, but it is no longer a logical impossibility. Said another way, when we model individuals in some representation space, we may fail to record idiosyncratic details of the individual, and so we are no longer speaking of individual probabilities, but rather average outcomes over the reference class defined by people who share the same representation. But this is not a sharp distinction, because our results have no dependence at all on the dimensionality or complexity of the representation we use for individuals. For this objection to have teeth, it must be that there is some crucial idiosyncrasy of an individual that we have failed to capture in our representation: if so, add this to our representation! Our results remain the same (not just qualitatively but also quantitatively) even if the representation of every individual records the position of every molecule in their body, a complete history of their life from birth until the present, or anything else,

[4] Of course, we do not take such radical under-determination of individual probability assignments from data about outcomes alone to in anyway impugn the objectivity of such assignments. Indeed, the primary virtue of our results from a philosophical perspective is that they provide an efficient method to guarantee inter-subjective agreement about individual probability assignments and thus secure their objectivity to this extent.

[5] It is perhaps worth remarking that in this case Dawid's uniqueness result [12], cited earlier, implies, with probability one with respect to God's choices, that there *is* an asymptotically unique *computable* assignment of probabilities that is computably calibrated with the data generated by God's choices. There is no paradox here: insofar as God's choices are generated uniformly at random, her (deterministic) probability forecast is, with probability one, algorithmically random, and thus not computable.

and so does not rely even implicitly on having only an impoverished representation of an individual to work with rather than "the real thing".

2.1 Additional Related Work

Our work is related to a number of strands of literature across statistics, economics, and computer science. Aumann [3] proved that two Bayesians who share a common prior, but may have made different observations, must agree on the posterior expectation of a random variable if their posterior distributions are common knowledge. Although Aumann's original result was nonconstructive, subsequent work has shown that agreement can be reached with finite, communication efficient protocols [1, 23]. Despite similarity in its conclusions, this line of work is quite distinct from ours. In the Bayesian setting that this line of work focuses on, it is immediate that two agents who share the same set of observations and prior beliefs must share the same posterior beliefs (as a posterior distribution is determined, via Bayes rule, as a function only of the prior distribution and observations). Aumann's agreement theorem instead shows that if agents have arrived at common knowledge of their posterior distributions, then their posteriors must agree *even if they have not directly shared their observations*. In contrast, in a frequentist setting, individual probabilities are not uniquely determined from data, which forms the basis of the reference class problem [26] and the model multiplicity problem [8]. Our work considers how two frequentist agents who agree on the same set of data (or the distribution from which it was drawn) must come to agree on individual probabilities — a problem which would not arise in the first place if they were Bayesian agents with a common prior.

[11] proposed calibration as a desirable frequentist condition for evaluating probabilistic forecasts: roughly speaking that the outcome being forecast should have appeared with empirical frequency p conditional on the forecaster predicting probability p of the outcome, simultaneously for all predictions p. Subsequently, [12] studied a substantial strengthening of this condition called *computable calibration* that requires calibration to hold simultaniously on all computable subsets of the data.Dawid proved that in the infinite data limit, two computably calibrated forecasters must approximately agree in their predictions almost everywhere — that is, except on a finite subset of the data [12]. He notes explicitly that this criterion is not of practical use in finite data scenarios, and speculates about the desirability of restrictions of computable calibration to finite sample scenarios (anticipating *multicalibration* [27]). Multicalibration [27] asks for calibration on a restricted class of subsets of the data. [27] gave algorithms for learning multicalibrated predictors with data requirements that scale only modestly with the number of subsets of the data on which calibration is required (and efficient algorithms whenever it is possible to efficiently optimize over these subsets)—but multicalibrated forecasts need not be unique. [28] generalized multicalibration (which aims to be consistent with *mean* outcomes) to moments and other properties of real valued outcomes, and gave efficient algorithms for obtaining these guarantees. [17] generalized multicalibration to notions of "outcome indistinguishability" that ask that a probabilistic forecaster be indistinguishable from a true probabilistic model with respect to a hierarchy of distinguishers that might have access not just to the predictions but to the implementation details of the forecaster itself. [17] explicitly connect outcome indistinguishability to philosophical questions surrounding individual probabilities. Multicalibration has proven to be an effective technique for improving individual predictions in several applications in predictive medicine [4, 5]

[21] gave the first algorithm to constructively make predictions of individual probabilities guaranteed to generate calibrated forecasts against arbitrary sequences of outcomes (and so necessarily without any knowledge of the "true individual probabilities", since the outcomes can be generated adversarially, with knowledge of the predictor's algorithm). [33] show constructively how to achieve calibration in the infinite data limit on any computable subsequence of an arbitrary sequence of outcomes. [32] showed that *any* empirical test (not just calibration tests) that is guaranteed to pass an expert who is forecasting true individual probabilities can be passed by a prediction algorithm on any sequence of outcomes. This is closely related to the fact that individual probabilities are not uniquely specified by data — and so we cannot attempt to test an expert by computing unique individual probabilities ourselves. [25] gave computationally and sample efficient algorithms for achieving multi-calibrated forecasts against arbitrary sequences of outcomes — for means, moments, and quantiles. [7] gave practical implementations of quantile multicalibration algorithms in adversarial sequential settings, and applied them to give algorithms for producing prediction sets of various kinds of classifiers with calibrated, group-wise conditionally valid guarantees.

Although [32] showed that no empirical test of outcomes can distinguish a forecaster with knowledge of true individual probabilities from one without such knowledge in *isolation*, [2] and [20] showed that there are *comparative tests* that can distinguish between *two* forecasters, one of whom is forecasting true individual probabilities and one of whom is not. In particular, the test of [20] is based on checking for *cross-calibration* between two forecasters — i.e. calibration conditional on the predictions of *both* forecasters, and is driven by the fact that on a sequence of predictions such that one forecaster predicts a probability for an outcome p and the other predicts a probability $p' \neq p$, they cannot both be right, which is empirically verifiable if there are many such rounds. In the context of studying the utility of predictors for downstream fairness interventions, [22] study predictors that are *refinements* of one another (in the sense of [14, 15]). They give a simple algorithm ("Merge") that given any two predictors f_1, f_2, produces a predictor f_3 that is cross-calibrated with respect to f_1 and f_2, and hence is a refinement of both. A variant of the "Merge" algorithm of [22] could be used in place of our "Reconcile" algorithm in our arguments; the two algorithms have incomparable data requirements, but would lead to the same qualitative conclusions.

[24] proposes a framework in which models that are sub-optimal on different subsets of the population can be updated and improved as part of a "bias bounties" program by means of falsification; this is another setting in which models can be made to be contestable.

3 BASIC SETTINGS AND DEFINITIONS

We study prediction tasks over a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here $\mathcal{X}$ represents the *feature* domain and $\mathcal{Y}$ represents the label domain. To avoid dealing with measure-theoretic issues, we assume in this paper that $\mathcal{X}$ is a discrete set, but this is not essential to any of our results. For this paper we will restrict attention to binary prediction tasks, where $\mathcal{Y} = \{0, 1\}$ records the outcome of some binary event. Given a labelled example $(x, y) \in \mathcal{Z}$, we view x as encoding all observable characteristics of the instance (e.g. meteorological conditions in a weather prediction task, demographic attributes and medical history in a predictive medicine task, etc.), and y represents the binary outcome we are trying to predict (and when part of the training data, represents the outcome of the binary event that we have observed and recorded).

We model the world via a distribution $\mathcal{D} \in \Delta\mathcal{Z}$. Generally we will not have a direct description of the distribution, and instead have access only to a *sample* of n datapoints D sampled i.i.d. from $\mathcal{D}$, which we will write as $D \in \mathcal{Z}^n$. We will also sometimes identify a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ with the *empirical distribution* over D, which is simply the discrete distribution that places probability mass $1/n$ on each point (x_i, y_i) for $i \in \{1, \ldots, n\}$.

A model is some function $f : \mathcal{X} \to [0, 1]$, and our (typically unattainable goal) is to find a model f^* that has the property that for all $x \in \mathcal{X}$, $f^*(x) = \Pr_{(x,y)\sim\mathcal{D}}[y = 1|x]$ is the *conditional label expectation* given x, or (since we are assuming labels are binary) just "the individual probability" of the outcome for x.

Suppose someone purports to have a model for individual probabilities f. How can we evaluate whether f is any good? If our goal was purely prediction, we might evaluate f via its *squared error* — i.e. the expected (squared) deviation of its prediction from the true label. This is the objective we would minimize if we were solving (e.g.) a least squares regression problem:

Definition 3.1 (Brier Score). *The squared error (also known as* Brier score*) of a model f evaluated on distribution $\mathcal{D}$ is: $B(f, \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$*

Observe that when we treat a dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ as an empirical distribution, then we have: $B(f, D) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$

The Brier score can be accurately estimated given access only to samples from a distribution, and a justification for evaluating models via their Brier score is that amongst all models, the Brier score is minimized by the true individual probabilities encoded by a probability distribution.

Lemma 3.1. *Fix any probability distribution $\mathcal{D}$ and let $f^*(x) = \Pr_{(x,y)\sim\mathcal{D}}[y = 1|x]$ represent the true individual probabilities encoded by $\mathcal{D}$. Let $f : \mathcal{X} \to [0, 1]$ be any other model. Then: $B(f^*, \mathcal{D}) \leq B(f, \mathcal{D})$*

Thus if we have two models f_1 and f_2, and can verify from data that $B(f_1, \mathcal{D}) < B(f_2, \mathcal{D})$, this constitutes an empirical falsification that f_2 correctly encodes individual probabilities.

4 A RECONCILIATION PROCEDURE

Suppose we are given two models $f_1, f_2 : \mathcal{X} \to [0, 1]$ that purport to predict individual probabilities. Our principle concern is the "model multiplicity" problem — that f_1 and f_2 differ substantially in their predictions, and yet we cannot falsify either of the two models from the data. Thus we will be interested in regions in which these models disagree substantially in their predictions. We will define "substantially" by an arbitrarily small discretization parameter ϵ:

Definition 4.1. *Two models f_1 and f_2 have an ϵ-disagreement on a point $x \in \mathcal{X}$ if $|f_1(x) - f_2(x)| > \epsilon$.*

Let $U_\epsilon(f_1, f_2)$ be the set of points on which f_1 and f_2 have an ϵ-disagreement: $U_\epsilon(f_1, f_2) = \{x : |f_1(x) - f_2(x)| > \epsilon\}$

Informally, we will say that if f_1 and f_2 do not have an ϵ-disagreement on x that they agree on x.

One way that we can empirically falsify a model f is by measuring the average outcome y on some *subset* or *group* defined on the data, and comparing it to the average prediction of the model f on the same subset. If the two differ substantially, the model must be incorrect. But from finite data we will only be able to accurately measure these averages on groups that are sufficiently large. We will model groups g as indicator functions $g : \mathcal{X} \to \{0, 1\}$ that specify whether or not each data point x is in the group ($g(x) = 1$) or not ($g(x) = 0$). We will use the following notation to measure the size of a group as measured on the underlying distribution $\mathcal{D}$:

Definition 4.2. *Under a distribution $\mathcal{D}$, a group $g : \mathcal{X} \to \{0, 1\}$ has probability mass $\mu(g)$ defined as: $\mu(g) = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1]$.*

Given a model f and a group g, we can define a quantitative extent to which the average prediction of the model on points in g compares to the average (expected) outcome on points in g. In expectation over the distribution, these two quantities should agree exactly if $f = f^*$ actually encodes true individual probabilities, but from data we will only be able to estimate these quantities approximately. Thus we define an approximate notion of agreement, which we call approximate group conditional mean consistency. Models f that can be shown not to satisfy approximate group conditional mean consistency on any group g have been falsified, in that this constitutes a proof that $f \neq f^*$ (i.e. f must not encode true individual probabilities).

Definition 4.3. *A model $f : \mathcal{X} \to [0, 1]$ satisfies α-approximate group conditional mean consistency with respect to a group $g \in \mathcal{G}$ if:*

$$\left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f(x)|g(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y|g(x) = 1]\right)^2 \leq \frac{\alpha}{\mu(g)}$$

Note that we parameterize α-approximate group conditional mean consistency so that it asks for a weaker condition the smaller the size $\mu(g)$ of the group g. Informally, for a fixed value of α, it asks that the deviation between the average prediction of f and the actual expected outcomes y on a group g differ by at most an error parameter that is proportional to $1/\sqrt{\mu(g)}$. This will turn out to be the "right" scaling because it corresponds to the precision to which we can measure these quantities from data.

We will show a quantitative version of the following statement. It must be the case that *either*

(1) f_1 and f_2 agree on almost all of their predictions, or
(2) f_1, or f_2, or both can be proven from the data to violate a group conditional mean consistency condition on a large set of points. In this case, the falsified model can be "patched" with a simple update in a way that improves its accuracy.

The result is that there can be no substantial disagreements about individual probabilities by people who are willing to be convinced by the evidence of the data before them: models which disagree on a substantial fraction of their predictions witness for each other places in which their predictions are falsified by the data, and provide the means to correct (and improve) each other. Thus disagreements can be leveraged to produce improved models, and this process necessarily converges only when the models agree.

To formalize this, we start by partitioning the set of ϵ-disagreements $U_\epsilon(f_1, f_2)$ into two additional sets that will be important — the set of disagreements on which $f_1(x) > f_2(x)$, and the set of disagreements on which $f_1(x) < f_2(x)$.

Definition 4.4. *Fix any two models $f_1, f_2 : \mathcal{X} \to [0,1]$ and any $\epsilon > 0$. Define the sets:*

$$U_\epsilon^>(f_1, f_2) = \{x \in U_\epsilon(f_1, f_2) : f_1(x) > f_2(x)\}$$
$$U_\epsilon^<(f_1, f_2) = \{x \in U_\epsilon(f_1, f_2) : f_1(x) < f_2(x)\}$$

Based on these sets, for $\bullet \in \{>, <\}$ and $i \in \{1, 2\}$ define the quantities:

$$v_*^\bullet = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | x \in U_\epsilon^\bullet(f_1, f_2)] \; v_i^\bullet = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_i(x) | x \in U_\epsilon^\bullet(f_1, f_2)]$$

Our analysis will proceed by showing that if $U_\epsilon(f_1, f_2)$, the set of ϵ-disagreements of f_1 and f_2 is large, then at least one of the two sets $U_\epsilon^>(f_1, f_2)$ and $U_\epsilon^<(f_1, f_2)$ will witness a large violation of group conditional mean consistency for at least one of the two models.

Lemma 4.1. *Fix any two models $f_1, f_2 : \mathcal{X} \to [0,1]$ and any $\epsilon > 0$. If the fraction of points on which f_1 and f_2 have an ϵ disagreement has mass $\mu(U_\epsilon(f_1, f_2)) = \alpha$ then for some $\bullet \in \{>, <\}$ some $i \in \{1, 2\}$, we have that:*

$$\mu(U_\epsilon^\bullet(f_1, f_2)) \cdot \left(v_*^\bullet - v_i^\bullet\right)^2 \geq \frac{\alpha\epsilon^2}{8}$$

In other words, at least one of the sets $U_\epsilon^>(f_1, f_2)$ and $U_\epsilon^<(f_1, f_2)$ is a group that witnesses an $\frac{\alpha\epsilon^2}{8}$-mean consistency violation for at least one of the models f_1 and f_2.

Proof. Since $U_\epsilon(f_1, f_2)$ can be written as the disjoint union:

$$U_\epsilon(f_1, f_2) = U_\epsilon^>(f_1, f_2) \cup U_\epsilon^<(f_1, f_2)$$

we must have that for at least one value of $\bullet \in \{>, <\}$ we have that:

$$\mu(U_\epsilon^\bullet(f_1, f_2)) \geq \frac{\alpha}{2}.$$

Since the points in $U_\epsilon^\bullet(f_1, f_2)$ are ϵ-separated, we must have that $|v_1^\bullet - v_2^\bullet| \geq \epsilon$. Therefore, for at least one of $i \in \{1, 2\}$ we must have that

$$|v_i^\bullet - v_*^\bullet| \geq \frac{\epsilon}{2}$$

Combining these two claims, we must have that:

$$\mu(U_\epsilon^\bullet(f_1, f_2)) \cdot (v_i^\bullet - v_*^\bullet)^2 \geq \frac{\alpha\epsilon^2}{8}$$

□

Let's consider the significance of this Lemma. Most basically, if we have two models f_1 and f_2 that disagree substantially, this lemma gives an easily constructable set ($U_\epsilon^>(f_1, f_2)$ or $U_\epsilon^<(f_1, f_2)$) that falsifies by a substantial quantitative margin either the assertion that f_1 encodes true conditional label expectations or the assertion that f_2 does. Next, we show that not only do these sets falsify that at least one of f_1 or f_2 are a "correct" model — they provide a directly actionable way to improve one of the models. We prove the following lemma (which is closely related to the kinds of updates used to obtain multicalibrated predictors [27]) which shows us how to improve a model given a group g on which the model fails to satisfy approximate group conditional mean consistency.

Lemma 4.2. *Fix any model $f_t : \mathcal{X} \to [0,1]$, group $g_t : \mathcal{X} \to \{0,1\}$, and distribution $\mathcal{D}$. Let*

$$\Delta_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | g_t(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_t(x) | g_t(x) = 1]$$

and

$$f_{t+1} = h(x, f_t; g_t, \Delta_t)$$

where h is a "patch" defined as:

$$h(x, f; g, \Delta) = \begin{cases} f(x) + \Delta & g(x) = 1 \\ f(x) & \text{otherwise} \end{cases}$$

Then:

$$B(f_t, \mathcal{D}) - B(f_{t+1}, \mathcal{D}) = \mu(g_t) \cdot \Delta_t^2$$

In other words: given any model f_t and a group g_t that witnesses a violation of α-approximate group conditional mean consistency on f_t, we can efficiently produce a model f_{t+1} that has Brier score that is smaller by exactly α.

Proof. By the definition of the patch $h(x, f_t; g_t, \Delta_t)$, models f_t and f_{t+1} differ in their predictions only for x such that $g_t(x) = 1$. Therefore we can calculate:

$$\begin{aligned}
& B(f_t, \mathcal{D}) - B(f_{t+1}, \mathcal{D}) \\
= \; & \Pr[g_t(x) = 0] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[(f_t(x) - y)^2 - (f_{t+1}(x) - y)^2 | g_t(x) = 0\right] \\
& + \Pr[g_t(x) = 1] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[(f_t(x) - y)^2 - (f_{t+1}(x) - y)^2 | g_t(x) = 1\right] \\
= \; & \mu(g_t) \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[(f_t(x) - y)^2 - (f_t(x) + \Delta_t - y)^2 | g_t(x) = 1\right] \\
= \; & \mu(g_t) \left(2\Delta_t \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y - f_t(x) | g_t(x) = 1] - \Delta_t^2\right) \\
= \; & \mu(g_t) \left(2\Delta_t^2 - \Delta_t^2\right) \\
= \; & \mu(g_t)\Delta_t^2
\end{aligned}$$

□

Summarizing, whenever we have two models that have ϵ disagreements on an α-fraction of points, we can always constructively falsify at least one of the models, and update it to improve its Brier score by at least $O(\alpha\epsilon^2)$.

Finally, to make our argument that in-sample quantities (i.e. as measured on the data samples) translate to out of sample quantities (measured on the distribution), it will be useful for our algorithm to not use arbitrarily precise values when patching models, but instead values that are rounded to a finite grid:

Definition 4.5. *Fix any integer m. Let $[1/m]$ denote the set of $m + 1$ grid points:*

$$\left[\frac{1}{m}\right] = \left\{0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m-1}{m}, 1\right\}$$

For any value $v \in [0,1]$ let $Round(v; m) = \operatorname{argmin}_{v' \in [1/m]} |v - v'|$ denote the closest grid point to v in $[1/m]$.

Observe that for $v' = \mathrm{Round}(v; m)$ we always have that $|v - v'| \leq \frac{1}{2m}$

We put this all together in Algorithm 1 (Reconciler). For simplicity of exposition, we initially describe and analyze Algorithm 1 as if it has direct access to distributional quantities. In practice, of course, we will have access only to samples from the distribution, and we will have to run an algorithm on a dataset $D \in \mathcal{Z}^n$ consisting of these samples. We can do so by interpreting D as the uniform distribution over the samples contained within it. In Section **??**, we show that when we run Reconcile$(f_1, f_2, \alpha, \epsilon, D)$ on a dataset $D \sim \mathcal{D}^n$ consisting of n i.i.d. samples from $\mathcal{D}$, then the guarantees we prove in Theorem 4.1 with respect to the empirical distribution D translate over to the true distribution $\mathcal{D}$ with error terms quickly tending to 0 as the number of samples n grows large.

Algorithm 1: Reconcile$(f_1, f_2, \alpha, \epsilon, \mathcal{D})$

```
Let $t = t_1 = t_2 = 0$ and $f_1^{t_1} = f_1, f_2^{t_2} = f_2$.
Let $m = \lceil \frac{2}{\sqrt{\alpha}\epsilon} \rceil$
while $\mu(U_\epsilon(f_1^{t_1}, f_2^{t_2})) \geq \alpha$ do
  For each $\bullet \in \{>, <\}$ and $i \in \{1, 2\}$ Let:
        $v_*^\bullet = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | x \in U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2})]\ \ v_i^\bullet$
            $= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_i^{t_i}(x) | x \in U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2})]$
  Let:
        $(i_t, \bullet_t) = \mathop{\mathrm{argmax}}_{i\in\{1,2\},\bullet\in\{>,<\}} \mu(U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2})) \cdot \left(v_*^\bullet - v_i^\bullet\right)^2$
  breaking ties arbitrarily.
  Let:
        $g_t(x) = \begin{cases} 1 & x \in U_\epsilon^{\bullet_t}(f_1^{t_1}, f_2^{t_2}) \\ 0 & \text{otherwise} \end{cases}$
  Let:
  $\tilde{\Delta}_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | g_t(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_{i_t}^{t_{i_t}}(x) | g_t(x) = 1]$
        $\Delta_t = \mathrm{Round}(\tilde{\Delta}_t; m)$
  Let: $f_i^{t_i+1}(x) = h(x, f_i^{t_i}, g_t, \Delta_t)$, $t_i = t_i + 1$, $t = t + 1$.
Output $(f_1^{t_1}, f_2^{t_2})$.
```

Theorem 4.1. *For any pair of models $f_1, f_2 : \mathcal{X} \to [0, 1]$, any distribution $\mathcal{D}$, and any $\alpha, \epsilon > 0$, Algorithm 1 (Reconcile) runs for $T = T_1 + T_2$ many rounds and outputs a pair of models $(f_1^{T_1}, f_2^{T_2})$ such that:*

(1) $T \leq (B(f_1, \mathcal{D}) + B(f_2, \mathcal{D})) \cdot \frac{16}{\alpha\epsilon^2}$

(2) $B(f_1^{T_1}, \mathcal{D}) \leq B(f_1, \mathcal{D}) - T_1 \cdot \frac{\alpha\epsilon^2}{16}$ and $B(f_2^{T_2}, \mathcal{D}) \leq B(f_2, \mathcal{D}) - T_2 \cdot \frac{\alpha\epsilon^2}{16}$

(3) $\mu(U_\epsilon(f_1^{T_1}, f_2^{T_2})) < \alpha$.

Remark 4.1. *The third conclusion of Theorem 4.1 states that the final models output $(f_1^{T_1}, f_2^{T_2})$ approximately agree on their predictions of individual probabilities almost everywhere. The first conclusion states that the reconciliation procedure converges quickly. The second condition of Theorem 4.1 focuses on one way that the output models $(f_1^{T_1}, f_2^{T_2})$ are superior to the input models (f_1, f_2) — they are more accurate. But there is also another way: Every intermediate model $f_1^{t_1}$ and $f_2^{t_2}$ for $t_1 < T_1$ and $t_2 < T_2$ considered by the reconciliation procedure but ultimately not output has been falsified via the demonstration of a set $U_\epsilon^\bullet(\cdot, \cdot)$ on which it fails to satisfy α-approximate group conditional mean consistency for a large value of α.*

Proof. By Lemma 4.1, for each round $t < T$ we must have that:

$$\mu(U_\epsilon^{\bullet_t}(f_1^{t_1}, f_2^{t_2})) \cdot \left(v_*^{\bullet_t} - v_{i_t}^{\bullet_t}\right)^2 \geq \frac{\alpha\epsilon^2}{8}$$

Let $\tilde{f}_t^{t_i+1} = h(x, f_i^{t_i}, g_t, \tilde{\Delta}_t)$ — i.e. the update that would have resulted at round t had the algorithm used the unrounded measurement $\tilde{\Delta}_t$ rather than the rounded measurement Δ_t. By Lemma 4.2, we have that:

$$B(f_t^{t_i}, \mathcal{D}) - B(\tilde{f}_t^{t_i+1}, \mathcal{D}) \geq \frac{\alpha\epsilon^2}{8}$$

.

We can now compute

$$\begin{aligned} & B(f_t^{t_i}, \mathcal{D}) - B(f_t^{t_i+1}, \mathcal{D}) \\ = \quad & (B(f_t^{t_i}, \mathcal{D}) - B(\tilde{f}_t^{t_i+1}, \mathcal{D})) - (B(f_t^{t_i+1}, \mathcal{D}) - B(\tilde{f}_t^{t_i+1}, \mathcal{D})) \\ \geq \quad & \frac{\alpha\epsilon^2}{8} - (B(f_t^{t_i+1}, \mathcal{D}) - B(\tilde{f}_t^{t_i+1}, \mathcal{D})) \end{aligned}$$

So it remains to upper bound $(B(f_t^{t_i+1}) - B(\tilde{f}_t^{t_i+1}))$. Let $\hat{\Delta} = \tilde{\Delta}_t - \Delta_t$. We make several observations: First, $\tilde{f}_t^{t_i+1} = h(x, f_t^{t_i+1}, g_t, \hat{\Delta})$. Second,

$$\begin{aligned} \hat{\Delta} &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | g_t(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_i^{t_i}(x) | g_t(x) = 1] - \Delta_t \\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | g_t(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_i^{t_i+1}(x) | g_t(x) = 1] \end{aligned}$$

Third, by definition of the Round operation, $|\hat{\Delta}| \leq \frac{1}{2m}$. Therefore we can again apply Lemma 4.2 to conclude that:

$$\begin{aligned} B(f_t^{t_i+1}, \mathcal{D}) - B(\tilde{f}_t^{t_i+1}, \mathcal{D}) &= \mu(g_t)\hat{\Delta}^2 \\ &\leq \frac{1}{4m^2} \end{aligned}$$

Combining this with our initial calculation lets us conclude that:

$$B(f_t^{t_i}, \mathcal{D}) - B(f_t^{t_i+1}, \mathcal{D}) \geq \frac{\alpha\epsilon^2}{8} - \frac{1}{4m^2} \geq \frac{\alpha\epsilon^2}{16}$$

Here we are using the fact that we have set $m \geq \frac{2}{\sqrt{\alpha}\epsilon}$. Applying this lemma for each of the T_1 and T_2 updates for f_1 and f_2, respectively, we get that: $B(f_1^{T_1}, \mathcal{D}) \leq B(f_1, \mathcal{D}) - T_1 \cdot \frac{\alpha\epsilon^2}{16}$ and $B(f_2^{T_2}, \mathcal{D}) \leq B(f_2, \mathcal{D}) - T_2 \cdot \frac{\alpha\epsilon^2}{16}$. Since Brier scores are non-negative, we conclude that $T_1 \leq B(f_1, \mathcal{D})\frac{16}{\alpha\epsilon^2}$ and $T_2 \leq B(f_2, \mathcal{D})\frac{16}{\alpha\epsilon^2}$. Thus $T = T_1 + T_2 \leq (B(f_1, \mathcal{D}) + B(f_2, \mathcal{D})) \cdot \frac{16}{\alpha\epsilon^2}$

Finally the halting condition of the algorithm implies that:

$$\mu(U_\epsilon(f_1^{T_1}, f_2^{T_2})) < \alpha.$$

□

Thus if we start with any two models that have substantial disagreement, we are guaranteed to be able to efficiently produce *strictly improved* models that almost agree almost everywhere. In particular, we can never be in a position in which we have two equally accurate *but unimprovable* models that have substantial disagreements: in this case, we can always improve the models. The only time we can have substantial model disagreement is if we refuse to improve the models even in the face of efficiently verifiable and actionable evidence that one of the models is suboptimal and improvable.

We observe that any pair of models that have gone through the "Reconcile" process must also produce very similar estimates for the conditional label expectation over any sufficiently large reference class (i.e. any subset of the feature space). In particular, for any sufficiently large reference class, either both models are consistent with the data or they are not — but they cannot substantially disagree.

Corollary 4.1. *Let $E \subset \mathcal{X}$ be any subset of the feature space. Let f_1 and f_2 be any two models that have been output by Algorithm 1 (Reconcile) with parameters ϵ and α. Let:*

$$p_1(E) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_1(x)|x \in E] \text{ and } p_2(E) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[f_2(x)|x \in E]$$

be the estimates for $\Pr[y = 1|x \in E]$ implied by models f_1 and f_2 respectively. Then:

$$|p_1(E) - p_2(E)| \leq \frac{\alpha}{\mu(E)} + \epsilon$$

Proof. Let $S_\epsilon(f_1, f_2) = \{x : x \notin U_\epsilon(f_1, f_2)\}$ be the set of points on which f_1 and f_2 do not have an ϵ-disagreement. Recall that $\mu(S_\epsilon(f_1, f_2)) \geq 1 - \alpha$. We compute:

$$\begin{aligned}
& \mu(E)|p_1(E) - p_2(E)| \\
= & \left|\sum_{x\in E} \mu(\{x\}) \cdot (f_1(x) - f_2(x))\right| \\
= & \Bigg| \sum_{x\in E\cap U_\epsilon(f_1,f_2)} \mu(\{x\}) \cdot (f_1(x) - f_2(x)) \\
& + \sum_{x\in E\cap S_\epsilon(f_1,f_2)} \mu(\{x\}) \cdot (f_1(x) - f_2(x))\Bigg| \\
\leq & \ \alpha + \mu(E \cap S_\epsilon(f_1, f_2))\epsilon \\
\leq & \ \alpha + \mu(E)\epsilon
\end{aligned}$$

Dividing by $\mu(E)$ yields the corollary. □

5 CONTESTABLE MODELS

Thus far we have considered the problem of reconciling *two* models f_1 and f_2, and have shown that we require only $O(1/(\alpha^3\epsilon^2))$ many points to obtain strictly improved models f_1', f_2' that have ϵ disagreements on at most an α measure of points. But what if someone then proposes a third model, f_3, and then another f_4, etc? We could run the reconciliation process again each time—and perhaps if we had k models, repeatedly in a pairwise fashion until all k of the models approximately agreed—but this would naively require a fresh set of samples for each new reconciliation procedure. In this section, we show how to do better: we attach to f just a single sample of "contestation" data that is of size polynomial in our target reconciliation parameters α and ϵ (and independent of the complexity of the model or distribution). Using this data, we show that we can then put f through a reconciliation procedure with a very large (exponential in the size of its contestation data set) number of models, with the same guarantees as if we had run the models through Algorithm 1 each time. Driving this result is the observation that each time a particular model f is updated using the patch operation defined in Lemma 4.2, f's squared error drops, independently of which reconciliation process the update is a part of — and thus the total number of large updates made to a single model is bounded independently of the number of other models that are "reconciled" with it. This, together with results from adaptive data analysis that allow us to repeatedly re-use hold-out sets while preserving statistical validity [6, 16, 29] are enough to give the result.

We define a "contestable model" to be a model f attached to a fixed sample of "contestation data". A "contestable" model can be "contested" by identifying any subset of the data identified by an indicator function $g : \mathcal{X} \to \{0, 1\}$. The guarantee of a contestable model is that if it is contested using a subset of the data that is simultaneously *large* and on which the expectation of the model's predictions is substantially different than the expectation of the label, then the model will be updated in a way that corrects the discovered error on the identified subset of data, and strictly improves the squared error of the model. These contestations are *accepted.* Contestations can also be *rejected* on the grounds either that the group identified is too small, or that the model already predicts on average a value over that group that is sufficiently close to the true label mean over that group. We aim to design contestable models that can receive a number of contestations over their lifetime that is exponential in the size of their contestation dataset.

Definition 5.1. *A* contestable model $\mathbf{f}$ *consists of a current model $f_c : \mathcal{X} \to [0, 1]$, a dataset $D \in \mathcal{Z}^n$, and has two operations: $\mathbf{f}$.predict(x) which takes as input a data point $x \in \mathcal{X}$ and $\mathbf{f}$.contest(g) which takes as input the indicator function for a group $g : \mathcal{X} \to \{0, 1\}$. $\mathbf{f}$.predict(x) outputs $f_c(x)$, where f_c is the current model belonging to $\mathbf{f}$, and $\mathbf{f}$.contest(g) may update the current model f_c to a new model f_{c+1} according to Algorithm 2.*

Algorithm 2, which follows, is a randomized algorithm: it samples from the *Laplace* distribution. We write Lap(b) to denote the sampling operation for the centered Laplace distribution with scale parameter b, which is the distribution that has probability density function $f(x; b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right)$.

The analysis of Algorithm 2 is in the full version and goes through *differential privacy* [18]. Differential privacy was originally introduced as a strong notion of privacy that could be satisfied while still carrying out high accuracy statistical analyses, but has since found many other uses. Our interest in differential privacy will be because of the transfer theorems of [6, 16, 29] which informally state that analyses that are both differentially private and accurate on a sample of data D drawn i.i.d. from an underlying distribution $\mathcal{D}$ must also be accurate on the underlying distribution. Algorithm 2 is an instantiation of Algorithm 3 (NumericSparse) from [19], from which it follows that the algorithm is differentially private in the contestation dataset D. We then apply the version of the "transfer theorem" given in [29], which establishes that its estimates of statistics on D are representative of their true values on the underlying distribution $\mathcal{D}$ from which D was drawn. Our use of differential privacy here to get out of sample guarantees closely mirrors its use in [27] to get a generalization theorem for multicalibration algorithms. In fact, if the "contestation" sets g_t submitted to $\mathbf{f}$.Contest(g_t) were the groups with respect to which

Algorithm 2: $\mathbf{f}$.Contest

```
Given: A failure probability $\delta$, a dataset $D \in \mathcal{Z}^n$, an initial
model $f = f_0$, a threshold $T$ to accept attempted contestations, a
target total number of contestations $K$, and an upper bound $C$
on the total number of accepted contestations.
Let $t = 0$ denote a count of the number of attempted
contestations and $c = 0$ denote the count of the number of
accepted contestations.
Let privacy parameter $\epsilon = \sqrt{\frac{\log \frac{K}{\delta}\sqrt{C \ln \frac{1}{\delta}}}{n}}$
Let $\epsilon_1 = \frac{\sqrt{512}}{\sqrt{512}+1}\epsilon$, $\epsilon_2 = \frac{1}{\sqrt{512}+1}\epsilon$
Let $\sigma(\epsilon) = \frac{\sqrt{32C\log(1/\delta)}}{\epsilon n}$
Let $\hat{T}_0 = T + \mathrm{Lap}(\sigma(\epsilon_1))$
while there is another model $g_t$ given as input to $\mathbf{f}$.contest($g_t$)
and $c < C$ do
    Compute an empirical estimate of
    $\mu(g_t) \cdot \mathbb{E}[y - f_c(x)|g_t(x) = 1]$:
        $$\eta_t(f_c, g_t) = \frac{1}{n}\sum_{(x,y)\in D}(y - f_c(x)) \cdot g_t(x)$$
    Let $\hat{\eta}_t = |\eta_t(f_c, g_t)| + \mathrm{Lap}\left(2\sigma(\epsilon_1)\right)$
    if $\hat{\eta}_t \geq \hat{T}_c$ then
        The contestation is accepted.
        Let:
        $$\tilde{\mu}_t = \frac{1}{n}\sum_{(x,y)\in D} g_t(x) + \mathrm{Lap}(2\sigma(\epsilon_2)) \quad \tilde{\eta}_t = \eta_t(f_c, g_t) + \mathrm{Lap}(2\sigma(\epsilon_2))$$
        Let $\tilde{\Delta}_t = \frac{\tilde{\eta}_t}{\tilde{\mu}_t}$
        Let $f_{c+1}(x) = h(x, f_c, g_t, \tilde{\Delta}_t)$, $c = c + 1$.
        Let $\hat{T}_c = T + \mathrm{Lap}(\sigma(\epsilon_1))$
    else
        The contestation is rejected.
    Let $t = t + 1$.
Halt.
```

which multicalibrated predictors are required to satisfy group conditional mean consistency, then Algorithm 2 would essentially (up to some details) be the multicalibration algorithm originally given by [27]. But a contestable model can take as input the indicator function g_t of *any* group, including those groups $U_\epsilon^\bullet(f_1^{t_1}, f_2^{t_2})$ used as updates within Reconcile (Algorithm 1). Thus, contestable models will be able to be reconciled with many other models in a data efficient way (in addition to being "contested" on other groups on which they fail to satisfy group conditional mean consistency).

Theorem 5.1. *Initialized with a dataset $D \sim \mathcal{D}^n$ of size n sampled i.i.d. from $\mathcal{D}$, a target number of contestations K, a failure probability δ, a threshold $T = \Theta\left(\frac{\left(\log\frac{K}{\delta}\right)^{1/3}\left(\ln\frac{1}{\delta}\right)^{1/6}}{n^{1/3}}\right)$ and a limit on successful contestations $C = \Theta\left(\frac{1}{T^2}\right)$, a contestable model will with probability $1 - 2\delta n$:*

1. *Process at least K contestations g_t without halting,*
2. *Guarantee that every accepted contestation g_t is such that: $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[g_t(x)(y - f_c(x))]\right| \geq \Omega\left(\frac{\left(\log\frac{K}{\delta}\right)^{1/3}\left(\ln\frac{1}{\delta}\right)^{1/6}}{n^{1/3}}\right)$ and produces an update that reduces the squared error of* $\mathbf{f}$ *by $B(f_c) - B(f_{c+1}) = \Omega(T^2) = \Omega\left(\frac{\left(\log\frac{K}{\delta}\right)^{2/3}\left(\ln\frac{1}{\delta}\right)^{1/3}}{n^{2/3}}\right)$.*
3. *Guarantee that every rejected contestation g_t is such that: $\left|\mathbb{E}_{(x,y)\sim\mathcal{D}}[g_t(x)(y - f_c(x))]\right| \leq O\left(\frac{\left(\log\frac{K}{\delta}\right)^{1/3}\left(\ln\frac{1}{\delta}\right)^{1/6}}{n^{1/3}}\right)$.*

The proof of Theorem 5.1 can be found in the full version.

We now observe how a contestable model with the guarantees of Theorem 5.1 can be repeatedly used as part of a reconciliation procedure akin to Algorithm 1.

Algorithm 3: Contestable-Reconcile($\mathbf{f}_1, \mathbf{f}_2, \alpha, \epsilon, \mathcal{D}$)

```
while $\mu(U_\epsilon(\mathbf{f}_1, \mathbf{f}_2)) \geq \alpha$ do
    For $\bullet \in \{>, <\}$ let:
        $$g^\bullet(x) = \begin{cases} 1 & x \in U_\epsilon^\bullet(\mathbf{f}_1, \mathbf{f}_2) \\ 0 & \text{otherwise} \end{cases}$$
    $\mathbf{f}_1.contest(g^>)$, $\mathbf{f}_1.contest(g^<)$, $\mathbf{f}_2.contest(g^>)$,
    $\mathbf{f}_2.contest(g^<)$
```

The idea is simple (and outlined in Algorithm 3): While we have two contestable models $\mathbf{f}_1$ and $\mathbf{f}_2$ that have ϵ-disagreements on more than an α fraction of the distribution, contest both models on the disagreement sets $U_\epsilon^>(\mathbf{f}_1, \mathbf{f}_2)$ and $U_\epsilon^<(\mathbf{f}_1, \mathbf{f}_2)$. By Lemma 4.1, if indeed $\mu(U_\epsilon^\bullet(\mathbf{f}_1, \mathbf{f}_2)) \geq \alpha$, then for at least one of the models $i \in \{1, 2\}$ and for at least one of the sets $\bullet \in \{>, <\}$, we must have:

$$\begin{aligned} & \left|\mu(U_\epsilon^\bullet(f_1, f_2)) \cdot \mathbb{E}[y - f_c(x)|x \in U_\epsilon^\bullet(f_1, f_2)]\right| \\ \geq \; & \mu(U_\epsilon^\bullet(f_1, f_2)) \cdot \mathbb{E}[y - f_c(x)|x \in U_\epsilon^\bullet(f_1, f_2)]^2 \\ \geq \; & \frac{\alpha\epsilon^2}{8} \end{aligned}$$

By Theorem 5.1, assuming that the contestation datasets of both models are of size:

$$n \geq \Omega\left(\frac{\log\frac{K}{\delta}\sqrt{\log\frac{1}{\delta}}}{\alpha^3\epsilon^6}\right)$$

then at least one of these contestations will succeed until (after at most a polynomial number of contestations in α and ϵ), the models are reconciled and $\mu(U_\epsilon^\bullet(\mathbf{f}_1, \mathbf{f}_2)) \leq \alpha$.

Solving for K, we find that a contestable model can be run through

$$K = \tilde{\Theta}\left(\delta \exp\left(\frac{n\alpha^3\epsilon^6}{\sqrt{\log\frac{1}{\delta}}}\right)\right)$$

many contestation procedures given a single contestation dataset of size n. Here the $\tilde{\Theta}$ hides terms that are logarithmic in $1/\alpha$ and $1/\epsilon$.

We emphasize that a contestable model can be contested using K many sets of any nature — these can include the disagreement regions that arise from our Reconcile procedure, but can also include arbitary regions on which the current model is found to be miscalibrated. Thus contestable models can be robustly and iteratively improved over an exponential number of contestations whenever they are falsified by being shown to fail to satisfy group conditional mean consistency on any group. Modest amounts of data, attached to a model as a contestation dataset, can make the model long-lived in an easily adaptable and improvable form.

6 CONCLUSION

Individual probability assignments are not determined by data; this lies at the heart of both the reference class problem and the predictive multiplicity problem. Insofar as individual probability assignments play a significant role in consequential decision-making, their underdetermination by data may give rise to practical problems when we have two or more seemingly equally good estimation methods that nevertheless result in models that differ substantially in the assignments they predict. We show that given modest amounts of data to resolve disagreements, such problems cannot arise at a substantial scale, because if two models disagree substantially in many places, then this large disagreement region itself points us to how to improve at least one of the models. The only way this process can conclude is with improved models that approximately agree almost everywhere. This does not "resolve" the reference class problem, the predictive multiplicity problem, or other puzzles about individual probability in that it does not claim a way to produce "correct" estimates of individual probabilities. But it does remove the practical bite of these problems in that it shows that two parties who agree on the data distribution and who have committed in good faith to make statistical estimates of individual probabilities cannot end up in a state where they substantially disagree on a large number of instances — and hence will rarely face any ambiguity in how they should act, given their statistical modeling.

REFERENCES

[1] Scott Aaronson. 2005. The complexity of agreement. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. 634–643.

[2] Nabil I Al-Najjar and Jonathan Weinstein. 2008. Comparative testing of experts. *Econometrica* 76, 3 (2008), 541–559.

[3] Robert J Aumann. 1976. Agreeing to Disagree. *The Annals of Statistics* 4, 6 (1976), 1236–1239.

[4] Noam Barda, Dan Riesel, Amichay Akriv, Joseph Levy, Uriah Finkel, Gal Yona, Daniel Greenfeld, Shimon Sheiba, Jonathan Somer, Eitan Bachmat, et al. 2020. Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nature communications* 11, 1 (2020), 1–9.

[5] Noam Barda, Gal Yona, Guy N Rothblum, Philip Greenland, Morton Leibowitz, Ran Balicer, Eitan Bachmat, and Noa Dagan. 2021. Addressing bias in prediction models by improving subpopulation calibration. *Journal of the American Medical Informatics Association* 28, 3 (2021), 549–558.

[6] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. 2021. Algorithmic stability for adaptive data analysis. *SIAM J. Comput.* 50, 3 (2021), STOC16–377.

[7] Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. 2022. Practical Adversarial Multivalid Conformal Prediction. *arXiv preprint arXiv:2206.01067* (2022).

[8] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 850–863. https://doi.org/10.1145/3531146.3533149

[9] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[10] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[11] A Philip Dawid. 1982. The well-calibrated Bayesian. *J. Amer. Statist. Assoc.* 77, 379 (1982), 605–610.

[12] A. P. Dawid. 1985. Calibration-Based Empirical Probability. *The Annals of Statistics* 13, 4 (1985), 1251 – 1274. https://doi.org/10.1214/aos/1176349736

[13] Philip Dawid. 2017. On individual risk. *Synthese* 194, 9 (2017), 3445–3474.

[14] Morris H DeGroot and Stephen E Fienberg. 1981. *Assessing Probability Assessors: Calibration and Refinement.* Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS.

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

[16] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. 2015. Generalization in adaptive data analysis and holdout reuse. *Advances in Neural Information Processing Systems* 28 (2015).

[17] Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. 2021. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 1095–1108.

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.

[19] Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[20] Yossi Feinberg and Colin Stewart. 2008. Testing multiple forecasters. *Econometrica* 76, 3 (2008), 561–582.

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

[22] Sumegha Garg, Michael P Kim, and Omer Reingold. 2019. Tracking and improving information in the service of fairness. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. 809–824.

[23] John D Geanakoplos and Heraklis M Polemarchakis. 1982. We can't disagree forever. *Journal of Economic theory* 28, 1 (1982), 192–200.

[24] Ira Globus-Harris, Michael Kearns, and Aaron Roth. 2022. An Algorithmic Framework for Bias Bounties. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1106–1124. https://doi.org/10.1145/3531146.3533172

[25] Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. 2022. Online Multivalid Learning: Means, Moments, and Prediction Intervals. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

[26] Alan Hájek. 2007. The reference class problem is your problem too. *Synthese* 156, 3 (2007), 563–585.

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

[28] Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. 2021. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*. PMLR, 2634–2678.

[29] Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. 2020. 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.

[30] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*. PMLR, 6765–6774.

[31] Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2022. Reconciling Individual Probability Forecasts. *arXiv preprint arXiv:2209.01687* (2022).

[32] Alvaro Sandroni. 2003. The reproducible properties of correct forecasts. *International Journal of Game Theory* 32, 1 (2003), 151–159.

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